MATHEMATICA Policy Research

REPORT

Evaluation of Health Care Innovation Awards (HCIA): Primary Care Redesign Programs

Third Annual Report

March, 2017

Greg Peterson
Laura Blue
Lorenzo Moreno
Keith Kranker
Boyd Gilman
Kate Stewart
Sandi Nelson
Kristin Geonnotti
Sheila Hoag
Andrew McGuirk
Ken Peckham

With the following teams: Impact, Implementation, Data Processing, Survey, Statistics, and Editorial and Production Coordination

Submitted to:
U.S. Department of Health and Human Services
Centers for Medicare & Medicaid Services
7500 Security Blvd.
Baltimore, MD 21244-1850
Project Officer: Timothy Day
Contract Number: HHSM-500-2010-000261/HHSM-500-T0015

Submitted by:
Mathematica Policy Research
P.O. Box 2393
Princeton, NJ 08543-2393
Telephone: (609) 799-3535
Facsimile: (609) 799-0005
Project Director: Greg Peterson
Reference Number: 40274.370

This page left blank for double-sided copying.

Impact Evaluation, Implementation Evaluation, Data Processing, Survey, Statistics, and Editorial and Production Coordination Teams

Primary Care Redesign Awardee	Impact Evaluation Team	Implementation Evaluation Team
Atlantic General Hospital	Keith Kranker	Linda Barterian, Rumin Sarwar
CareFirst Blue Cross Blue Shield	Greg Peterson	Kristin Geonnotti, Lauren Hula
Denver Health and Hospital Authority	Laura Blue	Lauren Hula, Tricia Higgins
Finger Lakes Health Systems Agency	Randall Blair	Rachel Shapiro, Rebecca Coughlin
Pacific Business Group on Health	Sean Orzol, Michael Barna	Rosalind Keith, Rumin Sarwar
PeaceHealth Ketchikan Medical Center	Purvi Sevak	Boyd Gilman, Victoria Peebles
Rutgers Center for State Health Policy	Purvi Sevak	Cara Stepanczuk, Katharine Bradley
Sanford Health One Care	Jelena Zurovac	KeriAnn Wells, Catherine DesRoches
TransforMED	Sean Orzol, Michael Barna	Rosalind Keith, Mynti Hossain
Wyoming Institute of Population Health at Cheyenne Regional Medical Center	Andrea Wysocki	KeriAnn Wells, Emily Ehrlich

Data Processing Team
Alex Bryce
Andrew McGuirk
Sandi Nelson
Ken Peckham
Patrick Wang

Survey Team
Catherine DesRoches
Lauren Harris
Rachel Kogan
Julita Milliner-Waddell
Betsy Santos

Statistics Team
Juan Diego Astudillo
Jared Coopersmith
Bonnie Harvey
Huihua Lu
Lauren Vollmer
Fei Xing
Frank Yoon

Editorial and Production Coordination Team
Mark Ezzo
John Kennedy
Patricia Ciaccio
Felita Buckner

CONTENTS

This page has been left blank for double-sided copying.

GLOSSARY

Awardee Name	Abbreviation
Atlantic General Hospital	AGH
CareFirst Blue Cross Blue Shield	CareFirst
Cooper University Hospital and Camden Coalition of Health Care Providers	CUH/CCHP
Denver Health and Hospital Authority	Denver Health
Finger Lakes Health Systems Agency	FLHSA
Pacific Business Group on Health	PBGH
PeaceHealth Ketchikan Medical Center	PeaceHealth
Research Institute at Nationwide Children's Hospital	NCH
Rutgers Center for State Health Policy	CSHP
Sanford Health	Sanford Health
TransforMED	TransforMED
University Hospitals of Cleveland Rainbow Babies and Children’s Hospital	UHC
Wyoming Institute of Population Health at Cheyenne Regional Medical Center	WIPH

This page left blank for double-sided copying.

EXECUTIVE SUMMARY

I. INTRODUCTION

In 2012, the Center for Medicare & Medicaid Innovation (CMMI) awarded cooperative agreements of up to $30 million to organizations that proposed compelling models for improving quality of care, improving health outcomes, and lowering medical spending for Medicare, Medicaid, and Children's Health Insurance Program (CHIP) beneficiaries. The purpose of these Health Care Innovation Awards (HCIAs) was to expand the source of innovation in health care delivery. CMMI is currently testing many models to improve quality and reduce spending. CMMI designed most of these models, which are fairly prescriptive in what participating providers must do. In contrast, external organizations developed the HCIA models with wide latitude in how to design the innovations. Each awardee proposed its own intervention and target population, leading to substantial variation across the HCIA portfolio in intervention content, who delivered it, who received it, and in what contexts (for example, physical location or type of health system). CMMI classified 14 of the 107 HCIAs issued in 2012 as primary care redesign (PCR) programs, an area of explicit focus under the Patient Protection and Affordable Care Act (ACA), the 2010 legislation that established CMMI.

This report presents findings on the impacts of the HCIA-PCR programs on quality of care, service use, and medical spending during the original three-year award period. This report also integrates these impact results with findings from the implementation evaluation—assessments of how each HCIA-PCR intervention worked, whether it was implemented as intended, and the barriers to and facilitators of successful program implementation. This report builds on earlier implementation findings, reported more extensively in the evaluation's second annual report (Moreno et al. 2016, available on CMMI's website).

This report presents impact and implementation findings during the original three-year award period for 10 of the 14 HCIA-PCR awardees:

1. Atlantic General Hospital (AGH)
2. CareFirst Blue Cross Blue Shield (CareFirst)
3. Denver Health and Hospital Authority (Denver Health)
4. Finger Lakes Health Systems Agency (FLHSA)
5. Pacific Business Group on Health (PBGH)
6. PeaceHealth Ketchikan Medical Center (PeaceHealth)
7. Rutgers Center for State Health Policy (CSHP)
8. Sanford Health
9. TransforMED
10. Wyoming Institute of Population Health at Cheyenne Regional Medical Center (WIPH)

Two of these awardees (CareFirst and FLHSA) received no-cost extensions to continue their interventions for 6 to 12 months after the end of the original three-year award period. For these two awardees, the impact conclusions are preliminary because they do not include these extension periods. An addendum to this report will report impact results including these extension periods and draw final impact conclusions. The impact conclusions for the other awardees, those not granted extensions, are final. Some awardees implemented multiple intervention components, each with its own target population and services. In these cases, we estimated the impacts for the single component that the awardee focused on most heavily, implemented well, and for which credible designs for estimating impacts were possible.

Of the four HCIA-PCR awardees not included in this report, we exclude three because we anticipate higher quality data for the evaluation in the future; we plan to present impact findings for them in the addendum to this report. These awardees are (1) Cooper University Hospital and the Camden Coalition of Healthcare Providers, for which too few program participants had enrolled by the cutoff date for this report to produce reliable impact estimates; (2) the Research Institute at Nationwide Children's Hospital, for which we plan to use Medicaid data, with a relatively long lag between the date of service and data availability; and (3) the University Hospitals of Cleveland Rainbow Babies & Children's Hospital, for which, similarly, we plan to use Medicaid data. We exclude a fourth awardee, the Foundation for California Community Colleges and the Transitions Clinic Network, because we could not develop a credible comparison group with available data to support a robust impact evaluation, as described in the second annual report (Moreno et al. 2016).

The results in this report can help inform decisions by CMMI and other stakeholders about whether and how to incorporate the tested models into their future PCR efforts. Under the ACA, CMMI has the authority to expand models proven to (1) improve quality without raising costs, (2) reduce costs without harming quality, or (3) improve quality and reduce costs simultaneously. Given this authority, CMMI might choose to take those HCIA-PCR models shown to be effective or promising and continue to test them, expand them to novel settings, or incorporate them (or components of them) into other primary care initiatives. Similarly, other stakeholders, such as commercial payers or accountable care organizations, could use the evidence in this report to decide whether to pursue tested innovations in their own distinct contexts. CMMI and other stakeholders might choose not to pursue programs for which the evaluation finds no impacts—particularly for evaluations with strong statistical power to detect impacts had they existed.

The rest of this executive summary describes the evaluation's methods (Section II), summarizes impact results across awardees (Section III), and describes key implementation and impact findings for individual awardees (Section IV), grouped by their overall impact assessment. Section V draws implications of the evaluation's findings for (1) designing tests of similar models to maximize chances of generating credible impact estimates and (2) future efforts to change the delivery of primary care to improve quality and reduce medical spending.

II. METHODS

Because each HCIA-PCR program had unique target populations and interventions, we evaluated each program separately. The evaluation for each program had three parts.

1. **Program implementation.** We examined the intervention itself—including its design, its theory of action (that is, how the awardee expected the intervention to improve patients' outcomes), the extent to which the program was implemented as intended, and barriers to and facilitators of implementation. We based this analysis on three sources of evidence. First, we reviewed documents from the awardee and CMMI's HCIA implementation and monitoring contractor (the Lewin Group). These documents included the awardee's original application to CMMI, quarterly progress reports from the awardee and Lewin, and self-monitoring metrics that the awardee reported about the intervention it delivered. Second, we conducted telephone and on-site interviews with program administrators and frontline staff implementing the interventions. For the on-site interviews, we visited two to four sites implementing the interventions in spring 2014 and again in spring 2015. Third, we surveyed program staff who received HCIA-funded training to identify the training provided and their perceptions of the benefits and limitations of the training for delivering intervention services.
2. **Clinicians' behavior and perceptions.** Most PCR programs used their HCIA to fund activities—such as training clinicians for new roles or providing new health information technology (IT)—that aimed to change how clinicians delivered care to their patients. Awardees expected those changes, in turn, to improve patients' outcomes. For these awardees, we assessed whether the anticipated changes in clinicians' behavior occurred. We surveyed primary care clinicians twice—in spring 2014 and summer 2015—to gauge their perceptions of the program's impact on the quality, timeliness, and other aspects of the care they provided to patients. When possible, we supplemented these survey data with metrics from the awardee about whether specific anticipated changes in providers' behavior occurred.
3. **Impacts on patients' outcomes.** We assessed program impacts on outcomes for Medicare fee-for-service (FFS) beneficiaries that we grouped into four domains: (1) quality-of-care processes, (2) quality-of-care outcomes, (3) service use, and (4) Medicare spending. Table 1 lists the full set of outcomes available for each evaluation domain. We selected outcomes that were measurable in Medicare FFS claims data, that many or all awardees expected to affect, and that the Centers for Medicare & Medicaid Services (CMS) listed as priority measures. For each program's impact evaluation, we further selected from the full set of outcomes available those that the awardee expected to affect. We focused on Medicare FFS beneficiaries in this report because timely claims data were not available for Medicaid or CHIP populations, and because claims do not reliably capture service use for Medicare beneficiaries enrolled in managed care plans.

Table 1. Domains and outcomes used in this evaluation

Domain	Outcome (units)	Calculated for only a subset of the treatment and comparison groups?	CMMI priority measure?[a]
Quality-of-care processes	Received all four recommended diabetes processes of care in the year (% of eligible beneficiaries/year)[b]	Yes, FFS Medicare beneficiaries with diabetes ages 18 to 75	No
	Received recommended lipid profile in the year (% of eligible beneficiaries/year)	Yes, FFS Medicare beneficiaries with IVD ages 18 or older	No
	All inpatient admissions within a quarter were followed by an ambulatory care visits within 14 days (% of eligible beneficiaries/quarter)	Yes, FFS Medicare beneficiaries with at least one inpatient stay in a quarter	No
Quality-of-care outcomes	Inpatient admissions followed by an unplanned readmission within 30 days (#/1,000 beneficiaries/quarter)	No	Yes
	Inpatient admissions for ACSCs (#/1,000 beneficiaries/quarter)	No	No
Service use	All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	No	Yes
	Outpatient ED visit rate—that is, ED visits that did not end in a hospital stay (#/1,000 beneficiaries/quarter)	No	Yes
Spending	Medicare Part A and B spending ($/beneficiary/month)	No	Yes
	Medicare spending for inpatient stays ($/beneficiary/month)	No	No

Note: This table lists all outcomes available for the impact evaluation. We selected all or a subset of these outcomes for the impact evaluation for each individual program, depending on whether the program expected to affect the outcome.

[a] Measures that CMMI has indicated are a priority for evaluations of all HCIA programs, not only those within primary care.

[b] The four recommended processes are dilated eye exam, hemoglobin A1c test, lipid profile, and nephropathy screening.

ACSC = ambulatory care-sensitive condition; CMMI = Center for Medicare & Medicaid Innovation; ED = emergency department; FFS = fee-for-service; HCIA = Health Care Innovation Award; IVD = ischemic vascular disease.

To estimate impacts, we used a core approach across all awardees that we tailored to meet the specific circumstances for each awardee. The core approach had seven design elements. First, we estimated impacts as the differences in outcomes for Medicare FFS beneficiaries in a treatment group to beneficiaries in a comparison group who were similar to one another before the intervention began. Whenever possible, we used a difference-in-differences model, which estimated impacts as the differences in outcomes for the treatment and comparison beneficiaries during the intervention period minus the differences in outcomes for these two groups before the intervention began. For the two awardees (CSHP and PBGH) whose treatment groups we defined as those who actually enrolled in the program, it was not possible to define a pre-intervention treatment group. In these cases, we used a contemporaneous differences model, which estimated impacts as the differences between the treatment and comparison group during

the intervention period only, using regressions to adjust for any measurable differences between the two groups at baseline.

Second, we defined the treatment group to align as closely as possible to the population the awardee expected to affect. That is, we defined the treatment group as either all Medicare FFS beneficiaries attributed to a treatment practice (for interventions targeting all of a practice's patients) or as beneficiaries who enrolled in the program or who met specific eligibility criteria (for interventions enrolling specific beneficiaries).

Third, we selected one of two time units for measuring outcomes to match the awardee's expected time path for affect. Awardees that enrolled whole practices to the intervention expected impacts to grow as practices transformed over time, so we measured outcomes relative to when the practices joined the intervention. In contrast, awardees that enrolled individual beneficiaries into care management or transitional care interventions expected impacts to concentrate in some period (for example, the first six months) after a person enrolled. In these cases, we measured outcomes relative to when an individual enrolled or otherwise met specific eligibility criteria that triggered the start of the intervention for the patient.

Fourth, whenever possible, we used statistical techniques to match the treatment groups to the comparison groups, matching at the same level at which the intervention was delivered. For example, if the intervention affected whole practices, we selected comparison practices. If the intervention enrolled individual beneficiaries, we selected comparison beneficiaries. In both cases, we matched on variables that could affect the likelihood of being selected for the intervention, the study outcomes, or both. For interventions affecting whole practices, we matched on characteristics of the practice, including its size, provider composition, and the average service use and health status of its Medicare patients. For interventions enrolling individual beneficiaries, we matched on beneficiary characteristics, including demographics, service use, and medical conditions.

Fifth, for each awardee, we used the program's theory of action to prespecify a limited number of primary tests—that is, the tests for which we most strongly expected to find evidence of impacts on patients if the program was indeed effective. The awardees had an opportunity to review and comment on these primary tests before we estimated impacts. Each primary test specified an outcome, population, time period, and expected direction of effect (that is, positive or negative), as well as the threshold that we considered substantively important. The substantive thresholds enabled us to identify (1) estimates that, although not statistically significant, were nonetheless promising because they were in the favorable direction and exceeded the threshold; and (2) programs that might have had unintended effects because the estimates were in the unfavorable direction and larger than the threshold. We used one-sided statistical tests (testing for evidence of favorable effects) and a threshold for statistical significance of $p < 0.10$. This reflects the evaluation's goal to identify promising programs or program components (which could be retested later), not only those with definitive evidence of impacts.

Sixth, we drew conclusions about program impacts in each of the four evaluation domains based on the results of the primary tests. When making conclusions, we considered the results of

a set of secondary tests (including robustness and regression model checks) and the consistency of the impact findings with implementation evidence. We planned to draw one of five conclusions within each domain:

1. Statistically significant favorable effect (the highest level of evidence)
2. Substantively important (but not statistically significant) favorable effect
3. Substantively important (but not statistically significant) unfavorable effect
4. No substantively large effect—that is, no evidence of substantively large effects despite good statistical power (at least 75 percent) to detect effects if they existed
5. Indeterminate effect—that is, no evidence of substantively large effects, but only poor or marginal statistical power (less than 75 percent) to detect them

We could not conclude that a program had a statistically significant *unfavorable* effect because, in consultation with CMMI, we decided to use one-sided statistical tests (which do not test for evidence of unfavorable effects). We used one-sided tests to increase the probability that, if a program truly did have impacts, we would be able to detect them. Moreover, in some cases (described in the next section), we were unable to draw any conclusions because the primary test results conflicted with findings from the robustness checks or the implementation evaluation evidence.

Seventh, after drawing impact conclusions, we further used the implementation evidence to highlight the core features of those interventions with evidence of favorable program impacts and the implementation contexts of these programs. We believe this operational and contextual detail could help guide future efforts to redesign primary care to improve quality while reducing medical spending. For awardees with no evidence of program impacts, we used the first two evaluation components (implementation and clinicians' behavior) to assess whether lack of patient-level impacts might be due to either (1) challenges to implementing the intervention as planned or (2) an inability to change clinicians' behavior as anticipated.

III. SUMMARY OF IMPACT CONCLUSIONS ACROSS AWARDEES

The impact conclusions varied substantially across the awardees and across the four outcome domains (Table 2).

1. **Four awardees showed statistically significant improvements in quality-of-care processes.** Three of these awardees—FLHSA, PeaceHealth, and Sanford Health—aimed to transform the way practices as a whole delivered care; however, the specific interventions varied substantially (Section IV). The fourth awardee, AGH, provided a transitional care intervention, focusing almost exclusively (at least in the program component we evaluated) on the transition from hospital to home among recently discharged beneficiaries. The specific care processes that the awardees improved varied across the awardees, reflecting the different focuses for their interventions. PeaceHealth and Sanford both improved diabetes care and AGH and FLHSA improved the percentage of people who received timely ambulatory care visit after being discharged from the hospital. Although we did not find that

other awardees improved the processes of care, this might be, in part, because our claims-based measures could not capture the full diversity of care processes that awardees aimed to improve.

2. **Only one awardee, CSHP, measurably improved quality-of-care outcomes.** CSHP provided care-intensive care management services to people with unusually complex medical and social needs. CSHP reduced rates of 30-day unplanned readmissions by an estimated 34 percent. For two awardees, CareFirst and AGH, we estimated that the program had unfavorable effects on quality-of-care outcomes. In both cases, substantively large increases in 30-day unplanned readmissions drove this unfavorable conclusion.
3. **Three awardees measurably reduced service use.** Two of these awardees delivered practice transformation interventions (Sanford Health and TransforMED) and the third, AGH, provided transitional care services. A reduction in outpatient ED visits drove the improvements for Sanford Health; reductions in outpatient ED visits and inpatient admissions drove those for AGH and TransforMED.
4. **Only one awardee, AGH, measurably reduced Medicare FFS spending.** AGH reduced total Medicare spending by an estimated 31 percent, or $1,333 per beneficiary per month.

We did not find evidence of effects—favorable or unfavorable—for any of the other awardees or outcome domains. In many instances, the tests had sufficient statistical power to detect substantively large effects, so the results indicate that the program likely had no large effects (Table 2 lists this conclusion as "No substantively large effect"). For example, our tests were generally well powered to detect substantively large effects on quality-of-care processes. However, in other cases, the statistical power to detect effects was not good (less than 75 percent), so we might not have seen effects either because (1) the program truly did not have effects; or (2) it did, but our tests did not detect them (Table 2 lists these conclusions as "Indeterminate"). This was particularly true for Medicare spending. None of the tests for spending were well powered to detect substantively large effects, in part due to the large variation in spending across Medicare beneficiaries.

Finally, in some cases we could not draw impact conclusions because the tests did not pass prespecified robustness checks, were inconsistent with implementation findings, or both (Table 2 lists these as "No conclusion"). Specifically, we were unable to draw conclusions in any domain for two awardees (PBGH and WIPH), and in one or more domains for two other awardees (Denver Health and PeaceHealth).

Table 2. Summary of impact conclusions across 10 HCIA-PCR awardees

Awardee	Intervention type (for component(s) included in the impact evaluation)	Impact conclusion, by domain				Conclusions are preliminary or final[a]
		Quality-of-care processes	Quality-of-care outcomes	Service use	Spending	
AGH	Transitional care	**Statistically significant favorable effect**	**Substantively important unfavorable effect**	**Statistically significant favorable effect**	**Statistically significant favorable effect**	Final
CareFirst	Practice transformation	No substantively large effect	**Substantively important unfavorable effect**	No substantively large effect	Indeterminate effect	Preliminary
CSHP	Care management for high-risk patients	Indeterminate effect	**Statistically significant favorable effect**	Indeterminate effect	Indeterminate effect	Final
Denver Health	Practice transformation	No substantively large effect	Indeterminate effect	No conclusion	Indeterminate effect	Final
FLHSA	Practice transformation	**Statistically significant favorable effect**	Indeterminate effect	No substantively large effect	Indeterminate effect	Preliminary
PBGH	Care management for high-risk patients	No conclusion	No conclusion	No conclusion	No conclusion	Final
PeaceHealth	Practice transformation	**Statistically significant favorable effect**	No conclusion	No conclusion	No conclusion	Final
Sanford Health	Practice transformation	**Statistically significant favorable effect**	No substantively large effect	**Statistically significant favorable effect**	Indeterminate effect	Final
TransforMED	Practice transformation	No substantively large effect	Not applicable	**Statistically significant favorable effect**	No substantively large effect	Final
WIPH	Practice transformation	No conclusion	No conclusion	No conclusion	No conclusion	Final

Source: Impact analyses using Medicare FFS claims data, as presented in individual report chapters.

Notes: We drew impact conclusions at the domain level. Section II describes the five possible conclusions we could draw. In some cases, we were unable to draw any conclusions (label "No conclusion" in the table) because the primary test results conflicted with findings from robustness checks or the implementation evaluation.

The outcomes included in each domain varied by awardee. We selected a set of possible outcomes in each domain for the evaluation as a whole based on available Medicare claims data and CMMI's priorities. Then, for each awardee, we selected outcomes from among that set based on the outcomes the awardee expected to affect (see Tables 3 and 4).

[a] Conclusions are preliminary for the two awardees (CareFirst and FLHSA) that received no-cost extensions to continue to provide interventions services after the original 3-year award period. An addendum to this report will include those extension months in the impact estimates and will draw final impact conclusions. For all other awardees in this table, the impact estimates are final because the award periods were not extended.

AGH = Atlantic General Hospital; CareFirst = CareFirst Blue Cross Blue Shield; CSHP = Rutgers Center for State Health Policy; Denver Health = Denver Health and Hospital Authority; FFS = fee-for-service; FLHSA = Finger Lakes Health Systems Agency; HCIA = Health Care Innovation Award; PBGH = Pacific Business Group on Health; PCR = primary care redesign; PeaceHealth = PeaceHealth Ketchikan Medical Center; WIPH = Wyoming Institute for Population Health.

IV. INDIVIDUAL AWARDEE FINDINGS, GROUPED BY IMPACT CONCLUSIONS

This section describes key impact and implementation findings for each awardee individually, grouped by the overall impact assessment. We first describe awardees that improved quality-of-care, service use, or spending outcomes, because these domains include CMMI's priority measures (Section A and Table 3). Next, we describe awardees that improved quality-of-care process but not outcomes in the other domains (Section B and Table 4). We then describe awardees that did not measurably improve outcomes in any domain (Section C and Table 5). Finally, in Section D, we describe awardees for which we could not draw any impact conclusions, including explanations for why these conclusions were not possible.

A. Awardees with favorable impacts for quality-of-care outcomes, service use, or spending

1. Atlantic General Hospital

AGH, a rural health care system in eastern Maryland with a 62-bed hospital, implemented care coordination and transitional care programs at the hospital and in its seven affiliated primary care practices. However, we estimated impacts only for the transitional care component due to difficulty constructing a credible comparison group for the care coordination component. Under the transitional care intervention, a nurse used hospital records to identify currently hospitalized patients who had an AGH primary care provider (PCP). The nurse then contacted patients by telephone within 72 hours of discharge and at least weekly thereafter for 30 days to review discharge instructions, schedule recommended office visits, and monitor patients' adherence to medications and treatment plans. AGH expected that its overall intervention would reduce (1) hospital admission rates by 20.0 percent, (2) ED visits by 20.0 percent, and (3) total cost of care by 15.5 percent. However, AGH did not set explicit targets for the transitional care component alone.

AGH largely implemented the transitional care intervention as planned, with the full-time care transitions care coordinator managing a full caseload ranging from 40 to 50 patients throughout the intervention. Of the 1,002 patients enrolled in the program, 90 percent participated for the full 30 days.

We found statistically significant favorable impacts on Medicare Part A and B spending—with average savings of $1,333 per beneficiary per month—and on service use (with the latter driven by a combined 14.7 percent decrease in outpatient ED visits and inpatient admissions). We further estimated that the program had a statistically significant favorable impact on quality-of-care processes (driven by an increase in the proportion of patients receiving ambulatory follow-up care within 14 days of discharge), but a substantively important *unfavorable* impact on quality-of-care outcomes (as measured by the 30-day unplanned readmission rate). This apparently unfavorable impact might have occurred if the nurse quickly identified patients who had to be readmitted, shifting readmissions into the 30-day window that otherwise would have occurred later.

Table 3. Intervention descriptions and impact results for awardees that improved quality-of-care outcomes, reduced service use, and/or reduced Medicare spending

	Atlantic General Hospital	CSHP	Sanford Health	TransforMED
Intervention type	Transitional care	Care management for high-risk patients	Practice transformation	Practice transformation
Awardee description	Health system with 62-bed hospital and 7 primary care practices	Research group at Rutgers University that guided implementation at four program sites	Large integrated health system serving 100 communities in 9 states	National learning and dissemination contractor that closed in 2015 (subsidiary of AAFP)
Award extended beyond June 2015?	No	No	No	No
Award amount	$1.1 million	$14.3 million	$12.1 million	$20.8 million
Location(s)	Eastern Maryland and southern Delaware (rural)	High-poverty areas in four cities[a]	Minnesota, North Dakota, and South Dakota (urban, suburban, and rural)	Multistate (urban, suburban, and rural)
Description of intervention (for component[s] included in impact evaluation)				
Target population	All patients with an AGH PCP discharged from AGH	Frequent users of hospital services (inpatient or outpatient ED)	All patients served by 33 of Sanford Health's practices, focusing on patients with at least 1 of 8 targeted conditions[b]	All patients served by 90 primary care practices that were part of 15 health systems
Intervention(s)	Transitional care for 30 days after discharge[c] • Nurse called patients within 72 hours of discharge; weekly thereafter • Nurse scheduled office visits for urgent needs; monitored patients' adherence to treatment plan, including medications	Care management to address patients' medical, behavioral, and social needs • Delivered by multidisciplinary care teams • Teams scheduled medical appointments and provided transportation • Patients coached on physician visits and self-management • Patients linked to social and behavioral health services (for example, SSDI benefits, substance abuse treatment centers)	Integrating behavioral health into primary care • Screenings for behavioral health conditions • Short-term counseling and/or referrals Care management for medical conditions • Patients coached on self-management skills • Symptoms and progress monitored Expanded health IT to support other award components	Health IT to help practices function as part of a patient-centered medical neighborhood • Software for managing health of patient panel and identifying cost drivers • Technical assistance (learning collaboratives and monthly calls) to use new health IT

Table 3 *(continued)*

		Atlantic General Hospital	CSHP	Sanford Health	TransforMED
Metrics of intervention delivered		• Enrolled 1,002 people (all insurance types) • 90 percent of patients participated for full 30-day intervention period	• Enrolled 1,068 people (all insurance types) • Among enrolled patients: - 10 contacts per month on average for 4.2 months - 66 percent met care goals and graduated[d]	• 290 staff members helped implement intervention • Hired 18 behavioral health triage therapists • Increased share of patients identified with depression from 13 to 17 percent and with anxiety from 10 to 14 percent	• 78 of 90 practices implemented population health management software • 96 percent of practices identified a health coach to serve as an expert for population management in each practice
Impact evaluation methods					
Core design		Difference-in-differences model with matched comparison group	Contemporaneous differences model with matched comparison group	Difference-in-differences model with matched comparison group	Difference-in-differences model with matched comparison group
Treatment group	Definition	Medicare FFS beneficiaries with an AGH PCP discharged from AGH	Medicare FFS beneficiaries who enrolled in the program	Medicare FFS beneficiaries attributed to 22 nonpediatric participating practices with baseline data	Medicare FFS beneficiaries attributed to 87 participating practices
	# of beneficiaries across quarters in the primary test period[e]	376 to 638	113 to 149	12,950 to 18,238	93,213 to 97,994
Comparison group definition		Matched Medicare FFS beneficiaries discharged from a nearby comparison hospital, or from AGH with a non-AGH PCP	Matched Medicare FFS beneficiaries living in same or similar geographic areas as treatment beneficiaries	Medicare FFS beneficiaries attributed to 91 matched comparison practices	Medicare FFS beneficiaries attributed to 286 matched comparison practices
Impact results: quality-of-care processes domain					
Ambulatory care visit within 14 days of discharge (% of beneficiaries/quarter)	Comparison mean[f]	67.6%	37.4%	62.3%	61.2%
	Impact estimate (% difference)	+5.9 pp (+8.8%)*	+3.6 pp (+9.7%)	+<0.1 pp (+0.1%)	+0.8 pp (+1.3%)

Table 3 *(continued)*

		Atlantic General Hospital	CSHP	Sanford Health	TransforMED
Received lipid test, for patients with IVD (% of beneficiaries/year)	Comparison mean[f]	n.a.	n.a.	n.a.	75.1%
	Impact estimate (% difference)	n.a.	n.a.	n.a.	+1.4 pp (+1.9%)
Received all four recommended diabetes processes of care (% of beneficiaries/year)[g]	Comparison mean[f]	n.a.	n.a.	44.7%	44.6%
	Impact estimate (% difference)	n.a.	n.a.	+3.8 pp (+8.6%)**	+0.5 pp (+1.2%)
Combined impact estimate[h]		n.a.	n.a.	+4.3%**	+1.5%
Impact conclusion		**Statistically significant favorable effect**	Indeterminate effect	**Statistically significant favorable effect**	No substantively large effect
Impact results: quality-of-care outcomes domain					
30-day unplanned hospital readmissions (#/1,000 beneficiaries/quarter, unless specified)	Comparison mean[f]	9.8%	365	10.9	n.a.
	Impact estimate (% difference)	+1.9 pp[i] (+18.9%)	-126 (-34.4%)*	-0.1 (-1.3%)	n.a.
Inpatient admissions for ACSCs (#/1,000 beneficiaries/quarter)	Comparison mean[f]	n.a.	215	12.7	n.a.
	Impact estimate (% difference)	n.a.	-27 (-12.4%)	+1.7 (+13.6%)	n.a.
Combined impact		n.a.	-23.4%**	+6.2%	n.a.
Impact conclusion		**Substantively important unfavorable effect**	**Statistically significant favorable effect**	No substantively large effect	n.a.
Impact results: service use domain					
All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	Comparison mean[f]	301	784	82.5	82.6
	Impact estimate (% difference)	-72 (-23.9%)	-116 (-14.8%)	+1.5 (+1.8%)	-5.8 (-7.1)

Table 3 *(continued)*

		Atlantic General Hospital	CSHP	Sanford Health	TransforMED
Outpatient ED visits (#/1,000 beneficiaries/quarter)	Comparison mean[f]	344	1,196	138.9	144.7
	Impact estimate (% difference)	-19 (-5.5%)	+57 (+4.8%)	-6.8 (-4.9%)*	-8.2 (-5.7)
Combined impact estimate[h]		-14.7%*	-5.0%	-1.6%	-5.5**[j]
Impact conclusion		**Statistically significant favorable effect**	Indeterminate effect	**Statistically significant favorable effect**	**Statistically significant favorable effect**
Impact results: spending domain					
Medicare Part A and B spending ($/beneficiary/month)	Comparison mean[f]	$4,325	$5,332	$898	$910
	Impact estimate (% difference)	-$1,333 (-30.8%)***	-$468 (-8.8%)	+$13 (+1.5%)	-$10 (-1.1%)[j]
Combined impact estimate[h]		n.a.	-5.0%	n.a.	0.40%[j]
Impact conclusion		**Statistically significant favorable effect**	Indeterminate effect	Indeterminate effect	No substantively large effect

Source: Chapters I, VII, VIII, and IX of this report.

Note: We drew impact conclusions at the domain level. Section II of this executive summary describes the possible impact conclusions and Appendix 3 of this report describes in detail the decision rules we used to draw impact conclusions. Appendix 1 describes how we calculated each of the study outcomes.

[a] Allentown, Pennsylvania; Aurora, Colorado; Kansas City, Missouri, San Diego, California.

[b] Asthma, anxiety, depression, diabetes, heart disease, hypertension, obesity, and substance abuse (alcohol and drug abuse).

[c] AGH's intervention included a second component—care coordination for people with chronic conditions—that we did not include in the impact evaluation.

[d] Graduated means that both program staff and the patient agreed the patient had met his or her care goals.

[e] For some outcome measures the sample is limited to a relevant subset of beneficiaries.

[f] The comparison mean is the estimate of the outcome the treatment group beneficiaries would have had if they had not received the intervention. It is equal to the mean for the treatment group over the intervention quarters (in the primary test period) minus the impact estimate.

[g] Lipid profile, hemoglobin A1c test, dilated eye exam, and nephropathy screening.

[h] The combined estimate is the average across all the individual estimates in each domain, in which the impact estimate for each individual outcome is expressed as percentage change relative to the comparison group.

[i] For AGH, we measured the percentage of people who were readmitted within 30 days of the index stay that qualified them for the treatment or comparison group.

[j] We also conducted primary tests for all-cause inpatient admissions, outpatient ED visits, and spending for a high-risk subset of the full sample (results in Chapter IX). The combined impact estimates in the service use and spending domains combine the estimates for the full and the high-risk populations.

* \ ** \ *** Significantly different from zero at the .10\ .05\ .01 levels, respectively (one-tailed test).

AAFP = American Academy of Family Physicians; ACSC = ambulatory care-sensitive condition; AGH = Atlantic General Hospital; CSHP = Center for State Health Policy at Rutgers University; ED = emergency department; FFS = fee-for-service; IT = information technology; IVD = ischemic vascular disease; PCP = primary care provider; pp = percentage point; SSDI = Social Security Disability Insurance

n.a. = not applicable.

These results, especially those for spending, are intriguing because the intervention was modest, consisting mainly of several telephone calls from the nurse to patients. The results suggest that a brief, well-timed transitional care intervention can have substantial impacts on patients' outcomes. However, two factors might influence the generalizability of the AGH findings to other settings. First, because the hospital and practices share a parent organization, the care transitions nurse coordinator (1) had timely, complete discharge information from the hospital with which to identify the target population within 72 hours of discharge; and (2) could offer potential program participants follow-up care in a familiar, recognizable setting—their existing primary care practices. Second, the hospital is located in Maryland, which, in 2014, required all payers to pay hospitals global budgets, regardless of the number of patients they saw, creating strong incentives for hospitals to reduce the number of admissions. This payment approach prompted strong leadership commitment to the patient-centered medical home (PCMH) model of care delivery and facilitated implementation of the care transitions program. It is unclear whether the program could be implemented as actually was and have the same effects in other regulatory environments.

2. Rutgers Center for State Health Policy

CSHP implemented a care management program in high-poverty areas in four cities: San Diego, California; Aurora, Colorado; Kansas City, Missouri; and Allentown, Pennsylvania. The program was designed to test whether a care management model for high-risk patients originally created by the Camden Coalition of Healthcare Providers could be successfully adapted and implemented in other settings. Multidisciplinary, community-based care teams aimed to connect frequent users of hospital services (so called high utilizers) to appropriate clinical and social services, help them manage their conditions, and overcome socioeconomic obstacles to care. Care teams addressed patients' medical, behavioral, and social needs, including securing primary care and specialist appointments, linking patients with social services, and educating patients to improve their capacity to manage their conditions. CSHP aimed to reduce average annual costs of care by 14.8 percent by the end of the award by reducing patients' use of inappropriate acute care—such as inpatient admissions and ED visits—and by increasing use of appropriate primary and specialty care.

All sites implemented the program largely as planned. Patients stayed in the program for an average of 4.2 months. Two-thirds of patients who enrolled graduated, meaning that they met the goals described in the care plans. However, the other third dropped out before meeting their goals, either because they moved out of the catchment area, became unreachable by care team staff, declined to participate further, or died.

We found statistically significant favorable impacts on quality-of-care outcomes, driven by a 34.4 percent reduction in the number of 30-day unplanned hospital readmissions. The intervention did not measurably improve outcomes in the three other outcome domains (quality-of-care processes, service use, and Medicare spending). However, because the statistical power to detect substantively large effects for these domains was poor to marginal (less than 60 percent for each outcome), it is possible the program had effects that we did not detect.

Because four different provider groups in four geographic areas implemented this intervention, our findings suggest that programs such as CSHP's might be broadly replicable and improve quality-of-care outcomes among patients with unusually complex needs in diverse settings. CSHP gave its four implementing sites flexibility in how to define their target populations and which bundle of care management services to include as part of the intervention, a decision that program staff said facilitated implementation. However, all four sites stayed true to the model's core design of enrolling patients considered to be at high risk of needing acute care. For example, on average, treatment group beneficiaries experienced 2.7 hospital admissions in the six months before program enrollment, more than 18 times the national Medicare FFS average. Therefore, to replicate these findings in other settings, programs would likely have to maintain this strong focus on high utilizers.

However, we found no evidence that the intervention succeeded in its goal of reducing spending in addition to improving quality-of-care outcomes. The lack of observed effects on Medicare spending—and on outcomes in the quality-of-care process and service use domains—could be due to three factors. First, we had insufficient statistical power to detect effects, as described previously. Second, CSHP experienced challenges sustaining long-term behavioral change in a patient population with complex medical and social needs. Third, the local health and social service systems likely lacked the effective resources that the program was designed to leverage. In addition, effects could have been concentrated among Medicaid and uninsured populations, which our impact estimates did not include.

3. Sanford Health

Sanford Health, a large integrated health system, implemented the One Care program, which consisted of a medical home intervention for 33 of its practices in Minnesota, North Dakota, and South Dakota. The intervention included (1) care management services, provided by nurse health coaches, to patients with asthma, diabetes, heart failure, hypertension, or obesity; (2) integration of behavioral health into primary care through screenings for behavioral health conditions, short-term counseling from newly hired behavioral health therapists (up to six sessions), or referrals for longer-term counseling; and (3) expanded health IT to support these interventions, for example through disease registries to identify patients with targeted conditions and track patients' receipt of recommended care processes. By the end of the award, Sanford Health aimed to reduce potentially preventable admission and outpatient ED visit rates by 20 percent and reduce total cost of care by 3 percent for Medicare, Medicaid, and CHIP beneficiaries with targeted conditions.

Sanford Health implemented the intervention largely as intended, engaging 290 staff members through training and intervention delivery. Further, the program appears to have largely engaged PCPs as planned. About three-quarters of surveyed clinicians said they were aware of the One Care program, and most said they believed that the program improved the patient-centeredness and quality of care they delivered.

The impact estimates indicate that the program improved quality-of-care processes and reduced service use. Specifically, the program increased the percentage of people with diabetes who received recommended care by 8.6 percent and reduced outpatient ED visits by 4.9 percent

(both estimates were statistically significant). However, there was no evidence that the program reduced inpatient admissions, improved quality-of-care outcomes, or reduced Medicare spending. The statistical power to detect substantively large effects on spending was marginal, so the program might have had effects that went undetected. The lack of measured effects on inpatient admissions and quality-of-care outcomes could be due to insufficient intensity or duration of nurse health coaching, or because other outcomes (such as outpatient ED visits) are easier to influence than others.

Several factors could influence the generalizability of Sanford Health's favorable impacts for quality-of-care processes and service use. First, many participating practices already had nurse health coaches who could play an expanded role under the intervention. Second, as a large integrated health system, Sanford Health's practices used a common health IT platform, which facilitated implementing new IT functions, such as disease registries and online screening tools. This integrated internal health IT system was critical to the integration of care management and behavioral health care. Finally, the participating practices served areas with shortages in behavioral health professionals, enhancing the value of providing behavioral health services within primary care.

4. TransforMED

TransforMED was a learning and dissemination contractor that was a subsidiary of the American Academy of Family Physicians, which closed in 2015. For its HCIA-funded intervention, TransforMED delivered health IT software and technical assistance to 90 primary care practices in 15 health systems to help them develop into patient-centered medical neighborhoods (PCMNs)—a variant of the PCMH concept. TransforMED selected practices that had used electronic health records (EHRs) for at least a year and had leadership and staff motivated to transform their practices in ways that the new health IT system supported. The intervention provided (1) population health management and cost-reporting software to practices to promote the use of data to improve clinical processes and (2) technical assistance to practices and health systems to use the software effectively. The health IT systems helped practices identify gaps in clinical care for their patients, develop care plans for their high-risk patients, and identify service use patterns that drove high costs. TransforMED expected that the practices would use these systems to identify and reach out to patients needing preventive care, improve care management for high-risk patients, and improve the coordination across providers. This, in turn, would reduce inpatient admissions and outpatient ED visits, and reduce redundant or unnecessary services, lowering total cost of care by 4 percent by the end of the award.

TransforMED implemented the population health management and technical assistance components of the intervention largely as planned. However, several difficulties prevented practices from using the cost-reporting software as intended, including technical challenges and data lags when generating reports, and financial competition between the convening health system and nonsystem practices. In addition, survey data suggest clinicians might not have been fully engaged throughout the award. Almost two-thirds (64 percent) of clinicians reported familiarity with the HCIA program and, among them, fewer than half reported that they believed the program improved patients' care.

We found statistically significant favorable impacts on service use, driven by a 7.1 percent reduction in the inpatient admission rate and a 5.7 percent decrease in the outpatient ED visit rate. The intervention did not measurably improve quality-of-care process measures, although this might be because different practices aimed to improve different processes, and our claims-based measures did not capture some of these. The intervention did not measurably reduce Medicare spending, even though the evaluation was well powered to detect substantively large effects in this domain. The lack of effects on spending is surprising because the program reduced inpatient spending, which accounts for a large share of total spending among Medicare beneficiaries. The savings from reduced inpatient admissions might have been partially offset by increases in outpatient spending due to greater use of primary care and other ambulatory services because of the intervention. We did not estimate the impact of the TransforMED intervention on the two measures in the quality-of-care outcomes domain because they were not part of the awardee's theory of action.

Because TransforMED implemented its intervention in different settings across multiple health systems, the favorable impacts on service use could be broadly generalizable. However, the practices shared common features that might restrict the types of practices for which these results could be expected. TransforMED selected the 90 participating practices based, in part, on their commitment to quality improvement, existing use of health IT, and ability to accommodate new software. These factors might help explain why health systems and practices were interested in, and largely able to, integrate the standard features of the new health IT systems into their workflows. For example, practices had to be using EHRs to use the new population health IT software because that software pulled data from the practice's EHR. Therefore, the favorable findings might be replicable in other practices that share similar levels of motivation and capacity to use the new health IT resources to transform their care.

B. Awardees with favorable impacts for quality-of-care processes, but not for other outcome domains

1. Finger Lakes Health Systems Agency

FLHSA, a regional community health planning and convening agency, implemented a PCMH intervention in 68 practices in the greater Rochester, New York, area. The intervention included (1) FLHSA practice improvement advisors working with practice champions and other practice staff to redesign primary care processes, culture, and workforce to transform practices into PCMHs; and (2) HCIA-funded care managers delivering intensive care management to high-risk Medicare and Medicaid beneficiaries. FLHSA aimed to reduce the total cost of care by 3 percent by improving intermediate health outcomes and quality of care for all patients—particularly high-risk Medicare and Medicaid beneficiaries—thus reducing potentially preventable hospital admissions, hospital readmissions, and avoidable ED visits. FLHSA also worked with payers in the region to develop a communitywide outcomes-based payment model to sustain the interventions after the award period ended. However, because the impact estimates in this report cover the original award period only, the payment model element of the intervention should not influence the impact estimates.

The intervention was delivered largely as intended. Practice champions at each practice spearheaded the transformation initiatives and all 68 participating practices successfully hired care managers to provide targeted, intensive care management. Further, FLHSA appears to have engaged clinicians largely, although not completely, as planned. By July 2015, all practices held weekly huddles and most used EHRs to identify gaps in care for their patients. More than 84 percent of PCPs surveyed were aware of the intervention, and slightly more than half thought the intervention improved the quality and patient-centeredness of care at their practices.

We found a statistically significant favorable impact on quality-of-care processes, driven by a 4.6 percent increase in the percentage of inpatient beneficiaries with a follow-up ambulatory care visit within 14 days of discharge (Table 4). The program did not measurably improve outcomes in the other three outcome domains. However, we did not have good statistical power in the quality-of-care outcomes and spending domains. Thus, it is possible the program had effects in these domains, but our tests did not detect them.

The modest improvement in quality-of-care processes is encouraging, though stakeholders should consider the extent to which other settings can replicate these favorable findings. FLHSA used a competitive application process to select practices that were highly motivated to undergo improvement efforts, particularly to become PCMHs, and that already used health IT systems to guide care. Therefore, these modest favorable impacts might generalize to other practices that are similarly committed to transformation and have IT systems in place that can facilitate this transformation.

The lack of measured effects in the other domains (quality-of-care outcomes, service use, and spending) could be due to one or more of four factors. First, practice champions and care managers reported that they had limited time to devote to practice transformation and intensive care management activities, respectively. Second, FLHSA's care management intervention might have been limited by its scope—only about 2 percent of all patients received care management services. Third, as indicated by the awardee's self-monitoring metrics, some of the practices already conducted key practices supported by the intervention (such as weekly huddles), meaning there was less opportunity for the intervention to improve performance in these areas, reducing the marginal impact of the interventions. Finally, although we set the primary test period to coincide with periods when the awardee expected effects, it is possible that program impacts take longer than anticipated to accrue. The final analysis, to be included in a future addendum to this report, will include an additional 12 months that FLHSA's award was extended beyond the original three-year award period.

Table 4. Intervention descriptions and impact results for awardees with favorable impacts for quality-of-care processes but not the other outcome domains

	FLHSA	PeaceHealth
Intervention type	Practice transformation	Practice transformation
Awardee description	Community health planning and convening organization in Rochester, New York	Medical center (with a 25-bed critical access hospital) and two affiliated primary care clinics
Award amount	$26.6 million	$3.2 million
Award extended beyond June 2015?	Yes (12 months)	No
Location(s)	6 counties in greater Rochester area[a] (urban, suburban, and rural)	Remote island communities in southeastern Alaska
Description of intervention (for component[s] included in impact evaluation)		
Target population	All patients served by 68 primary care practices, which enrolled in the intervention in three cohorts	All patients served by 2 primary care clinics, with some intervention components targeted to specific patients within those clinics
Intervention(s)	Identified care gaps among the full patient population at participating practices and developed care plans for high-risk patients • 5 HCIA-funded practice improvement advisors helped practices improve team communication, use EHRs to identify care gaps, and streamline workflows • PCPs were each paid $20,000 to participate in the intervention • 70 care managers hired to (1) coach high-needs patients on self-management, (2) coordinate care with providers, and (3) connect patients with social services	Conducted population health and disease management activities through several program components: • Transitional care, in which nurses (1) called each patient (once only) to review discharge instructions and medications, and assess need for further support; and (2) made additional calls to patients with CHF to assess signs of excess fluid and encourage follow-up with a PCP • Individualized care management for patients with specific conditions,[b] provided by 6 HCIA-funded nurses and a social worker • Expanded use of population health IT and scrub-and-huddle process[c]
Metrics of intervention delivered	• Weekly huddles at all practices by June 2015 • Care managers hired at all practices • Care manager services provided to 17,484 patients	• 12,600 direct encounters with 3,500 unique patients • 60 to 80 percent of targeted patients (depending on month) received transitional care

Table 4 *(continued)*

		FLHSA	PeaceHealth
Impact evaluation methods			
Core design		Difference-in-differences model with matched comparison group	Difference-in-differences model with comparison group (unmatched)[d]
Treatment group	Definition	Medicare FFS beneficiaries attributed to 37 practices FLHSA enrolled by July 1, 2013[e]	Medicare FFS beneficiaries attributed to 2 PeaceHealth treatment clinics
	# of beneficiaries across quarters during the primary test period[f]	9,271 to 15,638	996 to 1,101
Comparison group definition		Medicare FFS beneficiaries attributed to 108 matched comparison practices	Medicare FFS beneficiaries attributed to 57 (unmatched) comparison practices
Impact results: quality-of-care processes domain			
Ambulatory care visit within 14 days of discharge (% of beneficiaries/quarter)	Comparison mean[g]	67.6%	40.8%
	Impact estimate (% difference)	+3.1 pp (+4.6%)*	-14.7 pp (-36.0%)
Received lipid test, for patients with IVD (% of beneficiaries/year)	Comparison mean[g]	76.4	n.a.
	Impact estimate (% difference)	-0.6 pp (-0.7%)	n.a.
Received all four recommended diabetes processes of care (% of beneficiaries/year)[h]	Comparison mean[g]	NA[i]	20.1%
	Impact estimate (% difference)	n.a.[i]	+11.5 pp (+57.2%)**
Combined impact estimate[j]		+1.9%**	-5.4%[k]
Impact conclusion		**Statistically significant favorable effect**	**Statistically significant favorable effect**
Impact results: quality-of-care outcomes domain			
30-day unplanned hospital readmissions (#/1,000 beneficiaries/quarter)	Comparison mean[g]	14.3	n.a.
	Impact estimate (% difference)	+0.1 (0.7%)	n.a.

Table 4 *(continued)*

		FLHSA	PeaceHealth
Inpatient admissions for ACSCs (#/1,000 beneficiaries/quarter)	Comparison mean[g]	16.0	n.a.
	Impact estimate (% difference)	+0.3 (+1.6%)	n.a.
Combined impact estimate[j]		+3.7%[l]	n.a.
Impact conclusion		Indeterminate effect	No conclusion[m]
Impact results: service use domain			
All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	Comparison mean[g]	83	n.a.
	Impact estimate (% difference)	+3.1 (+3.7%)	n.a.
Outpatient ED visits (#/1,000 beneficiaries/quarter)	Comparison mean[g]	173.3	n.a.
	Impact estimate (% difference)	-3.5 (-2.0%)	n.a.
Combined impact estimate[j]		+0.6%	n.a.
Impact conclusion		No substantively large effect	No conclusion[m]
Impact results: spending domain			
Medicare Part A and B spending ($/beneficiary/month)	Comparison mean[g]	$825	n.a.
	Impact estimate (% difference)	+$11 (+1.3%)	n.a.
Combined impact conclusion[j]		0.8%	n.a.
Impact conclusion		Indeterminate effect	No conclusion[m]

Source: Chapters IV and V of this report.

Note: We drew impact conclusions at the domain level. Section II of this executive summary describes the possible impact conclusions and Appendix 3 of this report describes in detail the decision rules we used to draw impact conclusions. Appendix 1 describes how we calculated each of the study outcomes.

[a] Livingston, Monroe, Ontario, Seneca, Wayne, and Yates.

[b] Program staff initially targeted those with CHF or diabetes, then expanded to those with hypertension and high-risk pregnancies.

[c] Scrubbing involved reviewing a patient's medical records to identify outstanding care needs, such as laboratory tests, mammograms, immunizations, or colorectal screenings. The huddling process involved a team meeting to review a patient's needs before a regularly scheduled visit.

[d] The comparison group was unmatched because statistical matching did not meaningfully improve balance on prespecified matching variables relative to the full pool of potential comparison practices. We relied on the difference-in-differences model to account for any differences in outcomes that stemmed from persistent (time-invariant) differences between the treatment and comparison practices.

[e] Our impact evaluation covers 37 practices that enrolled in the intervention in the first two cohorts of participating practices. We excluded Cohort 3 practices because they joined late in the award period and neither we nor the awardee expected the program to affect patients' outcomes during the original 3-year award period. We will include Cohort 3 practices in our future final impact analyses.

Table 4 *(continued)*

[f] For some outcome measures the sample is limited to a relevant subset of beneficiaries.

[g] The comparison mean is the estimate of the outcome the treatment group beneficiaries would have had if they had not received the intervention. It is equal to the mean for the treatment group over the intervention quarters (in the primary test period) minus the impact estimate.

[h] Dilated eye exam, hemoglobin A1c test, lipid profile, and nephropathy screening

[i] We did not estimate impacts on receipt of all four recommended diabetes processes of care because FLHSA did not target all of these measures. Instead,we focused on the two processes FLHSA did target: HbA1c tests and lipid profiles. For both measures, the treatment group's outcomes were 1 to 3 percentage points higher than the comparison group's, but the differences were not statistically significant.

[j] The combined estimate is the average across all the individual estimates in each domain, in which the impact estimate for each individual outcome is expressed as percentage change relative to the comparison group.

[k] The combined impact includes the two estimates in this table plus one test not shown here (14-day ambulatory care follow-up visits for Medicare FFS beneficiaries with CHF, diabetes, and/or hypertension) but that is reported in Chapter VI.

[l] FLHSA's combined impact estimate for the quality-of-care outcomes domain comprises the estimates of two measures in this table (30-day unplanned readmissions and ACSC admissions) and two measures not reported in this table (30-day unplanned readmissions and ACSC admissions among only high-risk beneficiaries) but that are reported in the full chapter for FLHSA.

[m] We are unable to draw impact conclusions in three of the four study domains for reasons described in Chapter V.

* \ ** \ *** Significantly different from zero at the .10\ .05\ .01 levels, respectively (one-tailed test).

ACSC = ambulatory care-sensitive condition; CHF = congestive heart failure; ED = emergency department; FFS = fee-for-service; FLHSA=Finger Lakes Health Systems Agency; HCIA = Health Care Innovation Award; IVD = ischemic vascular disease; PCP = primary care provider; PCMH = patient-centered medical home; pp = percentage point.

NA = not available.

n.a. = not applicable.

2. PeaceHealth Ketchikan Medical Center

PeaceHealth, a medical center with a 25-bed critical access hospital, implemented a coordinated care program in its two affiliated primary care practices in remote island communities in southeastern Alaska. The program included (1) transitional care services for patients discharged from the PeaceHealth hospital, with particular emphasis on those with congestive heart failure (CHF); (2) care management (with varying duration, depending on the patient's need) for patients with chronic medical conditions; and (3) population health management, including improved scrub-and-huddle and outreach activities to improve preventive care. Over three years, PeaceHealth expected to reduce 30-day hospital readmission rates for patients with CHF by 20 percent, ED costs for patients with chronic conditions by 75 percent, and total costs for patients with chronic conditions by 15 percent.

According to interviews with frontline staff, PeaceHealth implemented the program largely as planned. Further, metrics from the awardee indicate that care coordinators contacted 60 to 80 percent of all patients discharged from the PeaceHealth hospital. However, metrics for the other intervention components were unavailable. In surveys, the small number of clinicians at the two participating practices reported that the program had a positive impact on quality of care, their ability to respond in a timely way to patients' needs, patients' safety, and the patient-centeredness of care they provided.

We can draw conclusions on program impacts in the quality-of-care processes domain only. We found statistically significant improvement in processes-of-care, driven solely by a 12 percentage point (or 57 percent) increase in the percentage of patients with diabetes who received all four recommended diabetes process-of-care measures. This improvement is consistent with the intervention's focus on improving diabetes care, including care coordinators' efforts to contact patients with diabetes who were overdue for an appointment and to use the scrub-and-huddle process to make sure routine tests were conducted before arrival. The program might also have achieved these large effects due to the low percentage of beneficiaries with diabetes who had received the recommended process-of-care measures before the intervention began (18 percent). These large, favorable results therefore might generalize only to other settings with similar gaps in care and to practices that put the same emphasis on improving diabetes care.

We are unable to draw conclusions in the quality-of-care outcomes, service use, and spending domains. For these outcomes, the robustness checks indicated that the treatment group's outcomes differed (in both favorable and unfavorable directions) from the comparison group's in the six months after the intervention began, when systems were being developed and few or no impacts were expected. These checks, together with the fact that the treatment and comparison practices differed on several important dimensions (such as size and likelihood of being owned by a hospital), raise concerns that observed differences during the primary test period could be due to limitations in the comparison group for these outcomes and do not represent true impacts.

C. Awardees without favorable impact estimates in any outcome domain

1. CareFirst Blue Cross Blue Shield

CareFirst, the largest commercial health insurer in the mid-Atlantic region, used its HCIA to extend an existing PCMH program designed for its commercial members to Medicare FFS beneficiaries in Maryland. The program targeted about 35,000 Medicare beneficiaries served by 52 primary care practices, which were grouped into 14 medical panels (the performance unit for the commercial PCMH program and the HCIA intervention). The program had three components: (1) care coordination for high-risk, clinically unstable patients; (2) financial incentives to medical panels for participating in care coordination and achieving savings and quality targets among their Medicare patients; and (3) technical assistance to medical panels to identify opportunities for reducing spending through changing referral patterns or shifting treatment to more cost-effective settings. CareFirst aimed to reduce total Medicare costs by 6 percent in the final intervention year by reducing patients' need for hospitalizations and ED visits and by encouraging PCPs to refer patients to lower-cost specialists and care settings.

After a one-year delay, the intervention was implemented largely as planned. CareFirst hired 44 nurse care coordinators, enrolled 3,276 beneficiaries into care coordination (which included roughly weekly nurse contact for an average of 260 days), provided ongoing technical assistance to panels, and paid financial incentives (called outcome incentive awards) to panels. Further, awardee data indicate that CareFirst engaged PCPs as planned, with 90 percent of PCPs enrolling at least one patient into care coordination services. Most (67 to 78 percent) PCPs reported they thought the intervention improved the quality, timeliness, and safety of the care they provided to patients.

During the original three-year award period, the program did not measurably improve outcomes in any of the four outcome domains. The lack of measured effects might be due to (1) challenges in identifying patients who were both at elevated risk of acute care service use and clinically unstable (therefore most able to benefit from care coordination services); (2) challenges in adapting care coordination strategies from commercial to Medicare populations; (3) limitations in the intervention design (such as challenges in reducing spending for all of a panel's patients when services are targeted to a small [about 10 percent] percentage of all patients); (4) modest statistical power to detect substantively large effects in two of the study domains (quality-of-care processes and spending); and (5) the relatively short intervention duration caused by implementation delays. Impact estimates might change after we include the final six months of program operations, the period when CareFirst expected to observe the largest impacts. We will report these findings in an addendum to this report.

Table 5. Intervention descriptions and impact results for awardees without favorable impacts estimates in any domain

	CareFirst	Denver Health
Intervention type	Practice transformation	Practice transformation
Awardee description	Largest commercial health insurer in the mid-Atlantic region	Integrated safety-net health system; largest provider to Medicaid and uninsured patients in Colorado
Award amount	$20.0 million	$19.8 million
Award extended beyond June 2015?	Yes (6 months)	No
Location(s)	Maryland, statewide (urban and suburban)	Denver, Colorado (urban)
Description of intervention (for component[s] included in impact evaluation)		
Target population	Approximately 35,000 Medicare FFS beneficiaries (excluding those also enrolled in Medicaid) served by 149 PCPs in 52 primary care practices grouped into 14 medical panels	All patients (about 250,000) meeting one of the following criteria: • Served by Denver Health's 8 FQHCs • In Denver Health's managed care plan • Used Denver Health's hospital or ED frequently
Intervention(s)	Extended a PCMH program developed for commercial members to Medicare FFS beneficiaries. The program included: • Care coordination, in which 44 HCIA-funded nurses worked with PCPs to develop and implement care plans for high-risk patients • Financial incentives to (1) reward panels that reduced total spending while meeting quality targets; and (2) pay PCPs to participate in care coordination • Technical assistance to panels to identify opportunities to generate savings though changes in referrals	Stratified patients into 4 risk tiers and, within those tiers, into clinically similar groups, to triage to other intervention services • Text message reminders about appointments • Enhanced primary care teams in 8 FQHCs, incorporating 23 HCIA-funded patient navigators and 3 clinical pharmacists • High-risk clinics that offered longer and more comprehensive appointments than typically covered by insurance
Metrics of intervention delivered	• Implemented care plans for 3,276 beneficiaries (almost 10 percent of panels' Medicare patients) • Nurses contacted patients in care plans roughly weekly for an average of 260 days • Paid panels $3,000 to $494,000 in 2015 to reward them because total Medicare spending for their Medicare patients was below projections	• 79,000 contacts with patient navigators • 19,000 contacts with clinical pharmacists • Text message reminders to 28,000 patients, with an average of 8 messages per person • High-risk clinics at capacity at intervention end

Table 5 *(continued)*

		CareFirst	Denver Health
Impact evaluation methods			
Core design		Difference-in-differences model with matched comparison group	Difference-in-differences model with comparison group (unmatched)[a]
Treatment group	Definition	Medicare FFS beneficiaries attributed to 14 treatment panels	Medicare FFS beneficiaries attributed to one of 8 FQHCs by the intervention start
	# of beneficiaries across quarters during the primary test period[b]	35,536 to 37,593	2,317 to 3,746
Comparison group definition		Medicare FFS beneficiaries attributed to 42 matched comparison panels participating in CareFirst's commercial PCMH program	Medicare FFS beneficiaries attributed to 15 comparison FQHCs by the start of the intervention
Impact results: quality-of-care processes domain			
Ambulatory care visit within 14 days of discharge (% of beneficiaries/ quarter)	Comparison mean[c]	66.0%	50.5%
	Impact estimate (% difference)	+0.3 pp (+0.4%)	+0.6 pp (+1.1%)
Received lipid test, for patients with IVD (% of beneficiaries/ year)	Comparison mean[c]	80.0%	n.a.
	Impact estimate (% difference)	-0.8 pp (-1.0%)	n.a.
Received all four recommended diabetes processes of care (% of beneficiaries/year)[d]	Comparison mean[c]	48.5%	n.a.
	Impact estimate (% difference)	-2.8 pp (-5.7%)	n.a.
Combined impact estimate[e]		-2.1%	n.a.
Impact conclusion		No substantively large effect	No substantively large effect
Impact results: quality-of-care outcomes domain			
30-day unplanned hospital readmissions (#/1,000 beneficiaries/ quarter)	Comparison mean[c]	8.6	15.1
	Impact estimate (% difference)	+1.3 (+16.3%)	+0.9 (+6.1%)

Table 5 *(continued)*

		CareFirst	Denver Health
Inpatient admissions for ACSCs (#/1,000 beneficiaries/quarter)	Comparison mean[c]	11.2	9.3
	Impact estimate (% difference)	+0.4 (+3.7%)	+0.3 (+3.3%)
Combined impact estimate[e]		+10.0%	+4.7%
Impact conclusion		**Substantively large unfavorable effect**	Indeterminate effect
Impact results: service use domain			
All-cause inpatient admissions (#/1,000 beneficiaries/ quarter)	Comparison mean[c]	70.9	n.a.
	Impact estimate (% difference)	+1.9 (+2.6%)	n.a.
Outpatient ED visits (#/1,000 beneficiaries/ quarter)	Comparison mean[c]	85.5	n.a.
	Impact estimate (% difference)	-2.6 (-3.1%)	n.a.
Combined impact estimate[e]		-0.2%	n.a.
Impact conclusion		No substantively large effect	No conclusion[f,g]
Impact results: spending domain			
Medicare Part A and B spending ($/beneficiary/ month)	Comparison mean[c]	$1,005	$948
	Impact estimate (% difference)	+$9 (+0.9%)	+$8 (+0.9%)
Combined impact estimate[e]		NA	+0.4%[g]
Impact conclusion		Indeterminate effect	Indeterminate effect

Source: Chapters II and III of this report.

Note: We drew impact conclusions at the domain level. Section II of this executive summary describes the possible impact conclusions and Appendix 3 of this report describes in detail the decision rules we used to draw impact conclusions. Appendix 1 describes how we calculated each of the study outcomes.

[a] The comparison group was unmatched because statistical matching did not meaningfully improve balance on prespecified matching variables relative to the full pool of potential comparison practices. We relied on the difference-in-differences model to account for any differences in outcomes that stemmed from persistent (time-invariant) differences between the treatment and comparison practices.

[b] For some outcome measures the sample is limited to a relevant subset of beneficiaries.

[c] The comparison mean is the estimate of the outcome the treatment group beneficiaries would have had if they had not received the intervention. It is equal to the mean for the treatment group over the intervention quarters (in the primary test period) minus the impact estimate.

[d] Dilated eye exam, hemoglobin A1c test, lipid profile, and and nephropathy screening.

[e] The combined estimate is the average across all the individual estimates in each domain, in which the impact estimate for each individual outcome is expressed as percentage change relative to the comparison group.

Table 5 *(continued)*

[f] We were unable to draw impact conclusions because the impact findings were not plausible given the implementation evidence. Specifically, there is no plausible explanation for why the program would have increased outpatient ED visit rates by the 14.2 percent suggested by the primary tests.

[g] The combined measure is the average of point estimates from two overlapping time periods specified in the primary tests (that is, the 5th through 11th intervention quarters and the 8th through 11th intervention quarters) as described in Chapter III..

ACSC = Ambulatory-care sensitive condition; CareFirst = CareFirst BlueCross BlueShield; Denver Health = Denver Health and Hospital Authority; ED = emergency department; EHR = electronic health record; FFS = fee-for-service; FQHC = federally qualified health center; HCIA = Health Care Innovation Award; IVD = ischemic vascular disease; PCP = primary care provider; PCMH = patient-centered medical home; pp = percentage point.

NA = not available.

n.a. = not applicable.

2. Denver Health and Hospital Authority

Denver Health, an integrated safety-net health system in Denver, Colorado, (1) developed a risk-stratification algorithm to group its roughly 130,000 patients into four risk tiers, based on their anticipated medical needs; (2) reorganized primary care delivery by incorporating newly hired support staff (such as patient navigators and clinical pharmacists) into its eight federally qualified health centers and creating three specialized clinics to provide intensive outpatient care to high-risk patients who met prespecified utilization or diagnostic criteria; and (3) upgraded its health IT infrastructure, enabling text messages to be sent to patients for appointment reminders. Denver Health expected these components would reduce the need for (and inappropriate use of) acute care services, reducing the overall cost of care by 2.5 percent by the end of its three-year award.

Denver Health implemented all program components largely as planned. In addition, survey data indicate Denver Health engaged PCPs as planned, as more than 90 percent of PCPs reported providing team-based care (consistent with the intervention model) and about 80 percent of respondents reported they thought the intervention improved the quality, timeliness, and patient-centeredness of care they provided to patients.

The treatment group for the impact analysis (Medicare FFS beneficiaries) comprised fewer than 5 percent of Denver Health's total target population—most of whom were covered by Medicaid or Medicare Advantage or had no insurance. For the Medicare FFS population, the impact estimates indicate largely indeterminate effects. We found no evidence of statistically significant or substantively large differences between the treatment and comparison groups in the quality-of-care processes, quality-of-care outcomes, and spending domains. However, for two of these domains—quality-of-care outcomes and spending—we had poor statistical power to detect effects. This means we cannot be sure whether the intervention truly had no effects in these domains, or whether it did have effects and our evaluation failed to detect them. We did not draw conclusions in the service use domain because we considered the primary test results implausible given the implementation evidence. Overall, it is difficult to generalize from the Denver Health impact results because (1) low statistical power renders our estimates imprecise and (2) Medicare FFS beneficiaries comprise a very small proportion of Denver Health's target population. The effects we report for Medicare FFS beneficiaries could differ from the effects for most patients participating in Denver Health's intervention.

D. Awardees for which impact conclusions were not possible in any domain

1. Pacific Business Group on Health

PBGH, a nonprofit coalition of businesses and public organizations that purchase health insurance for their employees, partnered with 23 physician medical groups to implement a care management program in five states: Arizona, California, Idaho, Nevada, and Washington State. The program targeted FFS and managed care Medicare beneficiaries who met several criteria (for example, having three or more hospitalizations in the previous six months) designed to indicate high risk of future acute care use. The intervention embedded care managers in primary care practices. These care managers assessed patients' needs, developed shared action plans with patients, and met frequently with them (about monthly) for about a year to educate patients on

self-care, connect patients with social and medical services, monitor patients' health, and alert clinicians of changes in health status that warranted changes in medications or treatment plans. By improving patients' self-care and clinical care, PBGH aimed to reduce the need for acute and post-acute care services, reducing admissions and total cost of care by 5 percent.

Due to delays and complexities in processing claims data, PBGH was unable to use claims to identify beneficiaries for enrollment as initially planned. As a result, it developed subjective approaches for identifying potential enrollees, such as clinicians' judgment and referral, and reduced its enrollment target from 27,000 to 15,000 participants. PBGH met this revised target and generally provided services as planned. However, physicians' engagement in the program was lower than expected, with only 30 percent of providers we surveyed aware of the program.

We were unable to draw conclusions about program impacts in any domain. Although the treatment and comparison groups were well matched at baseline, the outcomes for the treatment group were consistently worse than those for the comparison group during the intervention period. None of our implementation evidence suggested that the program had unfavorable effects in all domains. Rather, these differences likely stem from unobserved differences at baseline between the treatment and comparison groups that affected patients' outcomes during the intervention period. For example, because PBGH decided not to identify potential participants using claims data (as initially planned), clinicians made subjective decisions about whom to recruit for enrollment. We were unable to replicate these decisions in claims data, nor could we control for the possible selection bias that could occur when only some recruited patients voluntarily enrolled.

2. Wyoming Institute for Population Health

WIPH used its HCIA to implement a five-component program designed to transform care delivery in rural Wyoming. We focused our evaluation on the component that WIPH considered the centerpiece of its intervention—a PCMH intervention implemented at 20 primary care practices. Under this component, WIPH hired TransforMED (separately from that awardee's program) to instruct the 20 self-selecting practices on PCMH concepts and help them apply for PCMH recognition from the National Committee for Quality Assurance (NCQA). The PCMH component included (1) holding quarterly learning collaboratives, (2) conducting site visits and telephone calls with practices, (3) helping practices develop customized transformation plans, and (4) reviewing practices' PCMH application documents before submission to NCQA. WIPH intended that these practices would transform primary care by—among other things—increasing patients' access to care (for example, through evening hours or access to care managers), improving providers' adherence to clinical guidelines, improving care coordination among providers, and improving support for patients' self-management of chronic conditions. WIPH expected that these improvements would in turn reduce the need for acute care, reducing ED visits by 10 percent, hospitalizations by 5 percent, and total costs by 5 percent by the end of the award.

There is insufficient evidence to indicate whether WIPH implemented the intervention as planned. Although TransforMED led eight quarterly learning collaboratives, we have limited information about which practices participated in learning collaboratives, developed work plans,

or received application review services, and about the intensity and content of participating practices' interactions with TransforMED. Of the 20 participating practices, 10 eventually received NCQA certification. However, because WIPH selected practices that were already motivated to become, and in some cases were in the process of becoming, certified PCMHs, it is unclear to what extent the intervention contributed to these 10 practices achieving certification. In surveys, only 38 or 47 percent (depending on the round) of clinicians indicated that the program had a positive impact on the quality of the care they provided to patients, with the remainder saying the program had no impact or it was too soon to tell.

We attempted to estimate program impacts using a difference-in-differences model with a matched comparison group of 75 practices in Montana. We selected comparison practices from Montana, rather than regions of Wyoming similar to those of the treatment practices, because a large proportion of Wyoming practices were already involved in the WIPH intervention. The remaining pool of practices in Wyoming not participating in the intervention was small, and those practices—by virtue of choosing not to join the intervention—might have differed from the intervention practices in systematic but unobservable ways. However, robustness and model checks from our quantitative analyses (that is, the secondary tests) suggested that the comparison group in Montana did not provide an adequate estimate of the counterfactual for the treatment practices. One possibility is that a voluntary medical home program that Montana launched statewide in 2014 contributed to improvements in outcomes for the comparison practices that did not represent what would have happened for the treatment practices (in Wyoming) absent the intervention. In addition, we found unfavorable results from the primary tests that were implausibly large given the relatively modest scope of the PCMH intervention. For these reasons, we are unable to draw conclusions about program impacts.

V. CONCLUSIONS

The HCIAs intentionally tested a range of interventions with the goal of identifying promising interventions that merit expansion, further testing, or incorporation into other primary care reforms. The HCIAs differed from most CMMI models because the innovations were developed from the bottom up by the awardees themselves and, therefore, serve as a complementary source of innovative ideas to CMS's other model tests that are more centrally designed and administered. Given the range of models that HCIA-PCR tested, it is reasonable to expect that some programs would be promising in terms of having their intended effects, but others would not. A key goal of this evaluation is to identify those that are promising based on the available data.

Across the 8 PCR awardees (of 10 in this report) in which impact conclusions were possible, we found no clear pattern in the types of programs that had favorable effects. Several awardees improved quality-of-care processes or reduced service use (outpatient ED visits, inpatient admissions, or both), with fewer awardees measurably improving quality-of-care outcomes or reducing Medicare spending. The awardees with favorable effects in at least one outcome domain ranged from a transitional care intervention implemented in a single hospital in rural Maryland (AGH) to a health IT-based practice transformation effort (TransforMED) implemented in 90 practices across 15 states.

In this section, we draw two types of implications from these evaluation's findings. First, we discuss implications for future evaluations of primary care redesign, drawing lessons from the strengths and challenges we encountered estimating impacts for the HCIA-PCR portfolio of interventions examined for this report. Second, we discuss implications for future efforts to redesign primary care in ways that improve quality of care while reducing overall medical spending.

A. Implications for future impact evaluations in primary care redesign

Robust evaluation is critical for understanding models that are likely to achieve CMMI's aims. Our findings suggest several possibilities for improving future tests of care delivery models through refinements to the interventions themselves, selection of intervention participants, or evaluation design.

- **Reduce the possibility for confounding by increasing the number and diversity of treatment units.** Impact evaluation designs with a large number of treatment practices, particularly if they were located in different markets (such as TransforMED's), facilitated impact conclusions because they helped ensure that the HCIA-funded intervention was not confounded with other activities that were unique to the participating practices or their markets. In contrast, confounding was a significant risk for an awardee such as PeaceHealth, for which the number of treatment practices (two) was small, and the practices were unique in their locations and populations served.
- **Select proposed interventions that use (and adhere to) enrollment criteria that can be replicated in available data; if that is not possible, consider random assignment.** By replicating the treatment selection criteria, evaluators limit the likelihood that the treatment and comparison groups differ in ways—including those that might be unobservable—that affect outcomes but are unrelated to the intervention. For CareFirst, we could replicate key selection criteria by obtaining data on medical panels' performance in the commercial PCMH program, which CareFirst used to select panels for the HCIA intervention. In contrast, for PBGH, we could not replicate selection because the awardee relied heavily on providers' clinical judgment—which we could not mimic in claims or other available data. Because of this, unobservable factors (that is, factors not apparent in available data) likely drove patient and provider selection into the program, biasing impact estimates. To improve future tests of interventions like PBGH's, program administrators or evaluators could consider (1) having the intervention use measurable criteria, replicable in claims or other available data, to determine whom to enroll; or (2) if provider or patient selection is critical to the intervention design, randomly assigning beneficiaries within the target population to treatment and control groups to ensure they do not differ in systematic ways.
- **Encourage program administrators to set explicit targets and timelines for intervention implementation, and to measure progress.** The PCR awardees varied in the explicitness of their intervention protocols, targets for number of participants enrolled and services provided, and in the detail of their metrics about what their programs delivered. Although this was understandable in some cases, given some awardees' intentional flexibility to let their program designs evolve, the lack of targets sometimes made it difficult

to assess whether the programs delivered what was intended. This, in turn, made it hard to determine whether a lack of measured impacts was due to failures implementing the program as planned. Although some awardees collected detailed metrics, which enabled us to clarify exactly what intervention the impact results tested (regardless of whether the intervention delivered fit the original design), in other cases metrics were sparse and hindered how much we could learn from the impact findings.

- **Determine decision rules for drawing impact conclusions before analyzing results.** One strength of our evaluation has been the decision framework we used for drawing conclusions. Because we analyzed a large number of outcomes for each awardee, often with poor or marginal statistical power, it has been helpful to have clear methods for calling a program effective or ineffective. These decision rules not only prevented us from "chasing noise" in the data; they also provided a clear framework for communicating and vetting our hypotheses with the awardees and CMMI before estimating impacts. In the future, researchers who conduct evaluations that face similar challenges of small samples and a large number of outcomes of interest might wish to consider a similar framework.

Finally, CMS's current efforts to improve the timeliness of Medicaid data should help to improve future impact evaluations. Several PCR awardees targeted their interventions primarily to Medicaid beneficiaries. We were unable to include Medicaid beneficiaries (who were not also enrolled in Medicare) in the impact evaluation because the Medicaid claims data available did not cover any, or only a few months, of the intervention period. More current data will enable future evaluations to incorporate Medicaid and CHIP beneficiaries into the impact estimates and therefore increase the share of that target population captured by the treatment group, increasing both the representativeness of the treatment group and the statistical power to detect true program impacts.

B. Implications for future efforts to redesign the delivery of primary care

This evaluation identified six distinct interventions that show promise to improve the quality of care, reduce the need for acute care services, and/or reduce Medicare spending. Only one model (AGH), a short-term transitional care intervention delivered to beneficiaries recently discharged from the hospital, measurably improved quality-of-care processes while reducing service use and total Medicare spending. Three other models—(1) providing health IT and technical assistance to support practice transformation (TransforMED), (2) care management for frequent users of acute care services (CSHP), and (3) practice transformation emphasizing the integration of behavioral health and primary care (Sanford Health)—reduced service use or improved quality-of-care outcomes, without measurably reducing Medicare spending. Finally, two additional models of practice transformation—one in upstate New York (FLHSA) and the other in a frontier region of southeastern Alaska (PeaceHealth), improved quality-of-care processes, although with no evidence of improving outcomes in other domains.

The results from this evaluation have several implications for future efforts to redesign primary care delivery systems.

- **There is no single path to success in primary care redesign.** Across the HCIA-PCR portfolio, a range of intervention types had favorable effects in one or more outcome domains. This suggests that a range of interventions can be effective, and improved outcomes will depend on the specific context, delivery, and target population of the intervention.
- **The impacts from primary care redesign efforts could be modest.** The programs we identified as having impacts in quality-of-care outcomes, service use, or spending were typically effective in one or some domains, but not all—and only one intervention had measurable impacts on spending. For awardees that reduced service use (ED visits, inpatient admissions, or both) but not overall spending, it is possible that the program increased outpatient spending that at least partially offset reductions in spending on acute care. Further, for many awardees, the magnitude of the estimated impacts over the three-year intervention period was small to moderate: for example, improving the percentage of people receiving recommended care processes by less than 10 percent or reducing the outpatient ED visit rate by less than 5 percent. It is possible that, for some programs (especially those that received extensions to continue their interventions beyond the original three-year period), impacts will grow over time. However, our results suggest that core model designs for some awardees did not have their anticipated impacts. One possible explanation for why some awardees had small or no effects is that the health care environment is changing rapidly and beneficiaries in the comparison group might be receiving services that overlap, to some degree, with those provided by the HCIA interventions—limiting the ability of the intervention services to improve outcomes beyond those achieved in the comparison groups.
- **The specific context of model implementation might influence the generalizability of favorable findings.** One strength of the HCIAs is the awardees' ability to tailor their target populations and interventions to their organizations' distinct cultures and capacities and the needs of their particular patient populations. However, this targeting means that favorable estimates observed in one setting might not generalize to other settings. This makes it important to consider the internal and external factors that contribute to each program's success and might have to be identified, or fostered, in other settings before implementing the intervention elsewhere. For example, the Sanford Health program might have successfully reduced service use, in part, because the practices are located in areas with shortages of mental health professionals, increasing the value of its efforts to integrate behavioral health services into primary care.
- **There could be value in retesting successful interventions, in larger applications or different settings, before scaling them broadly.** Even interventions with statistically significant findings would likely benefit from retesting to (1) confirm that impacts are real, given that—across the many domains tested—some large favorable differences could have arisen due to chance; (2) assess programs' impacts in domains in which this evaluation's findings are indeterminate due to low statistical power; and (3) assess whether favorable results can be replicated in a broader (or different) setting.

Overall, the results in this report help CMMI meet its goal of assessing the many primary care redesign models tested by HCIA to identify those that are promising and those that are

not—at least over the time horizon measured. A future supplement to this report will expand the time period for some of the tested models and will increase the number of models with impact findings. Specifically, we plan to expand the primary test period by 6 to 12 months for the two awardees (CareFirst and FLHSA) included in this report that received no-cost extensions to continue their interventions beyond the original three-year award period. Further, we plan to include new impact results for three awardees, made possible by newly available Medicaid data or larger sample sizes due to an extended enrollment period. These three awardees tested, among other interventions, medical home services for children enrolled in Medicaid (University Hospitals of Cleveland Rainbow Babies & Children's Hospital), peer counseling and care coordination for Medicaid children with behavioral health needs (Research Institute at Nationwide Children's Hospital), and care management for very high-risk Medicare and Medicaid beneficiaries (Cooper University Hospital and the Camden Coalition of Healthcare Providers).

This page has been left blank for double-sided copying.

REFERENCE

Moreno, Lorenzo, Boyd Gilman, Greg Peterson, Catherine DesRoches, Sheila Hoag, Linda Barterian, Laura Blue, Katherine Bradley, Emily Ehrlich, Kristin Geonnotti, Lauren Hula, Keith Kranker, Rumin Sarwar, Rachel Shapiro, KeriAnn Wells, Joseph Zickafoose, Sandi Nelson, Frank Yoon with the Implementation Team, Impact Team, Data Processing Team, Surveys Team, and Production Coordination and Editorial Team. "Evaluation of the Health Care Innovation Awards (HCIAs): Primary Care Redesign Programs. Second Annual Report, Volumes I and II." Princeton, NJ: Mathematica Policy Research, March 2016. Available at https://downloads.cms.gov/files/cmmi/hcia-primarycareredesignprog-secondannualrpt.pdf. Accessed September 12, 2016.

This page has been left blank for double-sided copying.

CHAPTER 1

ATLANTIC GENERAL HOSPITAL

Keith Kranker, Linda Barterian, Rumin Sarwar, Greg Peterson, Boyd Gilman, Laura Blue, Kate Stewart, Sheila Hoag, and Lorenzo Moreno

This page has been left blank for double-sided copying.

ATLANTIC GENERAL HOSPITAL

CHAPTER SUMMARY

Introduction. Atlantic General Hospital (AGH), a rural health care system with a 62-bed hospital, received a $1.1 million Health Care Innovation Award (HCIA) to implement a patient-centered medical home (PCMH) intervention at the hospital and its seven primary care practices in partnership with the Worcester County Health Department (WCHD) of Maryland.

Objectives. This report describes and estimates the impacts of one key component of the intervention—care transitions from hospital to home—which ran from February 2012 to June 2015. We (1) describe the design and implementation of the intervention component; (2) assess impacts of the intervention on patient outcomes and Medicare Part A and B spending during the award, and (3) use implementation and impact findings to identify possible explanations for the observed impacts.

Methods. We reviewed AGH's program documents and self-monitoring metrics and conducted site visits and interviews with AGH leadership and program staff. We used a difference-in-differences design with a matched comparison group to estimate the impacts of the intervention on Medicare fee-for-service (FFS) beneficiaries. Using claims data, impact estimates measured the differences in post-discharge outcomes between the patients who had an AGH primary care provider (PCP) and were discharged from AGH during the intervention period (N = 638) and matched comparison beneficiaries, minus the differences in post-discharge outcomes between AGH patients with an AGH PCP discharged in a one-year period before the intervention began and comparison beneficiaries. The comparison group included beneficiaries discharged from AGH who did not have an AGH PCP or who were discharged from a nearby comparison hospital. The comparison beneficiaries were well matched to treatment group beneficiaries on demographics, health status, chronic conditions, reason for the hospitalization leading to eligibility for enrollment, and service use and spending one year before discharge.

Program design and implementation. A nurse care coordinator monitored the hospital's daily census to identify all admissions for patients with an AGH PCP and notified the PCP of the admission through the hospital's electronic medical record (EMR) system. The care coordinator collected information from the hospital's EMR on reasons for hospital stay, recent primary care visits, and discharge instructions. Within 72 hours of discharge, the care coordinator contacted the patients by telephone and scheduled a PCP follow-up visit. The care coordinator then contacted patients weekly for 30 days, monitoring and encouraging adherence to treatment plans, reconciling medications, and referring patients to home care or further PCP visits as needed. AGH's care transitions program was implemented as planned without major delays. AGH was able to deliver services to the target population as intended and achieve a steady increase in enrollment and low opt-out rate. One care coordinator worked for the care transitions program full-time and managed a caseload ranging from 40 to 50 patients. Other AGH administrative staff provided data support for the care coordinator. Finally, AGH completed training in the PCMH model for AGH PCPs and staff and added supplementary training in motivational interviewing to enhance care coordinator effectiveness in engaging patients.

Impacts on patient outcomes. The evidence indicates that the care transitions component of AGH's intervention achieved favorable impacts on patient outcomes in three of the four evaluation domains during the first six months after beneficiaries' enrollment: quality-of-care processes, service use, and spending. We estimate the intervention reduced composite service use by 14.7 percent (p = .095), which averages a 23.9 percent reduction in the inpatient admission rate and a 5.5 percent decrease in the outpatient emergency department visit rate. The intervention reduced spending by 30.8 percent (p = .002), or $1,333 per beneficiary per month. Furthermore, it increased the percentage of inpatient admissions followed by an ambulatory care visit with a PCP or specialist within 14 days by 8.8 percent (p = .097), one way AGH expected to affect admissions and spending. However, we found a large, unfavorable increase in the 30-day unplanned hospital readmission rate; this might be because the program had this effect on the outcome, or it could be due to chance because the statistical power to detect impacts for this outcome was poor.

Conclusions. The impact estimates indicate that the intervention improved patient outcomes in three of the four evaluation domains. The effects appear to be due to successful implementation of the program, including process improvements throughout the program to accommodate patient needs. Many studies have found that care transitions programs can improve patients' outcomes, but the current findings are new in illustrating that even a low-touch telephonic intervention in a small, rural health care system can be effective.

Summary of intervention and impact results for Atlantic General Hospital

Intervention description			
Awardee description		Health system with 62-bed hospital and 7 primary care practices	
Award amount ($ millions)		$1.1 million	
Award extended beyond June 2015?		No	
Locations		Eastern Maryland and southern Delaware (rural)	
Target population		All patients with an AGH PCP discharged from AGH	
Intervention component included in impact evaluation[a]		Transitional care for 30 days after discharge • Delivered by nurse care coordinator (by telephone) • Initial call within 72 hours; contact at least weekly thereafter • Care coordinator scheduled office visits for urgent needs; monitored patients' adherence to treatment plan, including medications	
Metrics of intervention delivered		• Enrolled 1,002 people (all insurance types) • 90% participated for full 30-day period	
Impact evaluation methods			
Core design		Difference-in-differences model with matched comparison group	
Treatment group	Definition	Medicare FFS beneficiaries with an AGH PCP discharged from AGH	
	# of beneficiaries during primary test period[b]	376 to 638	
Comparison group definition		Matched Medicare FFS beneficiaries discharged from a nearby comparison hospital, or from AGH with a non-AGH PCP	
Impact results: Quality-of-care processes domain			
Ambulatory care visit within 14 days of discharge (% of beneficiaries/quarter)[c]		Comparison mean[d]	67.6%
		Impact estimate (% difference)	+5.9 pp (+8.8%)*
Impact conclusion[e]		**Statistically significant favorable effect**	
Impact results: Quality-of-care outcomes domain			
30-day unplanned hospital readmissions (% of beneficiaries/quarter)[f]		Comparison mean[d]	9.8%
		Impact estimate (% difference)	+1.9 pp (+18.9%)
Impact conclusion[e]		**Substantively important unfavorable effect**	
Impact results: Service use domain			
All-cause inpatient admissions (#/1,000 beneficiaries/quarter)		Comparison mean[d]	301
		Impact estimate (% difference)	-72 (-23.9%)
Outpatient ED visits (#/1,000 beneficiaries/quarter)		Comparison mean[d]	344
		Impact estimate (% difference)	-19 (-5.5%)
Combined impact estimate[g]		-14.7%*	
Impact conclusion[e]		**Statistically significant favorable effect**	
Impact results: Spending domain			
Medicare Part A and B spending ($/beneficiary/month)		Comparison mean[d]	$4,325
		Impact estimate (% difference)	-$1,333 (-30.8%)***
Impact conclusion[e]		**Statistically significant favorable effect**	

Note: See the Atlantic General Hospital chapter for details on the intervention, impact methods and impact results.

[a] AGH's program had two primary components that were not included in the impact evaluation: (1) care coordination for participants with chronic conditions and (2) less-intensive support for participants discharged from care transitions or care coordination.

[b] Number of beneficiaries in the full treatment group across the quarters in the primary test period.

[c] Percentage of people who had an ambulatory care visit with a primary care or specialist provider within 14 days of the index stay that qualified them for the treatment or comparison group.

[d] The comparison mean is the estimate of the outcome the treatment group beneficiaries would have had if they had not received the intervention. It is equal to the mean for the treatment group over the intervention quarters (in the primary test period) minus the impact estimate.

[e] We drew conclusions at the domain level based on the results of pre-specified primary tests, secondary tests (robustness checks), and consistency with implementation evidence. For each domain, we could draw one of five conclusions: (1) Statistically significant favorable effect (the highest level of evidence), (2) Substantively important (but not statistically significant) favorable effect, (3) Substantively important (but not statistically significant) unfavorable effect, (4), No substantively large effect, and (5) Indeterminate effect. Section IV.A.8 of this report describes the decision rules we used to reach each of these possible conclusions.

[f] Percentage of people who were readmitted within 30 days of the index stay that qualified them for the treatment or comparison group.

[g] The combined estimate is the average across all the individual estimates in the domain, in which the impact estimate for each individual outcome is expressed as percentage change relative to the comparison group.

*/**/*** Significantly different from zero at the .10/.05/.01 levels, one-tailed test, respectively. No difference-in-differences estimates were significantly different from zero at the .05 level. We adjusted the *p*-values for the multiple (two) comparisons made within the service use domain.

AGH = Atlantic General Hospital; ED = emergency department; FFS = fee-for-service; PCP = primary care provider; pp = percentage point.

This page has been left blank for double-sided copying.

I. INTRODUCTION

This report presents findings from the evaluation of Atlantic General Hospital's (AGH's) patient-centered medical home (PCMH) initiative supported by a Health Care Innovation Award (HCIA), with a focus on program impacts on patient outcomes. Section II provides an overview of AGH's intervention and the design of the impact evaluation, which estimated impacts for one component (care transitions) of the overall PCMH intervention. Section III describes the design and implementation of the care transitions intervention, including how that intervention could be expected to affect study outcomes through changes in care transitions support services and patient behavior. Section IV describes our methods, results, and conclusions of estimating program impacts on patient outcomes in four domains: (1) quality-of-care processes, (2) quality-of-care outcomes, (3) service use, and (4) spending. Section V synthesizes the impact and implementation findings.

Unlike some other awardee reports, we do not include a section on primary care provider (PCP) perceptions of the program's impact on the care they provide to patients. This is because, as described in Section III.A.3, the theory of how the care transitions intervention could improve patient outcomes does not require PCPs to change their workflow or behavior.

II. OVERVIEW OF AGH'S HCIA-FUNDED INTERVENTION AND THE IMPACT EVALUATION

A. AGH's HCIA-funded intervention

AGH received $1.1 million in HCIA funding from the Center for Medicare & Medicaid Innovation (CMMI) to implement a PCMH model at AGH and its seven primary care practices in partnership with the Worcester County Health Department (WCHD) in Maryland. Table II.1 summarizes key features of the program. AGH's program had three primary components: (1) care coordination for participants with chronic conditions, (2) care transitions support for participants discharged from AGH with any diagnosis, and (3) less-intensive support for participants discharged from the first two components. HCIA program services began in the first quarter of 2013 and ran through June 2015 as planned.

AGH stated three objectives for its overall program: (1) reduce hospital admission rates by 20.0 percent, (2) reduce emergency department (ED) visits by 20.0 percent, and (3) reduce total cost of care by 15.5 percent (Table II.1). It also aimed to improve quality-of-care process and outcome measures (amounts not specified).

Table II.1. Summary of AGH's program and our evaluation for estimating its impacts on patient outcomes

Program description	
Award amount	$1,097,512
Award start date	June 2012
Implementation start date	January 1, 2013 (The first patient was enrolled into the care transitions component February 1, 2013.)
Award end date	June 30, 2015
Awardee description	AGH is a private, not-for-profit, community-based health care system comprised of Atlantic General Hospital, a 62-bed inpatient and outpatient facility in Berlin, Maryland, and seven primary care practices located throughout eastern Maryland and southern Delaware.
Intervention overview	AGH implemented a PCMH model at AGH and its seven primary care practices in partnership with the WCHD. The program is supported by health IT and community education and outreach.
Intervention components	1. **Care coordination.** Care coordinators contacted potential participants referred by providers to review medical conditions, assessments, goals, and care plans. Thereafter, care coordinators reviewed the participant's progress by monitoring lab results, attending the participant's office visits, and through weekly calls with participants (or two to three times a week for those with unstable conditions). Care coordinators referred participants needing additional assistance in the home to WCHD program staff to visit the participant in the home. 2. **Care transitions.** A care coordinator called potential participants within 48 to 72 hours of discharge to enroll the patient, and then called weekly during the 30-day period post discharge, increasing the frequency of calls for participants with unstable conditions. All participants were discharged after 30 days. Care coordinators referred participants needing additional assistance in the home to the care coordination program, through which they might have been referred to WCHD program staff, who visited the participant in the home. Care coordinators notified providers of participants who remained at high risk for readmission after 30 days. 3. **Keeping in Touch.** Volunteer nurses made brief weekly calls to participants to identify any emerging concerns and notified care coordinators and providers of any issues with participants' self-care.
Target population	1. **Care coordination.** Medicare beneficiaries with a primary diagnosis of COPD, CHF, or DM; expanded to others expected to benefit, such as those with other chronic conditions (for example, obesity or hypertension), or social needs or mental health issues who required assistance to adhere with medication regiments and care plans, even if non-Medicare or younger than 65. 2. **Care transitions.** All patients with an AGH PCP who were discharged from AGH with any diagnosis and any insurance type. 3. **Keeping in Touch.** Patients discharged from care coordination who required ongoing but less intensive follow-up support to manage their conditions.
Target impacts on patient outcomes	For the overall program: • Reduce hospital admission rates by 20.0 percent[a] • Reduce ED visits by 20.0 percent[a] • Reduce total cost of care by 15.5 percent[a] AGH's proposal contained no separately stated goals for the care transitions component.
Workforce development	The program was staffed with 3 care coordinators (RNs), 1 WCHD nurse and .5 FTE WCHD social worker, 1 program manager (last year only), and 1 data specialist (last year only). Although not all of the positions were funded through HCIA, the HCIA funding allowed AGH to hire new staff and to allow existing staff to focus on the program components. The Keeping in Touch component was implemented by two part-time unpaid (volunteer), retired nurses.
Location	Eastern Maryland and southern Delaware, rural

Table II.1 *(continued)*

Impact evaluation	
Core design	Difference-in-differences with matched comparison group
Treatment group	Medicare FFS beneficiaries who were discharged from AGH and were patients of an AGH PCP (for both the pre- and post-intervention cohorts)
Comparison group	Medicare FFS beneficiaries who were (1) discharged from PRMC, or (2) discharged from AGH but not patients of an AGH PCP and matched to a treatment group beneficiary (for both the pre- and post-intervention cohorts)
Intervention component(s) included in impact evaluation	Care transitions
Extent to which the treatment group reflects the awardee's target population (for the component(s) evaluated)	**Medium.** The impact evaluation consists exclusively of Medicare FFS beneficiaries, but the care transition component targeted other patients (for example, Medicare managed care, Medicaid, commercial). We used a claims-based, "intent-to-treat" framework to construct the treatment and comparison groups.
Study outcomes, by domain	1. **Quality-of-care processes.** 14-day follow-up to hospitalization 2. **Quality-of-care outcomes.** 30-day unplanned readmissions 3. **Service use.** All-cause inpatient admissions and outpatient ED visits 4. **Spending.** Medicare Part A and B spending

Source: Review of AGH reports, including their original application, operational plan, and 15 quarterly narrative reports to CMS.

[a] For Medicare patients with either a primary or admitting diagnosis of CHF, COPD, or DM.

AGH = Atlantic General Hospital; COPD = chronic obstructive pulmonary disease; CHF = congestive heart failure; DM = diabetes mellitus; ED = emergency department; FFS = fee-for-service; FTE = full-time equivalent, IT = information technology; PCMH = patient-centered medical home; PCP = primary care provider; PRMC = Peninsula Regional Medical Center; RN = registered nurse; WCHD = Worchester County Health Department.

B. Overview of impact evaluation

Our impact evaluation focused on the care transitions component of AGH's PCMH program. In consultation with CMMI, we decided not to attempt to estimate the impacts of the care coordination component of the AGH PCMH intervention. Although the care coordination component involved more extensive changes in primary care delivery than the care transitions component, the small number of practices participating in the care coordination program meant that our statistical models could not reliably detect even very large impacts. Furthermore, we could not fully replicate the process that AGH used to identify and enroll participants into the care coordination program using claims data, making it difficult to define a credible comparison group. For similar reasons, we also decided not to evaluate the smaller Keeping in Touch component of the PCMH program.

We used a difference-in-differences with matched comparison group design to estimate impacts of AGH's care transitions component. To implement the difference-in-differences framework in this report, we compared outcomes for Medicare fee-for-service (FFS) beneficiaries discharged from AGH's hospital after the program began and who met the program eligibility criteria (the post-intervention treatment group) and their matched comparison beneficiaries (the post-intervention comparison group). We adjusted for any differences in outcomes between beneficiaries discharged from AGH at least six months before the intervention began but who otherwise met the program eligibility criteria (the pre-intervention treatment group) and their matched comparison beneficiaries (the pre-intervention comparison group).

Using Medicare FFS claims data, we estimated impacts on outcomes in four domains: (1) quality-of-care processes, (2) quality-of-care outcomes, (3) service use, and (4) spending. Across the HCIA awardees in primary care redesign (PCR), we designed our impact evaluations to identify promising interventions or intervention components—consistent with evaluation goals from CMMI to find programs that could be scaled or re-tested as part of a future model test. Before conducting the analysis, we specified primary tests, describing the evidence we would need to conclude that the program was effective, and the awardee and CMMI reviewed these specifications. Each test specified a population, outcome, period, expected direction of effect, and threshold that we counted as substantively important. The purpose of these primary tests was to focus the impact evaluation on pre-specified hypotheses that would provide the most robust evidence about program effectiveness. We used the results from the primary tests and robustness checks to draw conclusions about program impacts in each of the four evaluation domains. Because we wanted to identify promising interventions, rather than only those programs with unequivocally demonstrated success, we conducted one-sided statistical tests (that is, testing only for program benefits) and used a threshold for statistical significance of 0.1, which is not as strict as the conventional standard of 0.05.

Our impact evaluation design does not capture impacts of AGH's HCIA program as a whole. There are two important reasons the evaluation did not reflect the effects of all intervention components among AGH's full HCIA target population. First, as mentioned earlier, we evaluated impacts of just one of the three main intervention components (care transitions). Second, the impact evaluation consisted exclusively of Medicare FFS beneficiaries, but the care transition component targeted other patients (for example, Medicare managed care, Medicaid, and commercial). On AGH's roster of patients enrolled in the care transitions component, 72 percent of patients had Medicare FFS listed as their type of insurance, and the remaining patients had other types of insurance.

III. PROGRAM IMPLEMENTATION

In this section, we first provide a detailed description of AGH's HCIA-funded care transitions intervention, highlighting how it evolved and its theory of action. Second, we assess the evidence on the extent to which the intervention was implemented as planned based on measures of program enrollment, service delivery, staffing, training, and timeliness. Third, we summarize the facilitators and barriers associated with implementation effectiveness.

We based our evaluation of AGH's program implementation on a review of the awardee's quarterly reports to CMMI and self-monitoring program metrics, telephone discussions and follow-up communications with program administrators, and information collected during site visit interviews with administrators and frontline staff conducted in April 2014 and April 2015. We did not verify the quality of the performance data reported by the awardee in its self-measurement and monitoring reports.

A. Program design and adaptation

1. Target population and patient identification, recruitment, and enrollment

Target population. The target population for AGH's HCIA-funded care transitions program included any patient with any diagnosis who was discharged from AGH and who had an AGH PCP. All AGH PCPs participated in the HCIA-funded PCMH.

Patient identification. The care coordinator, a nurse based in an administrative building near the hospital, monitored the hospital's daily census to identify all admissions of patients with an AGH PCP and notified the patient's PCP of the admission through the AGH electronic medical record (EMR). The care coordinator prioritized recruitment of patients who would benefit the most from care transitions services by identifying patients with elevated risk for readmission. The care coordinator determined elevated risk for readmission by reviewing discharge summaries and using the LACE index, a scoring system that predicts a patient's risk of unplanned readmission or death within 30 days after hospital discharge (Van Walraven et al. 2010). The care coordinator also identified potential participants already being monitored by another AGH program (for example, the cancer center) and removed them from the care transitions call list.

Patient recruitment and enrollment. During the first year, the care coordinator visited patients in the hospital to introduce them to the care transitions program and give them an informational brochure. Later, in-hospital visits were discontinued to avoid overburdening patients during the inpatient stay; instead, the care coordinator mailed an informational brochure to the patient's home before discharge and called the patient by telephone within 72 hours of discharge to explain the program, answer questions, and enroll the patient in the program. Potential participants might not have been enrolled if they opted out of care transitions services or if the care coordinator could not reach them by telephone after three tries.

2. Intervention description

The care transitions intervention began with the call to enroll patients in the program within 72 hours of discharge. During that call, the care coordinator talked with patients who agreed to enroll about their conditions, identified any immediate needs or barriers to self-care, and scheduled follow-up appointments with an AGH PCP. After the initial call, the care coordinator regularly communicated with participants to monitor their conditions and their compliance with treatment plans. Specifically, the care coordinator called participants weekly during the 30-day period post discharge. Participants with unstable conditions received more frequent calls during this period. The care coordinator identified these higher-risk patients during follow-up; AGH protocols did not define standard criteria that qualify a patient as unstable. Services that the care coordinator provided included scheduling office visits for urgent needs, monitoring the participant's adherence to treatment plans, and reviewing medications to ensure that the participant was taking the correct medication and dosage. In rare cases, the care coordinator contacted providers regarding urgent needs and referrals. For participants requiring additional support beyond that offered through the care transitions program (such as patients with complex medical conditions or social issues), the care coordinator referred participants to the care coordination program, which offered more intensive long-term follow-up support and in-home visits by WCHD program staff. AGH did not specify the exact criteria care transitions participants

needed to meet to be referred and enrolled in the care coordination program. Rather, the program relied on provider judgment to review referrals from the care coordinator and identify patients who would benefit from the program. Participants were enrolled in and received services from only one program component at a time: a care transitions participant who transferred to the care coordination program was disenrolled from the care transitions program.

AGH made one process improvement during the second year of the care transitions program implementation to improve outcomes. After realizing that readmission rates were highest for participants living in skilled nursing facilities, the care coordinator began developing a stronger working relationship with the facility staff. The care coordinator participated in rounds for these participants and coordinated follow-up for these patients with nursing facility staff during the 30 days after discharge from AGH.

3. Theory of action

Based on extensive review of AGH's program activities and goals, we developed a theory of action to depict the mechanisms through which program administrators expected the intervention (described in Section III.A.2) to improve quality-of-care processes and reduce hospital admissions, ED visits, and total cost of care after a patient was discharged from the hospital. The expected sequence of events is as follows:

1. **Coaching, monitoring, and scheduling office visits by a nurse care coordinator leads to improvement in patient adherence to the post-discharge treatment plan.** Services the care coordinator provides include monitoring the patient's adherence to medications (including taking the right medications and the right doses), scheduling timely visits with PCPs to make sure that the treatment regimen is appropriate and the patient is progressing as expected, recommending changes in diet or exercise, responding appropriately to early warning signs of worsening conditions, and monitoring patient adherence to any other post-discharge instructions. The impact evaluation captures one aspect of this potential improvement in patient adherence by assessing impacts on whether patients receive an ambulatory care visit within 14 days of discharge.
2. **Greater adherence to the treatment plan leads to fewer and less severe exacerbations of the patient's conditions.** This, in turn, should lower the need to go the ED or to be admitted to the hospital, particularly within 30 days of the initial discharge.
3. **If early warning signs do occur, the care coordinator should—through her frequent contact with patients—detect them and help the patient get timely outpatient care.** This can prevent an ED visit (by directing patients to the PCP instead of the ED) and, if the early symptoms are addressed quickly, prevent further exacerbations that would otherwise lead to hospitalizations.
4. **The reductions in ED visits, inpatient stays, and related post-acute care should, in turn, lower overall Medicare spending.** This would occur because inpatient stays and post-acute care account for a large share of total Medicare spending for the target population (as with Medicare FFS beneficiaries in general).

The care transitions program focused on helping participants stay connected to their PCPs by scheduling office visits and encouraging participants to attend office visits. However, it did not aim to change the workflow, behavior, or attitude of the PCP. The care transitions program could largely be expected to have its intended effects on participant outcomes even if it did not have any impact on behavior of the participant's usual providers. Therefore, unlike in some other awardee reports, this report does not include a section on PCP perceptions of the program's impact on the care they provided to participants.

4. Intervention staff and workforce development

Table III.1 provides key details about staff supporting the care transitions component of AGH's HCIA-funded PCMH program. AGH employed one registered nurse (RN) care coordinator who worked for the care transitions program full-time (the position existed before HCIA and was not supported by the award). After the first year of implementation, administrators identified a need for additional administrative support for the care transitions program and other program components. During the last year, they increased two full-time equivalents (FTEs) after hiring one person to supervise day-to-day operations and another to manage data collection and reporting. Both new hires supported all program components.

Table III.1. Care transitions intervention staff and responsibilities

Program component	Staff members	Staff/team responsibilities	Adaptations to originally planned roles?
Care transitions	Care coordinator	Assessed participants' care transitions needs, provided telephone follow-up, made participants' post-discharge follow-up appointments (position not supported by HCIA)[a] (one FTE)	No
All programs[b]	Program manager	Supervised day-to-day program operations	Yes—position added during last year of program implementation to support the clinical director[c]
All programs[b]	Data specialist	Completed mandatory program reporting, monitored outcomes	Yes—position added during last year of program implementation to support the clinical director and care coordinator[c]

Source: Interviews and document review.

[a] As of June 2015, AGH reported that the organization spent $82,956 for in-kind expenditures for staffing the program.

[b] AGH's PCMH program included three program components: (1) care transitions, (2) care management, and (3) Keeping in Touch.

[c] Originally, the AGH clinical director (a position not supported by HCIA) managed day-to-day operations and program data management and reporting. In the last program year, AGH used HCIA funding to add the program manager position to take over day-to-day program management and the data specialist position to take over data collection and reporting.

FTE = full-time equivalent; HCIA = Health Care Innovation Award; PCMH = patient-centered medical home.

AGH also created several new core clinical staff positions to support implementation of the other components of the HCIA-funded PCMH model at AGH and its seven primary care practices. AGH hired three care coordinators who were RNs with extensive clinical experience in inpatient and outpatient settings, as well as with experience using AGH's EMR. In addition, AGH used HCIA funds to support a nurse and social worker from the WCHD who conducted participant needs assessments and home visits as requested by providers, mainly for the care coordination program. Two retired nurse volunteers staffed the Keeping in Touch program.

B. Implementation effectiveness

In this section, we examine the evidence on implementation effectiveness—that is, we analyze measures of the intervention delivered and, when possible, compare those measures to the services the awardee intended to deliver. We assess the evidence on implementation effectiveness in five areas: (1) program enrollment, (2) service delivery, (3) staffing, (4) training, and (5) implementation timeliness. To conduct this analysis, we used data from interviews with program administrators and frontline staff, self-reported metrics included in AGH's self-monitoring and measurement reports to CMMI, and data from AGH on patients it enrolled in care coordination. We report metrics through June 2015, the end of the award period.

1. Program enrollment

The enrollment target for the entire PCMH program (1,314 enrollees) was exceeded by the end of the award period (1,460) (Table III.2). The cumulative number of enrollees in the care transitions component was 1,002, but AGH did not provide a separate enrollment target for the care transitions component. AGH did not provide a breakout of program participants by insurance type for each program component separately, although it did report the total number of Medicare FFS beneficiaries who received any of the program interventions. Throughout the implementation period, the awardee acknowledged challenges in collecting, managing, and analyzing data and consistently indicated the only measure it tracked separately for the care transitions program was readmissions, because reducing readmissions was the goal of the program.

Table III.2. Enrollment metrics–targets and actuals

Enrollment metrics	Awardee target	Value	Target met or exceeded
Cumulative number of Medicare FFS beneficiaries enrolled in any of three components of the PCMH (care coordination, care transitions, and Keeping in Touch)[a]	1,314	1,460	Yes
Cumulative number of patients (all insurance types) in care transitions component of the PCMH	None	1,002	n.a.

Source: Interviews and document review.

Note: All figures are as of June 30, 2015 as reported by the awardee.

[a]AGH did not provide distinct enrollment targets for different program components.

FFS = fee-for-service; PCMH = patient-centered medical home; n.a. = not applicable.

2. Service-related measures

AGH program administrators and staff reported that they faithfully adhered to the PCMH program model in delivering care transitions services; however, lack of defined targets for delivery of services for individual program components limited our ability to assess program implementation effectiveness. Next, we discuss the two service-related measures relevant to the care transitions program monitored and reported by AGH: (1) opt-out rate, and (2) encounters.

Opt-out rate. AGH reported that 119 patients opted out of any component of the program, including 114 who opted out of care transitions from January 2013 to June 2015 (Table III.3). Although AGH did not report a goal for the total number of patients who opted out of care transitions, it did set a target opt-out rate for all three program components of 1 percent. AGH did not meet this goal and had an average opt-out rate of 14.8 percent across all program components throughout the award period. The care transitions program had a slightly lower opt-out rate of 10.2 percent.

Encounters. AGH only reported the total volume of encounters across all program components and the proportion of follow-up conducted by telephone and in person. AGH was not able to break out these measures for the care transitions program. As of June 2015, care coordinators for the care management and care transitions program components had 7,422 encounters with participants (Table III.3). AGH estimated that about 90 percent of encounters consisted of follow-up telephone calls; the remaining 10 percent consisted of in-person encounters during participants' office visits at provider practices.

Table III.3. Service metrics (and targets, if applicable), by program component

Measure	Awardee target	Actual	Met target?	Which program components does this measure apply to?
Opt-out rate	1 percent	14.8 percent	No	All program components
Opt-out rate	Not specified	10.2 percent	--	Care transitions
Number of opt-outs	Not specified	114 patients	--	Care transitions
Participant encounters	Not specified	7,422 encounters	--	All program components

Source: Interviews and document review.

Note: Measures that apply to all program components include data from the care coordination program, the care transitions program, and the Keeping in Touch program.

3. Staffing measures

Staff. AGH employed 6.5 FTEs across its entire PCMH program, slightly exceeding its target of employing 4.5 FTEs across all program components (Table III.4). AGH did not provide a staffing target for its care transitions program.

Caseload. The care coordinator's caseload ranged from 40 to 50 patients, compared to AGH's target of 50 participants per care coordinator (Table III.4). The target caseload applied to each care coordinator in the care coordination and care transitions program components.

Table III.4. Staffing metrics (and targets, if applicable), by program component

Measure	Awardee target (source)	Actual (source)	Met target?	Which program components does this measure apply to?
Program staffing	4.5 FTEs	6.5 FTEs	Yes	All program components
Average care coordinator caseload	50 participants	40 to 50 participants	Close[a]	Care coordination and care transitions

Source: Interviews and document review.

[a] A higher caseload would not necessarily signal that the program was implemented more effectively. Rather, lower caseloads than the target might, all else equal, have permitted a more robust intervention for those who enrolled.

FTE = full-time equivalent.

4. HCIA-funded training

Motivational interviewing training was the only training provided by AGH that pertained directly to staff providing services in the care transition program. During the second year of implementation, program leaders and staff identified a need for care coordinators to learn ways to help participants improve their motivation and change their behaviors to better manage their conditions. To address this need, care coordinators, including the care transitions care coordinator, completed a course in motivational interviewing that provided guidance on how to engage participants and provide patient-centered care. Care transitions staff reported learning new skills to motivate their participants and help them reach their goals. AGH did not identify any other educational needs and did not conduct any additional training for the care transitions program. General training on the PCMH model was offered to AGH staff, providers, and partners at the beginning of the AGH initiative to staff, in preparation for the launch of the PCMH; this training was not specific to the care transitions program.

During our site visits in 2015, program leaders and frontline staff indicated that the motivational interviewing training helped improve the quality of services that frontline staff provided. Further assessment of trainings related to the care transitions program using the HCIA PCR trainee survey are not possible due to the small sample size (only one respondent participated in care transitions).

5. Program timeline

AGH conducted staff training (October 2012–June 2013) and began enrolling participants (December 2012–June 2013) to the care transitions program component according to the established timeline.

C. Summary of facilitators and barriers to implementation

Several factors helped implementation of AGH's HCIA-funded care transitions intervention, and others hindered implementation. We described those factors in detail in the second annual report (Moreno et al. 2015). Here, we summarize key facilitators and barriers, along with any new information since the second annual report that support those facilitators or barriers (Table III.5).

Two factors were particularly important in facilitating program implementation. First, AGH leadership faced strong financial incentives to successfully implement a PCMH model. These incentives included (1) new participation in shared savings programs after joining an accountable care organization (ACO) in January 2015, and (2) financial rewards in the last year of program implementation for avoiding unnecessary hospitalizations from Maryland's new global payment system. Second, the care transitions program had a highly trained and experienced care coordinator who provided high quality care transitions services to participants.

Two important barriers to implementation were (1) inadequate technological infrastructure that made it burdensome (and sometimes impossible) to retrieve timely patient information; and (2) the high needs of some participants, including social and financial barriers, that made it time-consuming (and sometimes impossible) to ensure patient compliance with post-discharge care plans.

Table III.5. Summary of key facilitators and barriers to the implementation of AGH's program

Item[a]	Description based on findings in the second annual report	Additional supporting data not available in the second annual report, if applicable
	Facilitators (domain)	
Resources (implementation process)	AGH had a trained, experienced nurse care coordinator who provided patient-centered care transition services. During the last year, the addition of a supervisor to oversee day-to-day program operations and a data specialist to manage data collection provided additional administrative support.	No new data
Program monitoring (implementation process)	AGH collected and monitored program metrics, including enrollment, utilization, and quality measures. The process of collecting data and producing reports proved time-consuming and labor-intensive but critical to informing program improvement decisions. Program leaders monitored outcome trends. Investigation of an observed increase in readmissions revealed that most readmissions occurred among participants admitted from skilled nursing facilities. In response, AGH built a relationship with a local skilled nursing facility to provide care transitions support. Finally, AGH tracked quality measures to identify opportunities to improve the quality of participants' care, although it did not provide specific quality measures that it tracked for care transitions.	No new data
Team characteristics (inner setting)	Program staff shared a strong commitment to teamwork and built a sense of camaraderie around the shared purpose of delivering high quality patient-centered care. Weekly team meetings provided opportunities for program staff across all components to discuss day-to-day processes, share problems encountered, and coordinate schedules. Frontline staff expressed a high level of comfort voicing concerns to administrators and a belief that their perspectives mattered in guiding program improvement.	No new data

Table III.5 *(continued)*

Item[a]	Description based on findings in the second annual report	Additional supporting data not available in the second annual report, if applicable
Payment model (outer setting)	During the first two years of program implementation, Maryland was promoting adoption of new provider payment models, which prompted strong leadership commitment to the PCMH model of care delivery and facilitated implementation of the care transitions program. Under a prior FFS model, reduction of admissions and ED visits translated to financial losses for AGH. In January 2014, Maryland shifted to a global payment model that rewards hospitals for avoiding unnecessary hospitalizations. Participation in other shared savings programs, including an ACO that AGH joined in January 2015, offered additional opportunities to achieve savings to support implementation of the PCMH model.	Administrators and executives believed the PCMH program (which includes the care transitions program) would help them achieve savings under the new payment model. They reported that, in the final quarter of implementation (April–June 2015), they were able to financially support the PCMH through the global payment model because they received credits for having already implemented a PCMH.
Barriers (domain)		
Technology (inner setting)	One internal factor—technological infrastructure capacity—was a barrier in implementing all three of AGH's program components, including the care transitions program. The care coordinator spent four hours each morning manually monitoring, collecting, and reporting data on AGH patients who visited the emergency room or were admitted to the hospital. The care coordinator's data-related tasks included identifying patients with AGH PCPs, identifying patients who participated in the care coordination program, and running reports of readmissions and calculating LACE scores. During the second site visit, one staff member described the challenge they faced: "The big barrier is data collection. We are a facility that needs help with EMRs and databases speaking to each other and pulling high quality trustworthy data. It is very difficult to pull data. It is done by hand, and it is very time-consuming."	No new data
Participant needs and resources (outer setting)	Throughout implementation of the care transitions and other programs, AGH encountered challenges related to participants' needs and resources. Some participants in the target population faced significant barriers to care. including low literacy, financial constraints, limited access to transportation, and lack of caregiver support. The care coordinator referred such participants to the care management program, where social work support provided by WCHD program staff helped connect participants to community resources to help meet these needs.	No new data

Source: Interviews and document review.

[a] We reviewed four domains associated with implementation experience: (1) program characteristics, (2) implementation process, (3) internal factors, and (4) external environment. Implementation research suggests that barriers and facilitators within these domains are important determinants of implementation effectiveness.

ACO = accountable care organization; ED = emergency department; EMR = electronic medical records; PCMH = patient-centered medical home; PCP = primary care provider; WCHD = Worchester County Health Department.

D. Conclusions about the extent to which the program, as implemented, reflects core design

AGH's care transitions program funded by the HCIA award was largely implemented as planned. There were no major delays in implementing the care transitions component, and AGH was able to deliver services to the target population, achieving a steady increase in enrollment. Although the patient opt-out rate was 10 percent, compared to AGH's goal of 1 percent, this still reflected a 90 percent participation rate overall. AGH was also able to deliver care transitions follow-up services as intended, adapting the frequency of follow-up to meet individual participant needs (according to interview respondents). Although it fluctuated month to month, AGH's care coordinator average caseload of 40 to 50 care transitions patients was close to the monthly targeted case load of 50 patients. Finally, AGH completed initially planned training and added supplementary training in motivational interviewing to enhance the effectiveness of care transitions.

The biggest challenge in program implementation was AGH's inadequate technological infrastructure capacity. The manual processes for monitoring, collecting, and reporting data were inefficient and time-consuming for frontline staff who could instead be providing services to participants. AGH alleviated some of the burden of manual data processes on the care coordinator by hiring a data specialist in the last year of program implementation to help with these tasks.

Although the care transitions program was implemented largely as planned, we again emphasize that the care transitions program was only a small part of the overall PCMH model that AGH implemented. The implementation of the other program components were not evaluated in this annual report because they could not be included in the impact evaluation; more information on the implementation of the entire AGH program can be found in the second annual report (Moreno et al. 2015).

IV. PROGRAM IMPACTS ON PATIENT OUTCOMES

In this section of the report, we draw conclusions, based on available evidence, about the impacts of AGH's care transitions component on patient outcomes in four domains: (1) quality-of-care processes, (2) quality-of-care outcomes, (3) service use, and (4) spending. We first describe the methods for estimating impacts (Section A) and then the characteristics of the treatment group at the start of the intervention (Section B). We next demonstrate that the treatment group was similar at the start of the intervention to the comparison group, which is important for limiting potential bias in impact estimates (Section C). Finally, we describe the quantitative impact estimates, their plausibility given implementation findings, and our conclusions about program impacts in each domain (Section D).

The findings in this report update the impact results from the Second Annual Report (Moreno et al. 2015) for AGH. Specifically, we: (1) include additional treatment group beneficiaries by extending the enrollment period for the post-intervention cohort, (2) rematched the treatment beneficiaries (in the post-intervention cohort only) to potential comparison beneficiaries so that all treatment enrollees had one or more matched comparison beneficiaries,

(3) extended the period that outcomes are measured in claims data by eleven months (from December 31, 2014, to November 30, 2015), and (4) added one outcome measure (inpatient admissions followed by an ambulatory care visit with a primary care or specialist provider within 14 days).

A. Methods

1. Overview

We estimated program impacts using a difference-in-differences framework. To implement this framework, we defined two cohorts of Medicare beneficiaries: (1) a *post-intervention cohort,* which included beneficiaries discharged from AGH after the program began on January 1, 2013, and who met the program eligibility criteria (the post-intervention treatment group) and their matched comparison beneficiaries (the post-intervention comparison group); and (2) a *pre-intervention cohort,* which included beneficiaries discharged at least six months before the intervention began but who otherwise met the program eligibility criteria (the pre-intervention treatment group) and their matched comparison beneficiaries (the pre-intervention comparison group). In each intervention quarter following the qualifying hospital discharge, we (1) calculated the difference in outcomes between the post-intervention treatment and comparison groups that quarter, and (2) subtracted any difference in outcomes between the pre-intervention treatment and comparison groups in the corresponding quarter.

We pre-specified primary tests, describing the evidence we would need to conclude that the program was effective, and the awardee and CMMI reviewed these tests. Each test specified a population, outcome, time period, expected direction of effect, and threshold that we counted as substantively important. The purpose of these primary tests was to focus the impact evaluation on hypotheses that would provide the most robust evidence about program effectiveness. We used the results from the primary tests and secondary tests (robustness checks) to draw conclusions about program impacts in each of the four evaluation domains: (1) quality-of-care process, (2) quality-of care outcomes, (3) service use, and (4) spending. In the remaining subsections, we describe each component of the impact evaluation in more detail.

2. Treatment group definition

Post-intervention treatment group. The post-intervention treatment group included Medicare FFS beneficiaries who met two criteria. First, they had to meet AGH's program eligibility criteria, to the extent that we could replicate them in claims. That is, the beneficiary had to (1) be discharged from AGH from February 1, 2013 (the date AGH enrolled its first patient into the care transitions component of its program), to May 31, 2015; and (2) be an AGH patient. We identified AGH patients as those who had their most recent primary care visit with an AGH provider (we received the list of providers from AGH) or who had the plurality of their primary care visits in the past two years with an AGH provider. This May 31, 2015, cutoff for the sample allowed all treatment group members to have potentially received a full dose of program services—that is, to have received 30 days of transitional care before the program ended on June 30, 2015. Second, a beneficiary had to be continuously enrolled in FFS Medicare for the four quarters before his or her qualifying discharge. This restriction improved the matching of

treatment to potential comparison beneficiaries by ensuring we could use a full year of claims to develop baseline indicators of service use and diagnoses for matching.

Pre-intervention treatment group. We defined the pre-intervention group using the same claims-based rule as for the post-intervention group, with one difference. The beneficiary had to have been discharged from AGH from July 1, 2011, to June 30, 2012, allowing us to follow each beneficiary for at least six months before the intervention began.

Additional sample restrictions in each intervention quarter. To be included in the analytic sample in any given quarter, each treatment group member had to meet two additional criteria to contribute an observation for the quarter. First, the beneficiary's outcomes had to be observable in Medicare claims for at least one day during the quarter. Outcomes were observable for beneficiaries who were enrolled in Medicare FFS (Part A and B), were alive, and had Medicare as their primary payer (including beneficiaries who were dually eligible for Medicaid). Second, all of a treatment beneficiary's matched comparison beneficiaries (see next section) also had to be in the sample during the quarter, so that the treatment beneficiary's outcomes could be compared to the outcomes for all of his or her comparison beneficiaries. (These sample restrictions were applied after constructing the comparison group, since only data from the period before each beneficiary's qualifying discharge were used for constructing the comparison group.)

Intent-to-treat criteria. These claims-based rules used to define the two treatment groups have two advantages over an alternative definition that includes only those who actually enrolled in the care transitions component of AGH's program. First, because AGH targeted any patients discharged from AGH with an AGH PCP, this definition corresponds to everyone the program intended to treat (that is, the definition follows an intent-to-treat design). Most notably, the claims-based definition includes Medicare patients who did not consent to participate in the program or who could not be contacted by the care coordinator. One limitation of the claims-based rules is that we include some Medicare patients AGH did not intend to treat—namely, those already being monitored by another AGH program (for example, the cancer center). We also included 25 patients who, at some point during the intervention period, were also enrolled in the other major component of AGH's PCMH program—care coordination for participants with chronic conditions. Second, we can use exactly the same definition to identify a pre-intervention treatment group, which is needed to implement the difference-in-differences design. Although the intent-to-treat results are most relevant for policymakers, some stakeholders could be interested in impacts among only those who received the treatment. When comparing our treatment group definition to the roster of actual AGH enrollees, we found that 62 percent of the treatment group members were actually enrolled in the care transitions component of the program. Therefore, any impacts measured among the full treatment group might understate the impacts among only those who actually enrolled. We did not conduct sensitivity analyses to estimate impacts among only those who enrolled because, without the ability to replicate individuals' enrollment decisions using claims data, we could not create a comparison group that would have made such sensitivity analyses meaningful.

3. Comparison group definition

Post-intervention comparison group. We constructed a comparison group of Medicare FFS beneficiaries who were similar to the post-intervention treatment group beneficiaries. This section describes how we constructed the matched comparison group; Section IV.C shows the balance we achieved between the two groups on the matching variables.

We used three steps to construct the comparison group:

First, we identified a pool of *potential* comparison members. This pool consisted of all Medicare FFS beneficiaries (1) discharged from February 1, 2013, to May 31, 2015, from Peninsula Regional Medical Center (PRMC) in Salisbury, Maryland, about 30 miles from AGH, but which did not implement the care transitions component; or (2) discharged from AGH (in the same time frame) but not attributed to an AGH provider (so the beneficiaries were not assigned to the post-intervention treatment group). We set the day following hospital discharge as the potential comparison beneficiary's pseudo-enrollment date (the date a beneficiary was assigned to the potential comparison pool). If a potential comparison beneficiary was discharged more than once, we set his or her pseudo-enrollment date to the day after the first discharge. The advantage of drawing comparison beneficiaries from those discharged from AGH or the nearby PRMC is that the comparison beneficiaries would be exposed to market forces similar to those of the treatment beneficiaries. This is particularly important because the Centers for Medicare & Medicaid Services (CMS) and other payers recently began paying Maryland hospitals based on global budgets, which creates a new type of incentive for reducing admissions that is not present outside the state.

Second, we used the Medicare Enrollment Database and a beneficiary's Medicare claims in the 12 to 36 months before his or her pseudo-enrollment date to develop baseline characteristics for each beneficiary.

Finally, we used propensity-score matching and exact matching techniques to limit the potential comparison pool to a list of matched comparison beneficiaries. Matching aims to reduce selection bias in observational studies by selecting comparison beneficiaries from the pool who are roughly equivalent to the treatment group across key baseline characteristics. The goal of matching is to achieve baseline equivalence between the treatment and matched comparison groups on the variables included in the matching process (Stuart 2010). For AGH, we matched on demographic characteristics, Medicare and Medicaid dual enrollment, original reason for Medicare entitlement, health status and chronic conditions, service use and spending 3 months before enrollment or pseudo-enrollment, and service use and spending 4 to 12 months before enrollment or pseudo-enrollment. Service use and spending before enrollment or pseudo-enrollment were included as matching variables because these variables are important predictors of the outcomes in the post-intervention period.

Within the family of propensity-score matching methods, we implemented a technique called *full matching* to form matched sets that contain one treatment beneficiary and one or more comparison beneficiaries. The important benefit of full matching is that it achieves maximum bias reduction on observed matching variables and, subject to this constraint, maximizes the size

of the comparison sample (Rosenbaum 1991; Hansen 2004). Each treatment beneficiary was matched to up to four beneficiaries from the potential comparison group. Ten of the 648 post-intervention treatment beneficiaries were dropped because they could not be matched to any potential comparison beneficiaries.

We used exact matching techniques to ensure matched comparison group beneficiaries had (1) a qualifying inpatient discharge within 90 days of the treatment beneficiary's enrollment date, (2) the same gender as the treatment beneficiary, and (3) the same reason for the hospitalization that caused a person to enter the treatment or comparison group. Specifically, we used 27 unique modified diagnosis-related group (MDRG) codes to define the types of hospital stays for most treatment beneficiaries. For the remaining treatment group beneficiaries, MDRG codes were too uncommon to provide sufficient matches in the comparison group; in this case, we used major diagnostic category (MDC) codes (instead of MDRG codes) for exact matching.

Pre-intervention comparison group. We constructed a comparison group of Medicare beneficiaries who were similar to the pre-intervention treatment group beneficiaries. The pool of potential comparison members consisted of all Medicare FFS beneficiaries (1) discharged from PRMC from July 1, 2011, to June 30, 2012; or (2) discharged from AGH (in the same time frame), but not attributed to an AGH provider. Because the sample sizes were smaller in the pre-intervention period, we exact-matched on 15 MDRG codes (instead of 27). Otherwise, the methods for constructing the pre-intervention comparison group were the same as described earlier for the post-intervention comparison group. Five of the 231 pre-intervention treatment beneficiaries were dropped because they could not be matched to any potential comparison beneficiaries.

Additional sample restrictions in each intervention quarter. To be included in the analytic sample, a comparison group beneficiary had to meet the same additional criteria as the treatment group members—that is, the beneficiary had to be observable in Medicare claims for at least one day of the quarter. Furthermore, the comparison beneficiary's matched treatment group beneficiary also had to be in the sample during the quarter, so that the comparison beneficiary's outcomes could be compared with the outcomes for his or her treatment beneficiary.

4. Construction of outcomes and covariates

We used Medicare claims for beneficiaries assigned to the treatment and comparison groups to develop two types of variables: (1) *outcomes,* defined for each person in each intervention quarter during which they were members of the treatment or comparison group; and (2) *covariates,* which describe a beneficiary's characteristics at the time of enrollment or pseudo-enrollment and were used in the regression models for estimating impacts to adjust for existing characteristics. We used one set of baseline covariates, without updating them each quarter, to avoid controlling in each intervention quarter for previous quarters' program effects, because this would bias the effect estimates away from detecting true impacts. For the post-intervention cohort of beneficiaries, the Medicare claims covered services provided from three years before the start of the intervention (February 1, 2010) through November 30, 2015. This guaranteed all beneficiaries in the post-intervention cohort could potentially be observed for two full quarters after their enrollment or pseudo-enrollment date, corresponding to the primary test period. For

the pre-intervention cohort, the claims covered services from July 1, 2008, to December 31, 2012. We ended on December 31, 2012, to avoid including outcomes for the pre-intervention cohort that actually occurred during the intervention period. Appendix 1 provides details on the methods we used to construct these variables.

Outcomes. We calculated four quarter-specific outcomes and grouped them into three domains:

1. Domain: Quality-of-care processes
    - Ambulatory-care follow-up visit within 14 days of a hospital discharge (binary variable for each beneficiary); for each person in the sample, this is a binary variable that equals one if the beneficiary had a visit with a primary care or specialist physician within 14 days of the discharge that qualified him or her for the treatment or comparison group, and zero if not.
2. Domain: Quality-of-care outcomes
    - 30-day unplanned readmission rate (binary variable for each beneficiary); for each person in the sample, this is a binary variable that equals one if the beneficiary had an unplanned readmission within 30 days of the discharge that qualified him or her for the treatment or comparison group, and zero if not.
3. Domain: Service use
    - All-cause inpatient admissions (number/quarter).
    - Outpatient ED visit rate (number/quarter); outpatient ED visits are defined as ED visits or observational stays that do not end in a hospital admission.
4. Domain: Spending
    - Total Medicare Part A and B spending ($/month).

Four of these outcomes—all but the quality-of-care process measure—are outcomes that CMMI has specified as "core" for the evaluations of all HCIA programs.

Covariates. The covariates, defined at the enrollment (treatment group) or pseudo-enrollment date (comparison group) include (1) demographics (age, age-by-gender interactions, race and ethnicity, and lives in a zip code with a poverty rate of 20 percent or higher); (2) whether dually enrolled in Medicare and Medicaid; (3) original reason for Medicare entitlement (old age, disability, or end-stage renal disease); (4) the number of months with Part A and B coverage 4 to 12 months before a beneficiary's pseudo-enrollment date; (5) Hierarchical Condition Category (HCC) score, which is a continuous score that CMS developed to predict a beneficiary's future Medicare spending; (6) whether a beneficiary has each of six chronic conditions (cancer, congestive heart failure [CHF], chronic obstructive pulmonary disease [COPD], chronic kidney disease, Alzheimer's disease-related disorders, or senile dementia), created by applying Chronic Condition Warehouse algorithms to claims in the 12 to 36 months (depending on the condition) before the beneficiary's enrollment or pseudo-enrollment date; and (7) service use and Medicare Part A and B spending in the prior 3 months, and 4 to 12 months.

Service use includes the number of unplanned readmissions, the number of inpatient discharges, the number of ED visits, and an indicator for one or more PCP visits.

5. Regression models

We used a regression model to implement the difference-in-differences framework. For each quarter-specific outcome, the model estimates the relationship between the outcome and predictor variables, assuming that each of the predictor variables has a linear (additive) relationship with the outcome. The predictor variables include (1) the beneficiary-level covariates (defined in Section IV.A.4); (2) an interaction of each beneficiary-level covariate with each intervention quarter; (3) indicators for each matched set (a treatment beneficiary plus his or her matched comparison beneficiaries) in each quarter; (4) whether the beneficiary was assigned to the treatment or comparison group; (5) an interaction of a beneficiary's treatment status with an indicator for being in the post-intervention cohort (as opposed to the pre-intervention cohort); (6) an interaction of a beneficiary's treatment status with each intervention quarter; and (7) a three-way interaction of a beneficiary's treatment status with each intervention quarter with an indicator for being in the post-intervention cohort. Appendix 2 provides details on the regression methods, including descriptions of the weights each beneficiary receives in the model and how the regressions account for correlation in outcomes across quarters for a given individual, and across individuals in the same matched set.

The estimated relationship between the three-way interaction term and an outcome in a given quarter provides the difference-in-differences estimate for that quarter and outcome. It measures the average difference between outcomes for post-intervention beneficiaries assigned to the treatment and comparison groups in a certain quarter, subtracting out any differences between the pre-intervention treatment and comparison groups during the same quarter. The model quantifies the uncertainty in the difference-in-differences estimates, allowing for statistical tests that determine whether observed differences are likely due to chance.

6. Primary tests

Table IV.1 shows the primary tests for AGH, by domain. Each test specified a population, outcome, time period, expected direction of effect, and threshold that we counted as substantively important. The purpose of these primary tests was to focus the impact evaluation on hypotheses that would provide the most robust evidence about program effectiveness (see Appendix 3 for details and a description of how we selected each test). We provided the awardee and CMMI an opportunity to comment on the primary tests.

Our rationale for selecting these primary tests is as follows:

- **Outcomes.** AGH's central goal is to reduce hospitalizations, ED visits, and Medicare Part A and B spending, so our primary tests address these three outcomes. In addition, the primary tests address one quality-of-care outcome the intervention was expected to affect: 30-day unplanned hospital readmissions. Finally, we included one quality-of-care process measure that directly aligned with AGH's theory of action: receipt of a follow-up ambulatory care visit with a primary care or specialist provider within 14 days of hospital discharge. AGH's original HCIA proposal contained no separately stated goals for the care transitions

component. Therefore, we assumed that AGH's target outcomes for the care transitions program were the same as those for the HCIA program as a whole.

- **Time period.** AGH's proposal contained no specific time frame for reaching the program goals, but the literature on transitional care interventions indicates effects on readmissions tend to be concentrated in the period following an initial, or index, hospital discharge (Peikes et al. 2012). For this reason, the primary tests measure impacts on the readmission rate in the 30 days following the (index) inpatient admission associated with the beneficiary's enrollment or pseudo-enrollment (that is, the stay that qualified a person for the treatment or comparison group). The time period is defined this way because the matching variables were balanced for the treatment and comparison groups at the beneficiaries' enrollment or pseudo-enrollment dates due to our matching approach, but they would not have been balanced for subsequent inpatient admissions. Similarly, the primary tests measure impacts on the follow-up ambulatory care visit rate in the 14 days following the (index) inpatient admission associated with the beneficiary's enrollment or pseudo-enrollment. We expected effects for the other three outcomes—hospitalizations, ED visits, and spending—to be concentrated in the first one to three months following the enrollment admission. For these three outcomes, however, we set the time period for the primary tests to the first two quarters immediately following the enrollment admission, because some studies show impacts of transitional care programs over longer periods (Peikes et al. 2012). Our report covers Medicare beneficiaries discharged from AGH through May 31, 2015, and includes outcome data constructed with claims data through November 30, 2016. This definition allows all treatment members to have potentially received at least one month of services before the program ended on June 30, 2015, and to be observed in claims at least six months following the enrollment admission.

- **Population.** AGH expected to have impacts for the population of beneficiaries enrolled in the care transitions component of its program. Therefore, the primary tests included all (observable) Medicare FFS beneficiaries who met the care transitions component's enrollment criteria. Although AGH did enroll patients with non-Medicare insurance, we do not have data for patients with Medicaid (without Medicare), commercial insurance, or no insurance. We did not include Medicaid beneficiaries in our primary tests (unless they were also enrolled in Medicare) because Medicaid data was not timely enough to cover the primary test period for a substantial number of Medicaid beneficiaries.

- **Direction (sign) of the impact estimate.** For the quality-of-care process measure, we expected the impact estimate to be positive, signaling an increase in the percentage of people receiving recommended care. For all other outcomes, we expected the impact estimates to be negative, indicating a reduction in service use or overall expenditures.

- **Substantive thresholds.** We express the substantive threshold as a percentage change from the counterfactual—that is, the outcomes that beneficiaries in the treatment group would have had if they had not received the intervention (that is, the treatment). The 11.6 to 15.0 percent thresholds we chose for substantive importance (depending on the outcome) are 75 percent of AGH's expected effects for the HCIA program as a whole. (We used 75 percent, recognizing that AGH could still be considered successful if it approached, but did not achieve, its fully anticipated effects. AGH's proposal contained no separately stated goals

for the care transitions component. Therefore, we assumed that AGH's target outcomes for the care transitions program were the same as those for the HCIA program as a whole.) The 15 percent threshold for the quality-of-care outcome and processes measures was extrapolated from the literature (Peikes et al. 2011, 2012; Rosenthal et al. 2016), because AGH did not specify by how much it expected to reduce these hospitalizations.

Table IV.1. Specification of the primary tests for AGH's care transitions component

Domain (number of tests in the domain)[a]	Outcome (units)	Time period for impacts (controlling for pre-intervention differences)[b]	Population	Substantive threshold (expected direction of the effect)[c,d]
Quality-of-care process (1)	Inpatient admissions followed by an ambulatory care visit with a primary care or specialist provider within 14 days (binary [yes or no]/discharge)	The 14 days immediately following the enrollment admission[e]	All observable Medicare FFS beneficiaries attributed to the treatment group with a qualifying enrollment admission (index stay)[e]	15.0% (+)
Quality-of-care outcomes (1)	30-day unplanned hospital readmissions (binary [yes or no]/discharge)	The 30 days immediately following the enrollment admission[e]	All observable Medicare FFS beneficiaries attributed to the treatment group with a qualifying enrollment admission (index stay)[e]	15.0% (-)
Service use (2)	All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	Average over the first two quarters immediately following the enrollment admission[e]	All observable Medicare FFS beneficiaries attributed to the treatment group	15.0% (-)
	Outpatient ED visit rate (#/1,000 beneficiaries/quarter)	Average over the first two quarters immediately following the enrollment admission[e]	All observable Medicare FFS beneficiaries attributed to the treatment group	15.0% (-)
Spending (1)	Medicare Part A and B spending ($/beneficiary/month)	Average over the first two quarters immediately following the enrollment admission[e]	All observable Medicare FFS beneficiaries attributed to the treatment group	11.6% (-)

[a] We adjusted the *p*-values from the primary test results for the multiple comparisons made within each domain, but not across domains.

[b] The regression models for estimating program impacts controlled for differences in outcomes between the pre-intervention treatment and comparison groups.

[c] For all-cause hospitalizations, the outpatient ED visit rate, and Medicare spending, we set the substantive threshold to 75 percent of AGH's expected effect. The 15 percent threshold for the quality-of-care outcome and process measures, for which AGH did not set explicit targets, was extrapolated from the literature (Peikes et al. 2011, 2012; Rosenthal et al. 2016), because AGH did not specify by how much it expected to improve these outcomes.

[d] The substantive threshold is the impact as percentage of the counterfactual. The counterfactual is the outcome the treatment group would have had in the absence of the HCIA-funded intervention.

[e] The enrollment admission is the inpatient discharge that led to a beneficiary being assigned to the treatment or comparison group.

AGH = Atlantic General Hospital; ED = emergency department; FFS = fee-for-service; HCIA = Health Care Innovation Award.

7. Secondary tests (robustness checks)

We also conducted secondary quantitative tests to help corroborate the findings from the primary tests. This is important, because some of the differences observed between the treatment and comparison groups in the primary test results could reflect limitations of the nonexperimental impact evaluation design or random fluctuations in the data. We have greater confidence in the primary results if they are generally consistent with the expected broader pattern of results from the secondary tests. Specifically, we repeated the primary tests above, but excluded from the sample 25 beneficiaries in the treatment group who were enrolled in the care coordination component of AGH's PCMH program, as well as their 81 matched comparison beneficiaries. If there were large differences between the primary tests and the secondary tests, it could suggest that impact estimates were being (fully or partially) driven by the care coordination component, not the care transitions component, of AGH's program.

8. Synthesizing evidence to draw conclusions about program impacts

Within each domain, we drew one of five conclusions about program effectiveness, based on the primary test results, the results of secondary tests, and the plausibility of those findings given the implementation evidence. These five possible conclusions are as follows:

1. Statistically significant favorable effect (the highest level of evidence)
2. Substantively important (but not statistically significant) favorable effect
3. Substantively important (but not statistically significant) unfavorable effect
4. No substantively large effect
5. Indeterminate effect

We cannot conclude that a program has a statistically significant unfavorable effect. This is because, in consultation with CMMI, we decided to use one-sided statistical tests (which do not test for evidence of unfavorable effects). We used one-sided tests to increase the probability that, if a program truly did have impacts, we would be able to detect them.

Appendix 3 describes our decision rules for each of the five possible conclusions. In short, we concluded that a program had a statistically significant favorable effect in a domain if (1) at least one primary test result in the domain was favorable and statistically significant, after adjusting the statistical tests to account for multiple tests (if applicable) within a domain; or (2) the average impact estimate across all primary tests in the domain was favorable and statistically significant. In both cases, we also needed to determine that the primary test results were plausible, given the results of the secondary tests and implementation evidence. We concluded that a program had a substantively important favorable effect if the average impact estimate in the domain was substantively important but not statistically significant, and if the result was plausible given the secondary tests and implementation evidence. In contrast, if the average impact estimate was unfavorable (opposite the hypothesized direction) and larger than the substantive threshold, and unfavorable effects were plausible given the other evidence, we concluded the program had a substantively important unfavorable effect. If the tests in a domain did not meet any of these criteria, we drew one of two conclusions. First, if the tests for at least one outcome in the domain (or all outcomes in the domain together) had sufficient statistical

power to detect an impact of the size of the substantive threshold with at least 75 percent probability, we concluded that there was not a substantively large effect, because we are reasonably confident that we would have detected a substantively large effect had there been one. Alternatively, if the power was not sufficient to detect this type of impact, we concluded the impact in the domain was indeterminate. Indeterminate means either that the program truly did not have effects that were substantively large, or that it did, but our statistical tests were not able to detect them.

B. Characteristics of the treatment group at baseline

This section describes the characteristics of the 638 beneficiaries in the post-intervention treatment group at the date of enrollment or pseudo-enrollment, shown in the first column of Table IV.2, panel A, and the characteristics of the 226 beneficiaries in pre-intervention treatment group (before the program began), shown in the first column of Table IV.2, panel B. For context, the last column shows the values of relevant variables for the national Medicare population, when available.

Post-intervention treatment group. Some demographic characteristics of the 638 Medicare FFS beneficiaries in the post-intervention treatment group (such as gender and age) are similar to benchmarks for the national Medicare population, but other characteristics in Table IV.2, panel A, indicate the treatment group has more health care needs than the general population. The HCC risk score for the treatment group is 2.47, indicating that the group could be expected to have Medicare spending that is 2.47 times higher than the national average (1.00) over the next year. The prevalence of COPD, chronic kidney disease, and CHF in the treatment group was more than twice the national average.

Treatment group members also had high service use (inpatient admissions and outpatient ED visits) and spending relative to national Medicare averages. For example, the treatment group beneficiaries had, on average, 1,092 hospitalizations (per 1,000 beneficiaries) in the quarter before their enrollment dates and 67 hospitalizations (per 1,000 beneficiaries per quarter) in the period 4 to 12 months before their enrollment dates, compared to a national average of 74 hospitalizations (per 1,000 beneficiaries per quarter). The program-targeting criteria explain the spike in this utilization outcome in the quarter before pseudo-enrollment. The program enrolled people who were in the hospital; therefore, the population hospitalization rate had to reach or exceed 1,000 (corresponding to at least one stay per person) in that quarter. These hospitalizations, and perhaps other utilization, drove up Medicare spending as well.

Pre-intervention treatment group. Although the pre-intervention treatment group was not required to be the same as the post-intervention treatment group by construction, the two groups were largely similar. The characteristics in panel B of Table IV.2 demonstrate the pre-intervention treatment group had significant health care needs, with average HCC scores of 2.73 and incidences of COPD, chronic kidney disease, and CHF higher than the national averages. The average service use and spending patterns over 12 months before enrollment for the pre-intervention treatment group were similar to patterns of the post-intervention treatment group. However, some differences between the two groups in the reasons for hospitalization were apparent.

Table IV.2. Characteristics at baseline of treatment and comparison beneficiaries in the pre- and post-intervention cohorts

Characteristic	Treatment group (n = 638)	Unmatched comparison pool (n = 9,905)	Comparison group (n = 2,232)	Absolute difference[a]	Standard-ized difference[b]	Medicare FFS average
Panel A: Post-intervention cohort						
Exact match variables[c]						
Female (%)	55.2	55.8	55.2	0	0	54.7[d]
Number of days from January 1, 2013, to enrollment	441.9	413.9	439.9	2.0	0.008	n.a.
Reason for hospitalization[e]						
MDRG 114: Intracranial hemorrhage or cerebral infarction (%)	4.7	5.4	4.7	0	0	NA
MDRG 409: COPD	4.9	3.7	4.9	0	0	
MDRG 410: Simple pneumonia and pleurisy (%)	6.7	6.1	6.7	0	0	NA
MDRG 524: Heart failure and shock (%)	5.2	5.4	5.2	0	0	NA
MDRG 807: Major joint replacement	5.8	8.0	5.8	0	0	NA
MDRG 1110: Renal failure (%)	4.1	3.0	4.1	0	0	NA
MDRG 1808: Septicemia (%)	6.6	6.0	6.6	0	0	NA
Propensity matched variables[f]						
Demographic characteristics						
Age (years)	76.8	74.9	76.5	0.3	0.029	71[g]
Race: white (%)	92.3	82.1	90.8	1.5	0.051	81.8[d]
Zip code poverty rate greater than 20 percent (%)	2.0	12.3	3.2	-1.1	-0.069	NA
Medicare-related characteristics						
Dual status at enrollment	12.9	21.4	12.9	0.0	-0.001	22[h]
Original reason for entitlement (%)						
Disability	16.9	23.6	17.9	-1.0	-0.025	16.7[d]
ESRD	0.2	1.4	0.3	-0.1	-0.022	0.13[d]
Health status and chronic conditions						
HCC risk score	2.47	2.63	2.58	-0.10	-0.066	1.0
Chronic conditions[i] (%)						
Alzheimer's	8.0	6.5	7.3	0.7	0.027	4.9[j]
Alzheimer's disease, related disorders, or senile dementia	16.8	16.3	16.0	0.7	0.020	11.1[j]
Cancer	17.4	17.4	18.8	-1.4	-0.037	NA
CHF	37.3	38.6	38.6	-1.3	-0.027	15.3[j]
COPD	29.9	31.5	32.0	-2.1	-0.044	11.8[j]
CKD	41.8	46.4	44.0	-2.2	-0.044	16.2[j]
Diabetes	42.5	43.3	42.1	0.3	0.007	28.0[j]
Service use and spending 3 months before enrollment or pseudo-enrollment						
Number of unplanned readmissions (#/1,000 beneficiaries/quarter)	44	59	30	14*	0.078	NA
Number of hospitalizations (#/1,000 beneficiaries/quarter)	1,092	1,114	1,071	21*	0.080	74[k]
Number of ED visits (#/1,000 beneficiaries/quarter)	406	367	395	11	0.014	105[l]
Primary care (%)[m]	96.4	95.9	96.1	0.3	0.016	NA
Medicare spending ($/month)	6,125	6,982	6,097	28	0.005	860[n]
Service use and spending 4 to 12 months before enrollment or pseudo-enrollment						
Number of unplanned readmissions (#/1,000 beneficiaries/quarter)	3	17	3	0	0	NA
Number of hospitalizations (#/1,000 beneficiaries/quarter)	67	101	63	4	0.030	74[k]
Number of ED visits (#/1,000beneficiaries/quarter)	241	239	229	12	0.029	105[l]
Primary care (%)[m]	95.1	85.2	94.3	0.9	0.037	NA
Medicare spending ($/month)	1,198	1,391	1,153	45	0.021	860[n]

Table IV.2 *(continued)*

Characteristic	Treatment group (n = 226)	Unmatched comparison pool (n = 4,395)	Comparison group (n = 1,008)	Absolute difference[a]	Standard-ized difference[b]	Medicare FFS average
Panel B: Pre-intervention cohort						
Exact match variables[c]						
Female (%)	56.2	57.8	56.2	0	0	54.7[d]
Number of days from January 1, 2013, to enrollment	-364.1	-380.3	-366.9	2.8	0.028	n.a.
Reason for hospitalization[e]						
MDRG 114: Intracranial hemorrhage or cerebral infarction (%)	4.4	5.1	4.4	0	0	NA
MDRG 409: COPD	4.4	4.5	4.4	0	0	
MDRG 410: Simple pneumonia and pleurisy (%)	6.2	5.6	6.2	0	0	NA
MDRG 524: Heart failure and shock (%)	8.0	6.9	8.0	0	0	NA
MDRG 1110: Renal failure (%)	6.6	3.4	6.6	0	0	NA
MDRG 615: GI hemorrhage (%)	4.9	9.1	4.9	0	0	
MDRG 807: Major joint replacement (%)	4.4	5.1	4.4	0	0	NA
Propensity matched variables[f]						
Demographic characteristics						
Age (years)	77.9	75.7	77.5	0.4	0.040	71[g]
Race: white (%)	93.8	82.7	91.2	2.6	0.093	81.8[d]
Zip code poverty rate greater than 20 percent (%)	4.9	11.5	6.1	-1.2	-0.049	NA
Medicare-related characteristics						
Dual status at enrollment	9.3	20.3	11.5	-2.2	-0.070	22[h]
Original reason for entitlement (%)					-0.011	
Disability	12.8	22.1	15.9	-3.1	-0.088	16.7[d]
ESRD	0	1.5	0.4	-0.4	-0.081	0.13[d]
Health status and chronic conditions						
HCC risk score	2.73	2.78	2.68	0.06	0.037	1.0
Chronic conditions[i] (%)						
Alzheimer's	8.4	8.5	8.7	-0.3	0.041	4.9[j]
Alzheimer's disease, related disorders, or senile dementia	24.3	19.9	23.5	0.9	-0.011	11.1[j]
Cancer	22.6	17.7	20.9	1.7	0.035	NA
CHF	43.4	44.7	41.6	1.7	-0.009	15.3[j]
COPD	33.2	35.3	33.6	-0.4	0.033	11.8[j]
CKD	52.7	50.4	51.0	1.7	-0.030	16.2[j]
Diabetes	44.7	43.7	46.2	-1.5	0.021	28.0[j]
Service use and spending 3 months before enrollment or pseudo-enrollment						
Number of unplanned readmissions (#/1,000 beneficiaries/quarter)	27	67	24	3	0.017	NA
Number of hospitalizations (#/1,000 beneficiaries/quarter)	1,084	1,129	1,078	6	0.022	74[k]
Number of ED visits (#/1,000 beneficiaries/quarter)	296	375	303	-7	-0.012	105[l]
Primary care (%)[m]	96.9	95.7	95.9	1.0	0.051	NA
Medicare spending ($/month)	6,603	7,203	6,116	486	0.081	860[n]
Service use and spending 4 to 12 months before enrollment or pseudo-enrollment						
Number of unplanned readmissions (#/1,000 beneficiaries/quarter)	7	33	6	1	0.031	NA
Number of hospitalizations (#/1,000/quarter)	106	160	103	4	0.019	74[k]
Number of ED visits (#/1,000 beneficiaries/quarter)	252	223	204	48*	0.136	105[l]
Primary care (%)[m]	95.6	87.0	93.5	2.0	0.083	NA
Medicare spending ($/month)	1,306	1,680	1,266	40	0.016	860[n]

Table IV.2 *(continued)*

Sources: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at CMS. Zip code poverty rate merged from the American Community Survey ZIP Code Characteristics.

Notes: Characteristics are measured at the date of the inpatient discharge from AGH or PRMC that led to a beneficiary's assignment to the treatment or comparison group (the beneficiary's enrollment or pseudo-enrollment date). The post-intervention cohort included beneficiaries whose enrollment or pseudo-enrollment dates were from February 1, 2013, to May 31, 2015, and the pre-intervention cohort included beneficiaries whose enrollment or pseudo-enrollment dates were from July 1, 2011, to June 30, 2012.The comparison group means were weighted based on the number of matched comparisons per treatment beneficiary. For example, if four comparison beneficiaries were matched to one treatment beneficiary, each of the four comparison beneficiaries had a matching weight of 0.25.

Absolute differences might not be exact due to rounding.

[a] The absolute difference is the difference in means between the matched treatment and comparison groups.

[b] The standardized difference is the difference in means between the treatment and comparison groups divided by the SD of the variable, which is pooled across the treatment and comparison groups.

[c] Variables on which we required treatment and comparison members to match exactly. For example, a treatment group beneficiary whose reason for hospital discharge was intracranial hemorrhage or cerebral infarction (MDRG 1114) could be matched only to a comparison beneficiary who had the same reason for discharge. The date of the qualifying inpatient discharge for matched comparison beneficiaries had to be within 90 days of the treatment beneficiary's enrollment date.

[d] Chronic Conditions Data Warehouse (2014a, Table A.1).

[e] The reason for the hospitalization that caused a person to enter the treatment or comparison group. We used MDRG codes to define the types of hospital stays. In addition to the 7 hospitalization types listed in the table, we exactly matched on 20 other MDRGs (for a total of 27 MDRG codes), which captured the reason for discharge for most treatment beneficiaries. For the remaining treatment group beneficiaries, MDRG codes were too uncommon to provide sufficient matches in the comparison group; in such cases, MDC codes (instead of MDRG codes) were used for exact matching. To pay acute care inpatient FFS claims, Medicare assigns discharges to Medicare severity diagnosis-related groups (MS–DRGs), which group patients with similar clinical problems expected to require similar amounts of hospital resources; MDRGs group one or more related DRG codes into larger categories. MDC codes, in turn, group one or more MDRG codes together into even larger categories. Because the sample sizes were smaller in the pre-intervention period, we exactly matched on 15 MDRG codes (instead of 27).

[f] Variables on which we matched through a propensity score, which captures the relationship between beneficiaries' characteristics and their likelihood of being in the treatment group. In addition to the variables shown, we also matched on the number of months with Part A and B coverage 0 to 3 months and 4 to 12 months before a beneficiary's enrollment or pseudo-enrollment date

[g] Health Indicators Warehouse (2014a).

[h] Health Indicators Warehouse (2014c).

[i] The chronic condition flags are calculated using one to three years of claims before the enrollment or pseudo-enrollment date (depending on the condition), using the Chronic Conditions Data Warehouse definitions.

[j] Chronic Conditions Warehouse (2014b, Table B.2).

[k] Health Indicators Warehouse (2014b).

[l] Gerhardt et al. (2014).

[m] Percentage of beneficiaries with any expenditures for primary care services in the 3 months before enrollment (or 4 to 12 months before enrollment).

[n] Boards of Trustees (2013).

* Significantly different from zero at the .10 level, two-tailed test. No differences were significantly different from zero at the .05 or .01 levels.

AGH = Atlantic General Hospital; CHF = congestive heart failure; CKD = chronic kidney disease; CMS = Centers for Medicare & Medicaid Services; COPD = chronic obstructive pulmonary disease; ED = emergency department; ESRD = end-stage renal disease; FFS = fee-for-service; GI = gastrointestinal; HCC = Hierarchical Condition Category; MDC = major diagnostic category; MDRG = modified diagnosis-related group; PRMC = Peninsula Regional Medical Center; SD = standard deviation.

NA = not available.

n.a. = not applicable.

C. Equivalence of the treatment and comparison groups at baseline

Demonstrating that the treatment and comparison groups are similar at the start of the intervention is critical for the evaluation design. This similarity increases the credibility of a key assumption underlying the difference-in-differences framework—that the change in outcomes for the pre and post-intervention comparison cohorts is the same that would have happened for the pre and post-intervention treatment cohorts had the intervention not occurred.

Post-intervention equivalence. Panel A of Table IV.2 shows that the treatment and comparison beneficiaries in the post-intervention period were similar at baseline (that is, before enrollment or pseudo-enrollment). By construction, there were no differences between the two groups on the exact matching variables: gender, date of discharge, and the reason for enrollment. There were some differences between the treatment group beneficiaries and matched comparison group beneficiaries on the variables we matched through propensity scores, but the standardized differences across the propensity-score matching variables are all well below our target of 0.25 standardized differences, and even within 0.10 standardized differences (the 0.25 target is an industry standard; for example, see Institute of Education Sciences 2014).

The propensity matching technique improved or did not affect the balance for most variables, but worsened the balance for a few. This can be seen in panel A of Table IV.2, which shows the means for the full comparison pool and for the selected comparison group. Key to our approach was improving balance on the reason for hospitalization (by MDRG or MDC), and the approach successfully removed *all* imbalance on this characteristic. Matching also improved the balance for other variables, particularly when the variables were not balanced before matching (such as zip code poverty rate, original reason for Medicare entitlement, and the number of unplanned readmissions, hospitalizations, and ED visits 4 to 12 months before enrollment) because those variables had relatively more predictive power in the propensity-score model. The improvements in balance on some variables came at the expense of small increases in the differences between the treatment and comparison beneficiaries on (1) the percentage with cancer, diabetes, COPD, and Alzheimer's disease or related disorders; (2) the number of hospitalizations in the 3 months before enrollment; and (3) Medicare spending 4 to 12 months before enrollment. However, as mentioned earlier, the imbalance for all these variables was less than 0.10 standard deviations after matching.

Pre-intervention equivalence. Panel B of Table IV.2 shows that the treatment and comparison beneficiaries in the pre-intervention period were also similar at baseline (that is, at pseudo-enrollment). We were able to exactly match comparison beneficiaries on gender, date of pseudo-enrollment, and reason for hospitalization. Some differences between the treatment group beneficiaries and matched comparison group beneficiaries remained after matching on the variables we matched through propensity scores, but the standardized differences across the propensity score matching variables were all well below our target of 0.25 standardized differences, and even within 0.15 standardized differences.

D. Beneficiary outcomes and intervention impacts

In this section, we first present sample sizes and mean outcomes, by quarter, for the pre-intervention and post-intervention treatment and comparison groups. These mean outcomes provide context for understanding the difference-in-differences estimates that follow; however, the differences in mean outcomes are not regression-adjusted and are not impact estimates by themselves. Next, we present the results of the primary tests, by domain. We then present the results of the secondary tests (robustness checks) and assess whether the primary test results are plausible given the secondary test results and the implementation evidence. We end with conclusions about program impacts in each domain.

1. Sample sizes

Post-intervention cohort. In the first intervention quarter (I1), the treatment group includes 638 treatment group beneficiaries and 2,232 comparison group beneficiaries. This is the same matched sample as shown in Table IV.2, panel A. The sample decreases to 376 treatment group beneficiaries and 1,280 comparison beneficiaries in the second intervention quarter (I2). This drop in sample occurs because (1) some treatment or comparison group members exited the sample due to death or becoming unobservable; and (2) if any member of a matched set dropped from the sample, we—in accord with the sample definitions—dropped all remaining members of the matched set. The sample sizes are smaller for the follow-up ambulatory care visit and the readmission outcomes than for the other outcomes because each sample is limited to beneficiaries where the hospital discharge that led to enrollment or pseudo-enrollment met the criteria for an index stay for the measure (see Appendix 1). The sample for the follow-up ambulatory care visit outcome was limited to 505 treatment and 1,729 comparison beneficiaries meeting the outcome-specific inclusion criteria, and the sample for readmissions was limited to 464 treatment and 1,573 comparison beneficiaries (Table IV.3). For the service use and spending outcomes, the sample sizes include all beneficiaries in the treatment and comparison groups (Table IV.4).

Pre-intervention cohort. The treatment group in I1 included 226 beneficiaries, and the comparison group included 1,008 beneficiaries (Table IV.4). This is smaller than the I1 sample for the post-intervention cohort, largely because the intake period for qualifying discharges was shorter for the pre-intervention cohort (365 days from July 1, 2011, to June 30, 2012) than for the post-intervention cohort (849 days from February 1, 2013, to May 31, 2015). As with the post-intervention cohort, and for the same reasons, the sample size drops from I1 to I2 for both the treatment and comparison groups. The sample for the follow-up ambulatory care visit outcome was limited to 163 treatment and 695 comparison beneficiaries meeting the outcome-specific inclusion criteria, and the sample for readmissions was limited to 147 treatment and 622 comparison beneficiaries (Table IV.3).

2. Mean outcomes for the treatment and comparison groups, by domain and quarter

Quality-of-care processes. The follow-up ambulatory care visit rate in the 14 days following the (index) inpatient admission associated with the beneficiary's enrollment or pseudo-enrollment for the comparison group members was 68.2 percent in the pre-intervention cohort and 67.8 percent in the post-intervention cohort (Table IV.3). The rate was moderately lower (by

1.2 percentage points) for the treatment group than the comparison group in the pre-intervention cohort, but higher (5.7 percentage points) for the treatment group than the comparison group in the post-intervention cohort.

Table IV.3. Unadjusted mean outcomes (quality-of-care processes and outcomes), by cohort and treatment status

	Inpatient admissions followed by an ambulatory care visit with a primary care or specialist provider within 14 days						30-day unplanned hospital readmissions					
	Number of Medicare FFS beneficiaries[a]			Rate (%)			Number of Medicare FFS beneficiaries[b]			Rate (%)		
Quarter	T	C (unweighted)	C (weighted)	T	C	Diff (%)	T	C (unweighted)	C (weighted)	T	C	Diff (%)
Pre-intervention cohort												
I1	163	695	163	68.1	68.2	-0.1 (1.2%)	147	622	147	10.9	11.7	-0.8 (-6.7%)
Post-intervention cohort												
I1	505	1,729	505	73.5	67.8	5.7 (8.4%)	464	1,573	464	11.6	11.4	0.3 (2.4%)

Sources: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at CMS.

Note: This table measures the follow-up ambulatory care visit rate in the 14 days following the (index) inpatient admission associated with the beneficiary's enrollment or pseudo-enrollment (that is, the stay that qualified a person for the treatment or comparison group). Similarly, it measures the readmission rate in the 30 days following the (index) inpatient admission associated with the beneficiary's enrollment or pseudo-enrollment.

The means are weighted: each treatment group beneficiary received a weight of 1; each comparison beneficiary received a weight equal to the reciprocal of the total number of comparison beneficiaries who matched to the same treatment beneficiary. The post-intervention cohort included beneficiaries whose enrollment or pseudo-enrollment dates were from February 1, 2013, to May 31, 2015, and the pre-intervention cohort included beneficiaries whose enrollment or pseudo-enrollment dates were from July 1, 2011, to June 30, 2012.

[a] The sample sizes are smaller for the follow-up ambulatory care visit rate (than the service use and spending outcomes) because the sample is limited to beneficiaries whose qualifying hospital discharges met the criteria for an index stay for the measure (see Appendix 1).

[b] The sample sizes are smaller for the readmission outcome (than the service use and spending outcomes) because the sample is limited to beneficiaries whose qualifying hospital discharges met the criteria for an index stay for the 30-day readmission measure (see Appendix 1).

C = comparison group; CMS = Centers for Medicare & Medicaid Services; Diff = difference; FFS = fee-for-service; I1 = first intervention quarter; T = treatment group.

Table IV.4. Unadjusted mean outcomes (service use and spending) measured for all Medicare FFS beneficiaries, by cohort, treatment status, and quarter

Quarter	Number of Medicare FFS beneficiaries [a]			All-cause inpatient admissions (#/1,000 beneficiaries/quarter)			Outpatient ED visit rate (#/1,000 beneficiaries/quarter)			Medicare Part A and B spending ($/beneficiary/month)		
	T	C (unweighted)	C (weighted)	T	C	Diff (%)	T	C	Diff (%)	T	C	Diff (%)
Pre-intervention cohort												
I1	226	1,008	226	385.0	349.1	35.8 (10.3%)	378.8	366.6	12.2 (3.3%)	$5,662	$3,949	$1,712 (43.4%)
I2	116	481	116	232.8	184.3	48.4 (26.3%)	362.1	258.3	103.7 (40.2%)	$2,586	$2,117	$469 (22.1%)
Post-intervention cohort												
I1	638	2,232	638	319.7	323.5	-3.8 (-1.2%)	352.7	349.2	3.4 (1.0%)	$4,229	$4,393	$-163 (-3.7%)
I2	376	1,280	376	138.3	195.7	-57.4 (-29.3%)	297.9	254.1	43.7 (17.2%)	$1,755	$2,215	$-461 (-20.8%)

Sources: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at CMS.

Note: The quarters are three-month periods after a beneficiary's enrollment date (treatment group) or pseudo-enrollment date (comparison group); that is, the first intervention quarter (I1) is the first three months after enrollment or pseudo-enrollment, and the second intervention quarter (I2) is months four to six. The means are weighted: each treatment group beneficiary received a weight of 1; each comparison beneficiary received a weight equal to the reciprocal of the total number of comparison beneficiaries who matched to the same treatment beneficiary. The post-intervention cohort included beneficiaries whose enrollment or pseudo-enrollment dates were from February 1, 2013, to May 31, 2015, and the pre-intervention cohort included beneficiaries whose enrollment or pseudo-enrollment dates were from July 1, 2011, to June 30, 2012.

[a] The sample sizes are smaller in I2 than I1 because (1) some treatment or comparison group members exited the sample due to death or becoming unobservable; and (2) if any member of a matched set dropped from the sample, we—per the sample definitions—dropped all remaining members of the matched set.

C = comparison group; CMS = Centers for Medicare & Medicaid Services; Diff = difference; ED = emergency department; FFS = fee-for-service; T = treatment group.

n.a. = not applicable.

Quality-of-care outcomes. The 30-day unplanned readmission rate for the comparison group members was 11.7 percent in the pre-intervention cohort and 11.4 percent in the post-intervention cohort (Table IV.3). The readmission rate was moderately lower (by 0.8 percentage points) for the treatment group than the comparison group in the pre-intervention cohort, but moderately higher (0.3 percentage points) for the treatment group than the comparison group in the post-intervention cohort.

Service use. For both the pre- and post-intervention cohorts, the mean hospitalization rates and outpatient ED visit rates in I1 and I2 for the comparison group were relatively high (for example, 323.5 all-cause admissions and 349.2 ED visits per 1,000 beneficiaries per quarter in I1 for the post-intervention cohort), signaling that patients remain vulnerable to acute events in the six months after hospital discharge (Table IV.4). The hospitalization rates for the treatment group were 10 to 26 percent *higher* than the comparison group in the pre-intervention cohorts, but 1 to 29 percent *lower* than the comparison group in the post-intervention cohort. In contrast, outpatient ED visits were similar for the treatment and comparison groups in I1 for both the pre- and post-intervention cohorts. The treatment group's rate in I2 was much higher (40 percent) than the comparison group's rate in the pre-intervention cohort, but was also higher (17 percent) than the comparison group's rate in I2 for the post-intervention cohort.

Spending. Medicare spending for the comparison group was higher in I1 than in I2, for both the pre- and post-intervention cohorts ($3,949 per beneficiary per month and $4,393 in I1 compared with $2,117 and $2,215 in I2, respectively). Spending was 22 to 43 percent *higher* in the treatment group than the comparison group in the pre-intervention cohort, but 4 to 21 percent *lower* in the post-intervention cohort (Table IV.4).

Because we were not able to match at the hospital level for this awardee, we expected to observe differences in outcomes between AGH's treatment group and the comparison group in the pre-intervention period. These differences likely reflected real, preexisting differences between the two groups in patients' post-hospitalization outcomes, which might stem from differences in post-hospitalization care and other factors. (In the post-intervention period, 73 percent of the comparison group beneficiaries were discharged from PRMC and the rest were discharged from AGH, and all beneficiaries in the treatment group were discharged from AGH. We are less concerned about differences between treatment and comparison group beneficiaries, because we matched carefully on demographics, diagnoses, reasons for hospitalization, and other variables.) Our evaluation relies on the difference-in-differences regression model to cancel out these preexisting differences between the treatment and comparison groups. That is, the model subtracts pre-intervention differences in outcomes (between the treatment and comparison groups) from the post-intervention differences in outcomes, under the assumption that the pre-intervention differences would have persisted in the post-intervention period if the HCIA program had not existed.

3. Results for primary tests, by domain

Overview. For three of the study domains—quality-of-care processes, service use and spending—the regression-adjusted differences between the treatment and comparison groups were favorable and statistically significant (Table IV.5). In contrast, in the quality-of-care outcomes domain, we found substantively large and *unfavorable* differences between the treatment and comparison groups. The large standard error for the estimate in this domain, however, means that the unfavorable impact was estimated imprecisely.

Quality-of-care processes. The follow-up ambulatory care visit rate in the 14 days following enrollment was 73.5 percent, 5.9 percentage points higher than the estimate of the counterfactual implied by the difference-in-differences regression model. (The estimate of the counterfactual—the outcome the treatment group members would have had in the absence of the intervention—is the treatment group mean minus the regression-adjusted difference-in-differences estimate.) This was an 8.8 percent difference and was statistically significant with a one-sided test (p = .097). The statistical power to detect substantively large effects was good (83 percent) for this measure.

Quality-of-care outcomes. The treatment group's 30-day unplanned readmission rate following enrollment was 11.6 percent, 1.9 percentage points higher than the estimate of the counterfactual. This was an 18.9 percent difference, which is large enough to be considered substantively large because it was larger than the substantive threshold of 15 percent. We cannot conclude whether these unfavorable results are statistically significant because our one-sided statistical tests tested only for improvements in outcomes. The statistical power to detect effects the size of the substantive threshold was poor for the 30-day unplanned readmissions rate (20.0 percent).

Service use. The treatment group averaged 229 all-cause inpatient admissions per 1,000 beneficiaries per quarter over the first two quarters following the beneficiary's enrollment date, which was estimated to be 72 admissions fewer than the counterfactual—a difference of about 24 percent. The favorable difference between the estimates of treatment group mean and the counterfactual was substantively large, but not statistically significant after accounting for multiple comparisons in the domain (p = .113). The large difference is because the treatment group's hospitalization rate was lower than that of the comparison group during the intervention period, but higher than that of the comparison group in the pre-intervention period, leading to a large difference-in-differences estimate. The rate of outpatient ED visits (per 1,000 beneficiaries per quarter) for the treatment group was similar to the estimated counterfactual. Specifically, the treatment group's rate of outpatient ED visits was 325 in I1 and I2, which was 19 visits fewer than the estimated counterfactual (a difference of about 5.5 percent) and not statistically significant (p = .465). The combined estimate across the two measures in the service use domain was -14.7 percent, a favorable point estimate that was statistically significant (p = .095) but just smaller than substantive threshold of -15 percent. The statistical power to detect substantively large effects was poor for the two measures individually (36 to 38 percent) and, in addition, combined across the measures (52 percent). (Statistically significant results were obtained because the effect on all-cause inpatient admissions was much larger than the substantive threshold.)

Table IV.5. Results of primary tests for AGH's care transitions component

Primary test definition					Statistical power to detect an effect that is[a]		Results			
Domain (# of test in domain)	Outcome (units)	Time period for impacts	Population	Substantive threshold (expected direction of the effect)[b]	Size of the substantive threshold	Twice the substantive threshold[c]	Treatment group mean	Regression-adjusted difference between treatment group mean and the counterfactual[b] (standard error)	Percentage difference[d]	*p*-value[e]
Quality-of-care process (1)	Inpatient admissions followed by an ambulatory care visit with a primary care or specialist provider within 14 days (binary [yes or no]/discharge)	The 14 days immediately following the enrollment admission[f]	All observable Medicare FFS beneficiaries attributed to the treatment group with a qualifying enrollment admission[f] (index stay)	15.0% (+)	82.8	99.9	73.5	5.9* (4.6)	8.8%	0.097
Quality-of-care outcomes (1)	30-day unplanned hospital readmissions (binary [yes or no]/discharge)	The 30 days immediately following the enrollment admission[f]	All observable Medicare FFS beneficiaries attributed to the treatment group with a qualifying enrollment admission[f] (index stay)	15.0% (-)	20.0	34.5	11.6	1.9 (3.3)	18.9%	0.711
Service use (2)	All-cause inpatient admissions (#/1,000 beneficiaries/ quarter)	Average over the first two quarters immediately following the enrollment admission[f]	All observable Medicare FFS beneficiaries attributed to treatment group	15.0% (-)	37.8	74.6	229	-72 (46)	-23.9%	0.113[g]
	Outpatient ED visit rate (#/1,000 beneficiaries/ quarter)	Average over the first two quarters immediately following the enrollment admission[f]	All observable Medicare FFS beneficiaries attributed to treatment group	15.0% (-)	36.1	71.5	325	-19 (56)	-5.5%	0.465[g]

Table IV.5 *(continued)*

Primary test definition					Statistical power to detect an effect that is[a]		Results			
Domain (# of test in domain)	Outcome (units)	Time period for impacts	Population	Substantive threshold (expected direction of the effect)[b]	Size of the substantive threshold	Twice the substantive threshold[c]	Treatment group mean	Regression-adjusted difference between treatment group mean and the counterfactual[b] (standard error)	Percentage difference[d]	*p*-value[e]
	Combined (%)	Average over the first two quarters immediately following the enrollment admission[f]	All observable Medicare FFS beneficiaries attributed to treatment group	15.0% (-)	52.1	91.7	n.a.	n.a.	-14.7%*[h]	0.095
Spending (1)	Medicare Part A and B spending ($/beneficiary/ month)	Average over the first two quarters immediately following the enrollment admission[f]	All observable Medicare FFS beneficiaries attributed to treatment group	11.6% (-)	43.1	82.5	2,992	-1,333*** (453)	-30.8%	0.002

Sources: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at CMS.

Notes: The results for each outcome are based on a difference-in-differences regression model that included one or two intervention quarter observations per beneficiary, as described in the text. For each quarter, the model calculated the regression-adjusted difference between outcomes for post-intervention period beneficiaries assigned to the treatment and comparison groups that quarter, subtracting out any differences between the pre-intervention treatment and comparison groups during the same intervention quarter. For three outcomes, the impact estimates from the first and second intervention quarters were averaged to obtain an average impact estimate for the first two quarters. The quarters are 91- or 92-day increments after the date of a discharge from AGH or PRMC that led to a beneficiary being assigned to the treatment or comparison group. For example, if a treatment beneficiary was discharged from AGH on July 15, 2013, and subsequently enrolled in the program on July 16, 2013, his or her first intervention quarter was July 16 through October 15, 2013; his or her second intervention quarter was October 16, 2013, through January 15, 2014. The estimates were adjusted for any differences in beneficiary-level covariates (defined in Section IV.A.4) in each intervention quarter, and for indicators for each matched set (a treatment beneficiary plus his or her matched comparison beneficiaries) for each quarter.

Table IV.5 *(continued)*

The treatment and comparison groups were limited to beneficiaries enrolled in FFS Medicare for each of the four quarters before the enrollment or pseudo-enrollment date. Furthermore, in each intervention quarter, the sample consisted of Medicare FFS beneficiaries who were (1) enrolled early enough to be potentially followed up for all 91 or 92 days in the quarter and (2) whose outcomes were observable for at least one day during the quarter. Outcomes were observable if the beneficiary is alive, enrolled in Medicare FFS (Part A and B), and has Medicare as his or her primary payer of medical bills. Outcomes were constructed through November 30, 2015. In each regression model, comparison group beneficiaries were weighted based on the number of matched comparisons per treatment beneficiary. For example, if four comparison beneficiaries were matched to one treatment beneficiary, each of the four comparison beneficiaries had a weight of 0.25. If either the treatment group beneficiary or *any* of the matched comparison group members in a matched set were not observable in a quarter, any remaining beneficiaries in the matched set were removed from the sample in that quarter.

[a] Statistical power is the probability of concluding that the program had a statistically significant favorable effect when the true effect was of the specified size. The power calculation was based on actual standard errors from analysis. For example, in the first row, a 15.0 percent effect on the follow-up ambulatory care visit rate (from the estimated counterfactual of 73.5 - 5.9 = 69.6 percent) would be a change of 10.1 percentage points. Given the standard error of 4.6 percent from the regression model, we would be able to detect a statistically significant result 82.8 percent of the time if the impact was truly 10.1 percentage points, assuming a one-sided statistical test at the $p = 0.10$ significance level.

[b] The substantive threshold is the impact as percentage of the counterfactual. The counterfactual is the outcome the treatment group would have had in the absence of the HCIA-funded intervention. Our estimate of the counterfactual is the treatment group mean minus the regression-adjusted difference-in-differences estimate.

[c] We show statistical power to detect a very large effect (twice the size of the substantive threshold) because this provided additional information about the likelihood that we would find effects if the program was indeed effective. If power to detect effects is less than 75 percent even for a very large effect, then the evaluation is extremely poorly powered for that outcome.

[d] Percentage difference is calculated as the regression-adjusted difference-in-differences estimate, divided by the estimate of the counterfactual.

[e] *p*-values test the null hypothesis that the regression-adjusted difference-in-differences estimate is greater than or equal to zero (a one-sided test).

[f] The enrollment admission is the inpatient discharge that led to a beneficiary being assigned to the treatment or comparison group.

[g] We adjusted the *p*-values from the primary test results for the multiple (two) comparisons made within the service use domain.

[h] The standard error for the combined percentage difference for the outcomes in the service use domain was 11.2 percentage points.

*/**/*** Significantly different from zero at the .10/.05/.01 levels, one-tailed test, respectively. No difference-in-differences estimates were significantly different from zero at the .05 level.

AGH = Atlantic General Hospital; CMS = Centers for Medicare & Medicaid Services; ED = emergency department; FFS = fee-for-service; HCIA = Health Care Innovation Awards; p.p. = percentage points; PRMC = Peninsula Regional Medical Center.

n.a. = not applicable.

Spending. Medicare Part A and B spending for the treatment group averaged $2,992 per beneficiary per month over the first two quarters following the beneficiary's enrollment date, which was estimated to be $1,333 lower than the counterfactual. This favorable difference is statistically significant (p = .002), and large (31 percent of the estimated counterfactual). The large difference is because the treatment group's spending was lower than the comparison group's during the intervention period, but higher in the pre-intervention period, leading to a large difference-in-differences estimate. The statistical power to detect substantively large effects was poor (43 percent), but a statistically significant estimate was obtained because the effect on spending was 2.7 times larger than the substantive threshold.

Aggregate estimates for CMMI's core measures. The estimates presented for the CMMI core outcomes—that is, for 30-day unplanned readmissions, all-cause inpatient admissions, the outpatient ED visit rate, and Medicare Part A and B spending—have so far been expressed as a percentage (for readmissions), 1,000 beneficiaries per quarter (for service use outcomes), or per beneficiary per month (for spending). Table IV.6 translates these estimates into estimates of aggregate impacts during the two-quarter long primary test period. We calculated these aggregate impacts by multiplying the point estimates by the average number of Medicare beneficiaries in the post-intervention treatment group and by the number of quarters or months during the primary test periods. The aggregate estimates give a sense of the scale of the impacts, given the number of beneficiaries in the treatment group. Most notably, the results in Table IV.6 indicate the care transitions component decreased Medicare Part A and B spending by $5.4 million, which is larger than the $1.1 million HCIA award from CMMI to AGH. The point estimates in Table IV.6 for the other outcome measures should be interpreted with caution, because the estimates are not statistically significant (the p-values for these aggregate estimates are the same as they are for the main results shown in Table IV.5).

4. Results for secondary tests

The results for the secondary tests were similar to those for the primary tests (Table IV.7). There were statistically significant differences for the follow-up ambulatory care visit rate, all-cause inpatient admissions, and Medicare Part A and B spending, similar to the primary tests. For the readmission rate and outpatient ED visit rate, the differences were similar and were not statistically significant. The primary test results were plausible given these secondary tests; the secondary tests suggest that the care coordination component of AGH's program was not a major factor in the primary test results.

5. Consistency of quantitative estimates with implementation findings and results for intermediate effects on provider behavior

Based on the implementation findings (Section III.D), it is plausible that AGH's care transitions program had its intended effect on patient outcomes. The care transitions program component was implemented as planned, with some minor challenges and process improvements along the way. The impact estimates in the primary tests showed favorable, statistically significant effects in the hypothesized direction for quality-of-care processes, service use, and spending (although not necessarily of the same magnitude as AGH intended).

Table IV.6. Results for primary tests for CMMI's core outcomes expressed as aggregate effects for all Medicare FFS beneficiaries in the treatment group

Outcome (units)	Aggregate impact estimate during the primary test period (the first two quarters, or 6 months, immediately following the enrollment admission[a])	*p*-value (one-sided)
30-day unplanned readmissions (#)	+9	0.711
All-cause inpatient admissions (#)	-98	0.113
Outpatient ED visits (#)	-26	0.465
Medicare Part A and B spending ($)	-$5,423,580	0.002

Source: Authors' calculation, based on analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at the Centers for Medicare & Medicaid Services.

Note: To estimate the aggregate impact during the primary test period (intervention quarters 1 and 2) we (1) multiplied the per beneficiary per quarter (or month) estimate from Table IV.5 by the average number of Medicare FFS beneficiaries in the post-intervention treatment group during the two primary test quarters (N = 678), then (2) scaled the estimate to the 6-month primary test period by multiplying the resulting product by 2 (or 6). For the readmissions measure, the aggregate estimate in I1 was multiplied by the number of Medicare FFS beneficiaries in the post-intervention treatment group, after limiting the sample to beneficiaries whose qualifying hospital discharges met the criteria for an index stay for the 30-day readmission measure (N = 464). We obtained similar results when we calculated aggregate effects using an alternative method that allowed difference-in-differences and sample sizes to vary by quarter. The *p*-values are taken from Table IV.5 and are therefore one-sided (testing that the program improved outcomes) and adjusted for multiple comparisons conducted within each outcome domain.

[a] The enrollment admission is the inpatient discharge that led to a beneficiary being assigned to the treatment or comparison group.

CMMI = Center for Medicare & Medicaid Innovation; ED = emergency department; FFS = fee-for-service.

The substantively large unfavorable impact estimate for the quality-of-care outcome is surprising, but not implausible. First, as we noted earlier, our statistical power to detect substantive effects on 30-day readmissions was poor, so it is plausible to observe large unfavorable impact estimates due to chance. Second, it is possible the care transitions component could have increased the 30-day unplanned readmission rate, even if the program decreased all-cause admissions over a longer measurement period (I1 to I2). For example, the care coordinator may have heard a participant report a health concern or identified signs of unstable conditions or an acute exacerbation through weekly monitoring. In these cases, the protocol called for the care coordinator to refer participants to their PCP or the ED, which may have increased the chance of an unplanned, but necessary readmissions. That is, close surveillance aimed at catching emerging medical issues early might also have resulted in increased short-term use or shifts in the timing of use.

Table IV.7. Results of secondary tests for AGH's care transitions component: Robustness checks

Secondary test definition (Robustness checks)				Results			
Domain	Outcome (units)	Time period for impacts	Population	Treatment group mean	Regression-adjusted difference between treatment group mean and the counterfactual (standard error)	Percentage difference[a]	*p*-value[b]
Quality-of-care process	Inpatient admissions followed by an ambulatory care visit with a primary care or specialist provider within 14 days (binary [yes or no]/discharge)	The 14 days immediately following the enrollment admission[c]	All observable Medicare FFS beneficiaries attributed to the treatment group with a qualifying enrollment admission (index stay)[c] **who were not enrolled in the care coordination component of AGH's intervention**	73.6	6.6* (4.6)	9.8%	0.075
Quality-of-care outcomes	30-day unplanned hospital readmissions (binary [yes or no]/discharge)	The 30 days immediately following the enrollment admission[c]	All observable Medicare FFS beneficiaries attributed to the treatment group with a qualifying enrollment admission (index stay)[c] **who were not enrolled in the care coordination component of AGH's intervention**	11.4	1.8 (3.4)	18.8%	0.704
Service use	All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	Average over the first two quarters immediately following the enrollment admission[c]	All observable Medicare FFS beneficiaries attributed to the treatment group **who were not enrolled in the care coordination component of AGH's intervention**	225	-73* (47)	-24.4%	0.059
	Outpatient ED visit rate (#/1,000 beneficiaries/quarter)	Average over the first two quarters immediately following the enrollment admission[c]	All observable Medicare FFS beneficiaries attributed to the treatment group **who were not enrolled in the care coordination component of AGH's intervention**	310	-30 (56)	-8.7%	0.298
Spending	Medicare Part A and B spending ($/beneficiary/month)	Average over the first two quarters immediately following the enrollment admission[c]	All observable Medicare FFS beneficiaries attributed to the treatment group **who were not enrolled in the care coordination component of AGH's intervention**	2,971	-1,304*** (456)	-30.5%	0.002

Source: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at CMS.

Table IV.7 *(continued)*

Notes: The analyses in Table IV.7 were conducted in the same way as the analyses in Table IV.5, except excluding 25 beneficiaries in the post-intervention treatment group who were enrolled in the care coordination component of AGH's program, and their 81 matched comparison beneficiaries.

[a] Percentage difference was calculated as the regression-adjusted difference-in-differences estimate, divided by the estimate of the counterfactual. The counterfactual is the outcome the treatment group would have had in the absence of the HCIA-funded intervention. Our estimate of the counterfactual is the treatment group mean minus the regression-adjusted difference-in-differences estimate.

[b] *p*-values test the null hypothesis that the regression-adjusted difference-in-differences estimate is greater than or equal to zero (a one-sided test). The *p*-values from the secondary test results were not adjusted for multiple comparisons within or across domains.

[c] The enrollment admission is the inpatient discharge that led to a beneficiary being assigned to the treatment or comparison group.

*/**/*** Significantly different from zero at the .10/.05/.01 level, one-tailed test. The *p*-values from the secondary test results were not adjusted for multiple comparisons within each domain or across domains.

AGH = Atlantic General Hospital; CMS = Centers for Medicare & Medicaid Services; ED = emergency department; FFS = fee-for-service; p.p. = percentage points.

6. Conclusions about impacts on patient outcomes within each domain

Based on all evidence currently available, we have drawn the following conclusions about program impacts in each domain during the primary test period. Table IV.8 summarizes these conclusions and their support.

- **The care transitions component had a statistically significant *favorable* effect on quality-of-care processes, service use, and spending.** The primary test(s) for each of these domains were all favorable and statistically significant, indicating favorable impacts on the 14-day follow-up ambulatory care visit rate, the average of the two outcomes in the service use domain (driven by a large favorable estimate for all-cause admissions), and Medicare Part A and B spending. The point estimate for the 14-day follow-up impact was about one-half the size of the substantive threshold (8.8 versus 15 percent), whereas the other point estimates were about the size of the substantive threshold (service use) or greater (spending). The secondary tests confirmed the plausibility of the primary tests; implementation findings indicate it is plausible that the care transition component was implemented in a manner that could have affected the outcomes in this way.
- **The care transitions component had a substantively large *unfavorable* effect on quality-of-care outcomes.** The primary test results showed a substantively large unfavorable estimate for the one outcome in the quality-of-care outcome domain: the 30-day unplanned readmission rate. However, the standard error for the primary test was large. Therefore, we have low confidence in the conclusion of substantively unfavorable impacts. Although the program may have increased the 30-day unplanned readmission rate (or decreased it), it is possible that the large observed point estimate was due to chance, rather than to true unfavorable impacts.

Table IV.8. Conclusions about the impacts of AGH's care transitions component on patient outcomes, by domain

		Evidence supporting conclusion		
Domain	Conclusion	Primary test result(s) that supported conclusion	Primary test result plausible given secondary tests?	Primary test result plausible given implementation evidence?
Quality-of-care process	Statistically significant favorable effect	Estimate for 14-day ambulatory care follow-up visit measure was favorable and statistically significant	Yes	Yes
Quality-of-care outcomes	Substantively important unfavorable effect	The estimated for 30-day unplanned readmissions measure is unfavorable and substantively important	Yes	Yes
Service use	Statistically significant favorable effect	Combined estimate was favorable and statistically significant	Yes	Yes
Spending	Statistically significant favorable effect	Estimate for Medicare Part A and B spending was favorable and statistically significant	Yes	Yes

Sources: Table IV.5 and Table IV.7.

V. DISCUSSION AND CONCLUSIONS

AGH used its $1.1 million dollar HCIA to implement a PCMH intervention in its hospital and seven primary care practices. The program aimed to reduce hospital admissions, ED visits, and total spending by helping participants manage their conditions. This report describes and estimates the impacts of one key component of the intervention: care transitions following a hospital discharge. Although the care transitions program was implemented largely as planned, it is important to remember the care transitions program was only a small part of the overall PCMH model that AGH implemented.

The results from our impact evaluation suggest that the care transitions program improved patient outcomes in three of the four evaluation domains: quality-of-care processes, service use, and spending. The impact estimates for the outcomes in these three domains were favorable and statistically significant; these results are plausible, given that AGH implemented the component successfully. The significant improvements in 14-day follow-up visits suggest that this was an important mechanism for the declines in hospitalizations and spending, as anticipated in the awardee's theory of action. We did not have direct measures of the other anticipated changes in intermediate outcomes, such as patient adherence to medications, that could mediate the overall program impacts on service use and spending.

We estimate that the program reduced Medicare Part A and B spending for the 638 FFS Medicare beneficiaries in the treatment group by $5.4 million from the care transitions intervention by itself. This reduction in Medicare spending is more than the cost of the total AGH award (which funded two additional components [care coordination and Keeping in Touch] and care transitions for additional patient populations, including Medicaid beneficiaries). Our difference-in-differences estimate captures the changes in CMS spending on Medicare Part A and B claims for patients in the treatment group in the six months following enrollment in care transitions.

Although the results indicate there may have been a large substantively important unfavorable effect on the one outcome in the quality-of-care outcomes domain (the 30-day unplanned hospital readmission rate), this result may have been due to chance.

Several measures capture the generally successful implementation of the program. These measures include that AGH enrolled the first patient in the care transitions component in the first quarter of 2013, met its enrollment targets and staffing targets, and completed the initially planned staff training. Key factors helping implementation included availability of resources to support the program operations, monitoring to identify process improvements, staff commitment to the program model, and Maryland's global payment model providing incentives to the hospital to improve quality of care and reduce costs. AGH made process improvements throughout the program and worked to overcome several barriers, including a lack of existing data collection and reporting infrastructure, as well as needs of participants with complex conditions and their noncompliance.

Many studies have found that care transitions programs can improve patients' outcomes (for example, see Feltner et al. 2014; Peikes et al. 2012; Vedel and Khanassov 2015), but the current

findings are new in illustrating that even a low-touch telephonic intervention in a small, rural health care system can be effective if it is well timed, builds on existing relationships between practices and their patients, and takes advantage of real-time information from the hospital about a patient's discharge instructions, medications, and treatment plan. These findings could guide the efforts of CMS, other payers, ACOs, and hospital systems such as AGH to improve efficiency and quality of care.

This page has been left blank for double-sided copying.

REFERENCES

Boards of Trustees, Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds. "2013 Annual Report of the Boards of Trustees of the Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds." Table V.D1. Washington, DC: Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds, 2013. Available at http://downloads.cms.gov/files/tr2013.pdf. Accessed August 13, 2014.

Chronic Conditions Data Warehouse. "Table A.1. Medicare Beneficiary Counts for 2003–2012." Baltimore, MD: Centers for Medicare & Medicaid Services, 2014a. Available at https://www.ccwdata.org/cs/groups/public/documents/document/ccw_website_table_a1.pdf. Accessed November 19, 2014.

Chronic Conditions Data Warehouse. "Table B.2. Medicare Beneficiary Prevalence for Chronic Conditions for 2003 Through 2012." Baltimore, MD: Centers for Medicare & Medicaid Services, 2014b. Available at https://www.ccwdata.org/cs/groups/public/documents/document/ccw_website_table_b2.pdf. Accessed November 19, 2014.

Feltner, Cynthia, Christine D. Jones, Crystal W. Cené, Zhi-Jie Zheng, Carla A. Sueta, Emmanuel J.L. Coker-Schwimmer, Marina Arvanitis, Kathleen N. Lohr, Jennifer C. Middleton, and Daniel E. Jonas. "Transitional Care Interventions to Prevent Readmissions for Persons with Heart Failure: A Systematic Review and Meta-Analysis." *Annals of Internal Medicine*, vol. 160, no. 11, 2014, pp. 774–784.

Gerhardt, Geoffrey, Alshadye Yemane, Keri Apostle, Allison Oelschlaeger, Eric Rollins, and Niall Brennan. "Evaluating Whether Changes in Utilization of Hospital Outpatient Services Contributed to Lower Medicare Readmission Rate." *Medicare & Medicaid Research Review,* vol. 4, no. 1, 2014, pp. E1–E13.

Hansen, Ben B. "Full Matching in an Observational Study of Coaching for the SAT." *Journal of the American Statistical Association,* vol. 99, no. 467, 2004, pp. 609–618.

Health Indicators Warehouse. "Average Age of Medicare Beneficiaries (Mean)." Hyattsville, MD: National Center for Health Statistics, HIW, 2014a. Available at http://www.healthindicators.gov/Indicators/Average-age-of-Medicare-beneficiaries-mean_308/Profile/ClassicData. Accessed November 19, 2014.

Health Indicators Warehouse. "Hospital Inpatient Medicare Admissions (per 1,000 beneficiaries)." Hyattsville, MD: National Center for Health Statistics, HIW, 2014b. Available at http://www.healthindicators.gov/Indicators/Hospital-inpatient-Medicare-admissions-per-1000-beneficiaries_2001/Profile/ClassicData. Accessed August 13, 2014.

Health Indicators Warehouse. "Medicare Beneficiaries Who Are Also Eligible for Medicaid (Percent)." Hyattsville, MD: National Center for Health Statistics, HIW, 2014c. Available at http://www.healthindicators.gov/Indicators/Medicare-beneficiaries-eligible-for-Medicaid-percent_317/Profile/ClassicData. Accessed August 4, 2015.

Institute of Education Sciences. "What Works Clearinghouse: Procedures and Standards Handbook, Version 3.0." Washington, DC: U.S. Department of Education, IES, 2014. Available at http://ies.ed.gov/ncee/wwc/DocumentSum.aspx?sid=19. Accessed September 15, 2014.

Moreno, Lorenzo, Boyd Gilman, Greg Peterson, Catherine DesRoches, Sheila Hoag, Linda Barterian, Laura Blue, Katherine Bradley, Emily Ehrlich, Kristin Geonnotti, Lauren Hula, Keith Kranker, Rumin Sarwar, KeriAnn Wells, Joseph Zickafoose, Sandi Nelson, Kate Stewart, and Frank Yoon. "Evaluation of the Health Care Innovation Awards (HCIA): Primary Care Redesign Programs. Quarterly Report for CMS: Second Annual Report." Report submitted to the Center for Medicare & Medicaid Innovation. Princeton, NJ: Mathematica Policy Research, December 11, 2015.

Peikes, Deborah, Stacy Dale, Eric Lundquist, Janice Genevro, and David Myers. "Building the Evidence Base for the Medical Home: What Sample and Sample Size Do Studies Need? White Paper." AHRQ Publication no. 11-0100-EF. Rockville, MD: Agency for Healthcare Research and Quality, October 2011.

Peikes, Deborah, Rebecca S. Lester, Boyd Gilman, and Randall Brown. "The Effects of Transitional Care Models on Re-Admissions: A Review of the Current Evidence." *Generations*, vol. 36, no. 4, winter 2012–2013, pp. 44–55.

Rosenbaum, Paul R. "A Characterization of Optimal Designs for Observational Studies." *Journal of the Royal Statistical Society, Series B,* 1991, pp. 597–610.

Rosenthal, M.B., S. Alidina, M. Friedberg, S. Singer, D. Eastman, Z. Li, and E. Schneider. "A Difference-in-Differences Analysis of Changes in Quality, Utilization, and Cost Following the Colorado Multi-Payer Patient-Centered Medical Home Pilot." *Journal of General Internal Medicine,* vol. 31, no. 3, March 2016, pp. 289–296.

Stuart, Elizabeth A. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science,* vol. 25, no. 1, 2010, pp. 1–21.

Van Walraven, C., I.A. Dhalla, C. Bell, E. Etchells, I.G. Stiell, K. Zarnke, P.C. Austin, and A.J. Forster. "Derivation and Validation of an Index to Predict Early Death or Unplanned Readmission After Discharge from Hospital to the Community." *Canadian Medical Association Journal,* vol. 182, no. 6, April 6, 2010, pp. 551–557.

Vedel, Isabelle, and Vladimer Khanassov. "Transitional Care for Patients with Congestive Heart Failure: A Systematic Review and Meta-Analysis." *Annals of Family Medicine*, vol. 13, no. 6, 2015, pp. 562–571.

CHAPTER 2

CAREFIRST BLUECROSS BLUESHIELD

Greg Peterson, Kristin Geonnotti, Lauren Hula, Boyd Gilman, Laura Blue, Keith Kranker, Kate Stewart, Sheila Hoag, and Lorenzo Moreno

This page has been left blank for double-sided copying.

CAREFIRST BLUECROSS BLUESHIELD

CHAPTER SUMMARY

Introduction. CareFirst BlueCross BlueShield (CareFirst) used its $20 million Health Care Innovation Award (HCIA) to extend a patient-centered medical home (PCMH) program designed for its commercial members to Medicare fee-for-service (FFS) beneficiaries in Maryland. The intervention targeted approximately 35,000 Medicare beneficiaries served by 149 primary care providers (PCPs) in 52 practices. These practices formed 14 medical panels, groups of 5 to 15 PCPs who participated in the intervention as a performance unit. CareFirst aimed to reduce total Medicare spending by 6 percent in the final intervention year (and by 3 percent in the second year) by reducing patients' need for acute care—such as inpatient admissions and emergency department visits—through care coordination and by changing PCPs' referral patterns.

Objectives. (1) To describe the design and implementation of CareFirst's HCIA-funded intervention, including the role of PCPs in the intervention and the extent to which anticipated changes in providers' behavior occurred; (2) to assess impacts of the intervention on patients' outcomes and Medicare Part A and B spending during the first three years of the award; and (3) to use both implementation and impact findings to identify possible explanations for the observed impacts.

Methods. We reviewed CareFirst's program documents and self-monitoring metrics, conducted interviews with CareFirst leadership and program staff, and surveyed participating clinicians. To estimate impacts, we compared outcomes for Medicare FFS beneficiaries served by the 14 treatment panels with outcomes for Medicare beneficiaries served by 42 matched comparison panels participating in CareFirst's commercial PCMH program (which does not serve Medicare beneficiaries), adjusting for any differences in outcomes for the two groups during a one-year baseline period.

Program design and implementation. The intervention had three components: (1) care coordination, in which nurse care coordinators hired by CareFirst worked with PCPs to develop and implement care plans for high-risk, clinically unstable patients; (2) financial incentives to panels for participating in care coordination and achieving savings and quality targets among Medicare patients; and (3) technical assistance to panels to identify opportunities for reducing spending through changing referral patterns or shifting treatment to more cost-effective settings. After an initial one-year delay, the intervention was largely implemented as planned. CareFirst hired 44 nurse care coordinators, enrolled 3,276 beneficiaries into care coordination services, provided ongoing technical assistance to panels, and paid financial incentives to panels. However, the method for targeting high-risk patients for care coordination services did not consistently identify patients who were clinically unstable, making them less likely to benefit from these services. In addition, it was sometimes difficult to sufficiently adapt care coordination strategies initially developed for the commercial population to an older Medicare population with generally more complex health needs.

Clinicians' perceptions of intervention effects on the care they provide. CareFirst's program design required PCPs to engage in care coordination—supported by nurse care coordinators—and to change their referral patterns. The available evidence suggests that CareFirst engaged PCPs as planned, with 90 percent of PCPs enrolling at least one patient into care coordination services. Further, most PCPs reported they thought the intervention improved the quality, timeliness, and safety of their care. However, no evidence is available to assess whether planned changes in referral patterns occurred.

Impacts on patients' outcomes. The impact estimates indicate that the intervention did not improve patients' outcomes in any of the four evaluation domains during the first three years of the award: quality-of-care processes, quality-of-care outcomes, service use, or spending. Specifically, there was no evidence of statistically significant or substantively large favorable effects in any domain. The statistical power to detect effects was good for the domains of quality-of-care processes and service use (outcomes include, for example, all-cause inpatient admissions, the outpatient ED visit rate, and the proportion of people discharged from the hospital who received a primary care or specialist visit within 14 days), but not for the other two evaluation domains.

Conclusion. Evaluation evidence indicates that CareFirst did not achieve its intended impacts on patients' outcomes during the original three-year award period. The lack of effects appears not be due to a failure to engage PCPs or generally implement the program as planned. Rather, the lack of effects might be due to (1) challenges identifying clinically unstable patients and adapting care coordination strategies from commercial to Medicare populations, (2) limitations in the intervention design itself, or (3) the relatively short intervention duration. Impact estimates might change after including the final six months of program operations, the period when CareFirst expected to observe the largest impacts. We plan to report final results, including these six months, in a future addendum to this report.

Summary of intervention and impact results for CareFirst

Intervention description			
Awardee description		Largest commercial health insurer in the mid-Atlantic region	
Award amount ($ millions)		$20.0	
Award extended beyond June 2015?		Yes (6 months)	
Location		Maryland, statewide (urban and suburban)	
Target population		Approximately 35,000 Medicare FFS beneficiaries (excluding those also enrolled in Medicaid) served by 149 PCPs in 52 primary care practices grouped into 14 medical panels	
Interventions		Extended a PCMH program developed for commercial members to Medicare FFS beneficiaries. The program included • Care coordination, in which 44 HCIA-funded nurses worked with PCPs to develop and implement care plans for high-risk patients • Financial incentives to (1) reward panels that reduced total spending while meeting quality targets; and (2) pay PCPs to participate in care coordination • Technical assistance to panels to identify opportunities to generate savings though changes in referrals	
Metrics of intervention delivered		• Implemented care plans for 3,276 Medicare FFS beneficiaries (almost 10% of panels' Medicare patients) • Care plans active for 260 days, on average, with roughly weekly nurse contact • Rewards from $3,000 to $494,000 to panels with spending below target in 2015	
Impact evaluation methods			
Core design		Difference-in-differences model with matched comparison group	
Treatment group	Definition	Medicare FFS beneficiaries attributed to 14 treatment panels (excluding those also enrolled in Medicaid)	
	# of beneficiaries during primary test period[a]	35,536 to 37,593	
Comparison group definition		Medicare FFS beneficiaries attributed to 42 matched comparison panels participating in CareFirst's commercial PCMH program (excluding those also enrolled in Medicaid)	
Impact results: Quality-of-care processes domain			
Ambulatory care visit within 14 days of discharge (% of beneficiaries/quarter)		Comparison mean[b]	66.0%
		Impact estimate (% difference)	+0.3 pp (+0.4%)
Received recommended lipid test, for patients with IVD (% of beneficiaries/year)		Comparison mean[b]	80.0%
		Impact estimate (% difference)	-0.8 pp (-1.0%)
Received all four recommended diabetes processes of care (% of beneficiaries/year)		Comparison mean[b]	48.5%
		Impact estimate (% difference)	-2.8 pp (-5.7%)
Combined impact estimate[c]		-2.1%	
Impact conclusion[d]		No substantively large effect	
Impact results: Quality-of-care outcomes domain			
30-day unplanned hospital readmissions (#/1,000 beneficiaries/quarter)		Comparison mean[b]	8.6
		Impact estimate (% difference)	+1.3 (+16.3%)
Inpatient admissions for ACSC conditions (#/1,000 beneficiaries/quarter)		Comparison mean[b]	11.2
		Impact estimate (% difference)	+0.4 (+3.7%)
Combined impact estimate[c]		+10.0%	
Impact conclusion[d]		**Substantively large unfavorable effect**	
Impact results: Service use domain			
All-cause inpatient admissions (#/1,000 beneficiaries /quarter)		Comparison mean[b]	70.9
		Impact estimate (% difference)	+1.9 (+2.6%)
Outpatient ED visits (#/1,000 beneficiaries/quarter)		Comparison mean[b]	85.5
		Impact estimate (% difference)	-2.6 (-3.1%)
Combined impact estimate[c]		-0.2%	
Impact conclusion[d]		No substantively large effect	
Impact results: Spending domain			
Medicare Part A and B spending ($/beneficiary/month)		Comparison mean[b]	$1,005
		Impact estimate (% difference)	+$9 (+0.9%)
Impact conclusion[d]		Indeterminate effect	

Note: See the CareFirst chapter for details on the intervention, impact methods, and impact results.

[a] Number of beneficiaries in the full treatment group across the quarters in the primary test period.

Summary of intervention and impact results for CareFirst *(continued)*

[b] The comparison mean is the estimate of the outcome the treatment group beneficiaries would have had if they had not received the intervention. It is equal to the mean for the treatment group over the intervention quarters (in the primary test period) minus the impact estimate.

[c] The combined estimate is the average across all the individual estimates in the domain, in which the impact estimate for each individual outcome is expressed as percentage change relative to the comparison group.

[d] We drew conclusions at the domain level based on the results of pre-specified primary tests, secondary tests (robustness checks), and consistency with implementation evidence. For each domain, we could draw one of five conclusions: (1) Statistically significant favorable effect (the highest level of evidence), (2) Substantively important (but not statistically significant) favorable effect, (3) Substantively important (but not statistically significant) unfavorable effect, (4), No substantively large effect, and (5) Indeterminate effect. Section V.A.8 of this report describes the decision rules we used to reach each of these possible conclusions.

*Significantly different from zero at the .10 level, one-tailed test.

**Significantly different from zero at the .05 level, one-tailed test.

***Significantly different from zero at the .01 level, one-tailed test.

ACSC = ambulatory care-sensitive condition; ED = emergency department; FFS = fee-for-service; HCIA = Health Care Innovation Award; IVD = ischemic vascular disease; PCP = primary care provider; PCMH = patient-centered medical home; pp = percentage point.

I. INTRODUCTION

This report presents findings from the evaluation of CareFirst BlueCross BlueShield's (CareFirst) Health Care Innovation Award (HCIA), with a focus on program impacts on patients' outcomes. Section II provides an overview of CareFirst's HCIA-funded intervention and the design of the impact evaluation. Section III describes the design and implementation of the intervention, including how the program could be expected to affect evaluation outcomes through changes in patients' and providers' behavior. In Section IV, we assess the evidence on the extent to which planned changes in providers' behavior occurred. Section V describes our methods for, and results and conclusions from, estimating program impacts on patients' outcomes in four domains: quality-of-care processes, quality-of-care outcomes, service use, and spending. Section VI draws conclusions by synthesizing the impact and implementation findings and describes next steps for the evaluation.

The impact estimates in this report are preliminary because they cover the original three-year award period of the HCIA (June 2012 through June 2015). Because CareFirst's HCIA program was extended beyond that date—through December 2015—we do not yet include the final six months of CareFirst's intervention. We plan to report final results, including these six months, in a future addendum to this report.

II. OVERVIEW OF CAREFIRST'S HCIA-FUNDED INTERVENTION AND THE IMPACT EVALUATION

A. CareFirst's HCIA-funded intervention

CareFirst, the largest commercial health insurer in the mid-Atlantic region (Maryland, Virginia, and Washington, D.C.), received $20 million in HCIA funding to expand its commercial patient-centered medical home (PCMH) program to Medicare fee-for-service (FFS) beneficiaries in Maryland (Table II.1, top panel). In its commercial program, which began in 2011, CareFirst provides support staff, technical assistance, and clinician incentive payments to 1,219 practices to reduce total medical spending for commercial members while improving quality of care. CareFirst grouped these 1,219 practices into 450 panels, defined as groups of 5 to 15 primary care providers (PCPs) (either physicians or nurse practitioners) who voluntarily agree to participate as a unit in terms of quality measurement and shared incentive payments. For the HCIA intervention, CareFirst selected 14 of the top-performing panels in the commercial program and extended the PCMH intervention to Medicare FFS beneficiaries, excluding those who were dually eligible for Medicaid. These 14 panels comprise 52 primary care practices that served about 35,000 Medicare FFS beneficiaries during each quarter of the award period. HCIA program services began on August 1, 2013, roughly 13 months later than originally planned, and were extended six months beyond the originally planned award end date (June 30, 2015) to end on December 31, 2015.

CareFirst's goals were to reduce hospital costs by 7.5 percent and total health care costs by 6.0 percent among Medicare FFS beneficiaries by the end of the award (Table II.1). CareFirst expected to achieve these outcomes through three intervention components: (1) care coordination for high-risk beneficiaries, (2) financial incentives to panels for participating in care coordination

services and for achieving savings and quality targets for all non-dually eligible Medicare FFS patients, and (3) technical assistance to panels' PCPs to identify opportunities for reducing spending through changing their referral patterns or shifting treatment to more cost-effective settings. CareFirst expected that these intervention components would reduce the need for hospitalizations and post-acute care among high-risk Medicare beneficiaries and encourage PCPs to refer patients to cost-effective providers and settings of care. The reductions in acute care and changes in use of specialty care were expected, in turn, to reduce total Medicare spending. (Section III.A.3 describes the awardee's theory of action in detail.)

Table II.1. Summary of CareFirst's HCIA program and our evaluation for estimating its impacts on patients' outcomes

Program description	
Award amount	$20,000,000[a]
Award start date	June 2012
Implementation date	August 1, 2013
Award end date	Original: June 2015 After no-cost extension: June 2016, with direct program services ending December 2015
Awardee description	CareFirst BlueCross BlueShield is the largest private health insurer in the Mid-Atlantic region (Maryland, Virginia, and Washington, DC).
Intervention overview	CareFirst extended a PCMH program for commercial members to Medicare FFS beneficiaries. The Medicare beneficiaries were served by 52 primary care practices grouped into 14 medical panels.
Intervention components	1. **Care coordination for high-risk beneficiaries.** LCC nurses worked with patients' PCPs to develop and implement care plans for high-risk patients. 2. **Financial incentives.** CareFirst financially rewarded panels that reduced spending for their Medicare beneficiaries while improving or maintaining quality, and paid physicians for developing and updating care plans. 3. **Technical assistance to panels.** Program consultants analyzed data to identify opportunities to reduce spending by having PCPs change their referral patterns, including recommending treatment in more cost-effective settings. Program consultants relied on web-based IT (iCentric platform, originally developed for the commercial program) to track global cost and utilization metrics, which they shared with panels.
Target population	About 35,000 Medicare FFS beneficiaries whom CareFirst attributed to the 14 participating medical panels (excluding those also enrolled in Medicaid)[b]
Target impacts on patients' outcomes	• Reduce all-cause inpatient admissions by 7.5 percent • Reduce outpatient ED visits by 7.5 percent • Reduce Medicare Part A and B spending by 6.0 percent • Improve quality-of-care process and outcome measures (amount not specified)
Workforce development	Created 49 positions fully funded by the award: 44 LCCs, 5 program consultants
Location	Maryland, statewide (urban and suburban areas)

Table II.1 *(continued)*

Impact evaluation	
Core design	Difference-in-differences with matched comparison group
Treatment group	Medicare FFS beneficiaries (excluding those also enrolled in Medicaid) whom we attributed to the treatment panels using CareFirst's attribution rules[b]
Comparison group	Medicare FFS beneficiaries (excluding those also enrolled in Medicaid) whom we attributed to 42 matched comparison panels. The comparison panels participated in CareFirst's commercial PCMH program but not its expansion to Medicare FFS beneficiaries and were matched to treatment panels on performance in the commercial program, characteristics of the panels' Medicare patients, and other panel characteristics.
Intervention component(s) included in impact evaluation	**All three components described above.** CareFirst expected the three program components to work in combination to affect outcomes for all Medicare FFS beneficiaries attributed to the treatment panels, although panels provided intensive care coordination services only to high-risk beneficiaries.
Extent to which the treatment group reflects the awardee's target population (for the component(s) evaluated)	**High**. The awardee's target population and the impact evaluation's treatment group both consist exclusively of Medicare FFS beneficiaries (excluding those also enrolled in Medicaid) attributed—according to CareFirst's attribution rules—to the treatment panels.
Study outcomes, by domain	1. **Quality-of-care processes.** Preventive care for diabetes, lipid testing for patients with IVD, and 14-day follow-up to hospitalization 2. **Quality-of-care outcomes.** 30-day unplanned readmissions and inpatient admissions for ambulatory care-sensitive conditions 3. **Service use.** All-cause inpatient admissions and outpatient ED visits 4. **Spending.** Medicare Part A and B spending

Source: Review of CareFirst reports, including its original application, operational plan, and 15 quarterly narrative reports to the Centers for Medicare & Medicaid Services.

[a] CareFirst was originally awarded $20 million to expand its PCMH program to Medicare beneficiaries in Maryland. An additional $4 million was allocated for use if CareFirst could find a partner to expand the program outside of Maryland, but this did not happen.

[b] CareFirst attributed a Medicare FFS beneficiary to a medical panel in a month if, based on claims data, the providers in that panel provided the plurality of the beneficiary's primary care services in the prior year (or prior two years if there were no primary care services in the prior year).

ED = emergency department; FFS = fee-for-service; HCIA = Health Care Innovation Award; IT = information technology; IVD = ischemic vascular disease; LCC = local care coordination; PCMH = patient-centered medical home; PCP = primary care provider.

B. Overview of impact evaluation

To estimate program impacts on patient outcomes, we compared outcomes for Medicare FFS beneficiaries served by the 14 panels participating in the HCIA intervention (treatment panels) with outcomes for beneficiaries served by 42 matched comparison panels, adjusting for any differences in outcomes between these two groups before the intervention began. Table II.1, bottom panel, summarizes our impact evaluation design. We designed the impact evaluation to estimate the marginal impact of the HCIA intervention—that is, the impact of expanding the existing PCMH model to Medicare FFS beneficiaries. While it is possible that the commercial PCMH program by itself has some positive spillover for Medicare beneficiaries that is not captured in our impact estimates, we anticipate such spillover to be small. The largest

intervention component in the PCMH program is care coordination for high-risk patients, and providing care coordination for one patient is likely to have little influence on care for other patients the panel served. Consisent with CareFirst's target population, we excluded Medicare FFS beneficiaries who were also enrolled in Medicaid from both the treatment and comparison groups.

We selected the 42 comparison panels for the evaluation from the pool of all panels in Maryland that participated in CareFirst's commercial PCMH program, but not its expansion of the program to Medicare beneficiaries. We selected panels that were similar to the 14 treatment panels in terms of their quality and financial performance in the commercial PCMH program and characteristics of their Medicare patients before the intervention began.

We estimated impacts on outcomes, as measured in Medicare FFS claims data, which we grouped into four domains: (1) quality-of-care processes, (2) quality-of-care outcomes, (3) service use, and (4) spending. Across the HCIA awardees in primary care redesign (PCR), we designed our impact evaluations to identify promising interventions or intervention components—consistent with evaluation goals from the Center for Medicare & Medicaid Innovation (CMMI) to find programs that could be scaled or retested as part of a future model test. Before conducting analyses, we specified a series of primary tests, describing the evidence we would need to conclude that the program was effective, and the awardee and CMMI reviewed these tests. Each test specified a population, outcome, period, expected direction of effect, and threshold that we count as substantively important. The purpose of these primary tests was to focus the impact evaluation on hypotheses that would provide the most robust evidence about program effectiveness. We used the results from the primary tests and robustness checks to draw conclusions about program impacts in each of the four evaluation domains. Because we sought to identify promise, rather than only those programs with unequivocally demonstrated success, we conducted one-sided statistical tests (that is, testing only for program benefits) and used a threshold for statistical significance of 0.1, which is not as strict as the conventional standard of 0.05.

Our impact evaluation design reflects the effects of all three intervention components among CareFirst's full HCIA target population; that is, impact estimates capture the effects of all three intervention components that constituted CareFirst's HCIA intervention for all attributed Medicare patients. The evaluation's treatment group includes all Medicare FFS beneficiaries (excluding those who are dually eligible for Medicaid) that the 14 treatment panels served. CareFirst expected the three intervention components—care coordination, financial incentives, and technical assistance—to work in combination to affect outcomes for all Medicare FFS beneficiaries served by treatment panels, even though the panels provided care coordination services only to high-risk patients. We used CareFirst's own attribution rules to attribute Medicare beneficiaries to treatment and comparison panels.

III. PROGRAM IMPLEMENTATION

This section first provides a detailed description of CareFirst's HCIA-funded intervention, highlighting how it evolved over time and its theory of action. Second, it assesses the evidence on the extent to which the intervention was implemented as planned based on measures of

program enrollment, service delivery, staffing, training, and timeliness. Third, the section summarize the facilitators and barriers associated with implementation effectiveness.

We based our evaluation of CareFirst's program implementation on a review of its quarterly reports to CMMI and self-monitoring program metrics, telephone discussions and follow-up communications with program administrators, and information collected during site visit interviews with frontline staff conducted in April 2014 and April 2015. We did not verify the quality of the performance data reported by CareFirst in its self-measurement and monitoring reports.

A. Program design and adaptation

1. Target population and patient identification, recruitment, and enrollment

In this section, we describe how CareFirst selected panels to participate in the HCIA intervention, identified the Medicare FFS patients these panels serve, and identified high-risk Medicare patients for care coordination services.

Identification of panels for participation. Panels are groups of 5 to 15 PCPs (either physicians or nurse practitioners) who voluntarily agree to participate as a unit in terms of quality measurement and shared incentive payments. These panels can be formed by solo or small, independent group practices that agree to work together (referred to as a virtual panel); independent group practices that already fall within the size range; or a subgroup of a large group practice. Health system-based practices, under common ownership of a hospital or health system, may also participate in the program. CareFirst selected 14 of the 450 panels in the commercial program to participate in the HCIA-funded intervention based on the following criteria. The panels must have (1) been located in Maryland, (2) joined the commercial program when it began in 2011, (3) served at least 1,000 CareFirst members in 2012, and (4) performed well—both financially and on quality measures—in the commercial program. CareFirst expected all of the PCPs working in the 14 panels to engage in the HCIA-funded intervention, including participating in care coordination services.

Target population. The target population for CareFirst's intervention was Medicare FFS beneficiaries with Parts A and B coverage for whom Medicare was their primary payer. Beneficiaries who were dually eligible for Medicare and Medicaid were not eligible for the intervention. CareFirst attributed Medicare beneficiaries to participating panels on a monthly basis. A beneficiary was attributed to a panel if the practice's PCPs provided the plurality of his or her primary care services in the previous 12 months (or in the previous 24 months if the beneficiary received no primary care services in the previous 12 months). In each program quarter, CareFirst had about 35,000 Medicare FFS beneficiaries attributed to the 14 participating panels.

Identification, recruitment, and enrollment of patients for care coordination. From the population of attributed Medicare beneficiaries, CareFirst targeted for care coordination services patients who were clinically unstable and had multiple chronic conditions and whom panel staff, using risk information from CareFirst, considered to have the highest risk of hospitalization or

other costly acute care services. CareFirst defined these participants as those who (1) had multiple chronic conditions, (2) were clinically unstable, or (3) were likely to have an acute exacerbation that might lead to either costly outpatient emergency department (ED) visits or an inpatient admission. CareFirst intended to identify these beneficiaries primarily using illness burden scores—as defined later—in addition to the clinical judgment of PCPs and local care coordination nurses (LCCs).

CareFirst stratified attributed beneficiaries into five illness bands based on their health status as measured by illness burden scores, using inpatient and outpatient diagnoses and demographic information to assess risk. These illness bands were meant to help program staff identify beneficiaries who were most in need of care coordination services; LCCs and PCPs were encouraged to target patients with the highest illness burden scores. LCCs and PCPs identified beneficiaries with high illness burden scores using iCentric (the online portal through which providers maintain care plans and track cost and utilization metrics, including illness burden scores). LCCs and PCPs also used other sources of information—in addition to, or sometimes instead of, illness burden scores—to identify patients for care coordination. This included the LCCs' and PCPs' clinical judgment about who would benefit most from care coordination services; their understanding of participants' medical and social needs; and complementary data sources (for example, the Chesapeake Regional Information System for Patients [CRISP]—Maryland's statewide health information exchange that provides real-time notifications based on admissions and discharge data to PCPs when their patients were hospitalized). LCCs and PCPs contacted eligible patients primarily by telephone or during office visits to invite them to participate in the intervention. Medicare beneficiaries verbally consented to receive care coordination services.

2. Intervention components

CareFirst's intervention had three components. Only one—care coordination for high-risk beneficiaries—delivered services directly to patients enrolled in this program component. The other two program components—financial incentives and technical assistance to panels—were delivered to PCPs and panels to help them improve the care of all attributed patients.

Care coordination for high-risk beneficiaries. Care coordination services focused on providing and implementing care plans for attributed high-risk Medicare beneficiaries with multiple chronic and/or unstable conditions and who were considered at high risk of hospitalization or other costly acute care services. Each individualized care plan, developed by LCCs in collaboration with PCPs, described a clinical strategy for a given program participant, and panel staff typically implemented the plan over the course of several months to a year, depending upon a patient's clinical needs. Care plans might describe a regimen of medications, specialty care, diet, exercise, and responses to early warning signs intended to bring a patient's chronic conditions under control. In developing care plans, LCCs reconciled medications, meaning that they (1) checked whether any medication prescriptions across providers were duplicative or conflicting and, if so, worked with the providers to simplify or harmonize the prescriptions; and (2) developed a list of medications the patient should be taking and assessed the extent of adherence. The care plans also included standard recommended clinical care guidelines for common chronic conditions that were developed by the LCC and PCP based on

the beneficiary's specific clinical needs. All care plans were documented in the iCentric online portal.

LCCs were expected to contact beneficiaries in active care plans at least once per week (almost always via telephone), and were required to make at least three attempts to contact the participant each week if the participant had not yet been reached that week. Phone calls were expected to last, on average, 5 to 30 minutes, depending on a beneficiary's needs. During these contacts, LCCs educated patients on self-management, encouraged patients to schedule appropriate visits with primary care or specialty providers, and collected information about changes in health status that they could relay to PCPs to make timely changes in clinical care as needed.

Patients participating in care coordination were also eligible for additional support services that CareFirst phased in throughout the award. Based on need (as determined by LCCs and PCPs) and interest, LCCs scheduled patients to receive home-based health assessments, remote monitoring of health conditions at home, and behavioral health services. Medicare paid for any support services already reimbursed by traditional FFS Medicare (for example, home health), but CareFirst paid these other providers to upload their data into iCentric. This provided PCPs and LCCs access to additional information about participants.

PCPs periodically reviewed each care plan and the patient's progress toward its goals; the frequency of those reviews varied depending on the beneficiaries' chronic condition or the timing of the patient's follow-up appointment with the PCP (which the LCC often attended).

For one and a half years (March 2014 through September 2015), CareFirst provided additional short-term care coordination through four case managers. These case managers contacted high-acuity patients—most of whom had recently been hospitalized—to provide post-acute care not only to participants receiving care plans, but also to other Medicare patients who were recently hospitalized, but not appropriate candidates to receive a care plan. These services were generally offered for a shorter duration than those provided through care plans (typically two to six weeks), until the acute needs could be stabilized. The case managers would sometimes refer patients to LCCs for longer-term care coordination through care plans. CareFirst began hiring the four case managers in March 2014, seven months after the intervention was implemented. As of September 2015, all four case managers had resigned their positions and were not replaced because CareFirst program administrators felt the LCCs could adequately care for patients with post-acute care needs.

Financial incentives. CareFirst offered financial incentives to facilitate program implementation and hold panels accountable for the cost and quality outcomes of their attributed patients. PCPs received $200 for developing each new care plan and $100 for updating an existing care plan as needed. Medical panels that could keep the total cost of care for their attributed Medicare patients (not only participants with a care plan) below a specified threshold qualified to earn an Outcome Incentive Award (OIA), with the size of the incentive payment proportional to a panel's performance on quality measures. The target for the cost of care was based on the number of the panel's Medicare FFS patients, the acuity of the panel's patients (as

measured by illness burden scores), and an expected growth rate in overall medical spending of 2.5 percent per year.

Technical assistance to panels. CareFirst employed five program consultants to help panel staff understand and interpret their patients' data, with the goal of influencing providers' behavior change and ultimately reducing spending and improving outcomes. CareFirst expected program consultants to focus on changing panels' referral patterns by encouraging the use of low-cost, high-value specialist providers and places of service. CareFirst developed a tiering system to rank specialists and places of service on costs, using its commercial claims data. (CareFirst did not verify that low-cost providers in the commercial setting were also low-cost providers to Medicare.) In addition, CareFirst encouraged program consultants to work with panels to (1) better identify appropriate candidates for care plans by interpreting the illness burden scores; (2) apply positive peer pressure to increase program engagement; and (3) address gaps in care, as identified by poor performance on quality measures. Program consultants had to meet with each panel at least quarterly, but they often communicated with individual PCPs more frequently.

3. Theory of action

Based on extensive review of CareFirst's program activities and goals, we developed a theory of action to depict the mechanisms through which program administrators expected the program to improve the outcomes we selected for the impact evaluation (see Table II.1 for a list of these outcomes). CareFirst expected that its HCIA-funded intervention would improve outcomes for Medicare patients through two pathways.

Primary pathway to improved outcomes. LCCs and PCPs provide care coordination services to high-risk patients, which reduces the frequency of acute exacerbations. Planned mechanisms of this pathway include the following:

1. **Improved care coordination leads to improved self-management among patients with multiple chronic conditions who are at high risk for acute care use.** These improvements, based on goals described in the care plan, could include increased patient adherence to medications, faster and more appropriate patient responses to early warning signs of acute exacerbations of their condition(s), and improved diet and exercise regimens.
2. **Improved care coordination among multiple providers—facilitated by LCCs—leads to improvements in clinical care.** This could include timely visits with PCPs and other providers and better receipt of routine recommended clinical care, such as receipt of recommended tests for diabetes, and complete lipid profile for ischemic vascular disease (IVD).
3. **Improved self-care and clinical care reduces the need for potentially avoidable and costly acute care services.** Specifically, these improvements reduce the frequency of acute exacerbations, reducing the need for outpatient ED visits and inpatient admissions. The benefits from improved clinical care and self-care should be most apparent among admissions for ambulatory care-sensitive conditions, which are often considered potentially preventable. Reducing the frequency of acute events for the highest-risk patients should also

reduce the frequency of events among the *full* Medicare population, given that high-risk patients account for a large proportion of all acute events. Further, because hospitalizations and post-acute care drive overall Medicare spending, reductions in acute events should reduce total Medicare Part A and B spending.

4. **When acute care is needed, LCCs and PCPs help patients manage post-acute needs to reduce the need for additional acute care later.** Care coordination following an acute exacerbation increases the proportion of people who receive ambulatory follow-up within 14 days of discharge. This further reduces outpatient ED visits (because patients have prompt follow-up care, so they do not have to seek immediate care from another setting, such as the ED). Appropriate ambulatory follow-up and medication reconciliation also reduces the number of 30-day unplanned hospital readmissions.

Secondary pathway to improved outcomes. Program consultants provide technical assistance to panels, with a focus on more efficient referral patterns. Referrals to more cost-effective providers and settings leads to lower total Medicare Part A and B spending. Planned mechanisms of this pathway include the following:

1. **Program consultants share data with panels to encourage PCPs to send their patients to lower-cost specialists and lower-cost settings of care.** PCPs should also be motivated to change their referral patterns based on the desire to earn an OIA, as well as by positive peer pressure from colleagues and program consultants.
2. **The changes in referral patterns lead to more Medicare patients receiving care from cost-effective providers or in cost-effective settings,** which lowers total Medicare Part A and B spending.

Text box III.1. Example from CareFirst illustrating the program's theory of action

"A Medicare beneficiary with diagnoses of diabetes, hypertension, coronary artery disease, hyperlipidemia, and sleep apnea and a history of prior strokes and seizures was selected for care coordination. The patient was on multiple medications and was experiencing bouts of dizziness, trouble with balance and had fallen last year. The LCC [local care coordination nurse] began care coordination of the member and found that the CPAP machine[a] ordered by the pulmonologist did not fit and was not being utilized. The beneficiary also had an overlap in medications, due to visits to multiple specialists who were unaware of what the others were prescribing.

"The LCC worked with the beneficiary's providers to ensure that he [the patient] had a CPAP machine that fit properly and could be utilized, had accurate prescriptions that are not duplicative, and that the member is compliant with specialist visits and physician orders. Today, the patient no longer has symptoms of dizziness or daytime fatigue, and his fall risk has been significantly reduced. The beneficiary is more stable in a home setting and continues to engage with his PCP [primary care provider] to manage his chronic conditions."

[[a] A continuous positive airway pressure (CPAP) machine supplies a constant and steady air pressure through a mask or nose piece. It is a common component of treatment for sleep apnea.]

Source: CareFirst's Eigth quarterly report to the Centers for Medicare & Medicaid Services.

4. Intervention staff and workforce development

Table III.1 provides key details about staff hired for the HCIA-funded intervention. Through a vendor, Healthways, CareFirst hired registered nurse LCCs to provide care coordination services by facilitating the development and implementation of care plans for high-risk participants. CareFirst added case managers in March 2014 to support the program by focusing on care transitions for a small subset of patients who were recently discharged from the hospital. CareFirst also hired program consultants at the start of the program, who analyzed data on each panel's attributed population to provide technical assistance to them.

Table III.1. Key details about intervention staff

Program component	Staff members	Staff/team responsibilities	Adaptations?
Care coordination	Local care coordinator	Through its vendor Healthways, CareFirst hired registered nurse LCCs to help develop and implement care plans for high-risk participants. LCCs were supposed to contact participants with active care plans at least once per week by telephone, making at least three attempts to contact the participant each week. The LCCs generally did not physically work in the primary care practices (that is, they were not embedded) but they did occasionally visit patients and PCPs at the practices. The extent to which LCCs were functionally integrated into panels varied. Some LCCs interacted with their PCPs regularly, whereas others did so far less frequently; this was based to some extent on the preferences of the PCP. A full caseload for an LCC was considered to be 45 active care plans.	No
Care coordination	Case manager	Case managers, who were registered nurses, provided care coordination services to participants experiencing a care transition after an acute care episode. They helped to ensure post-acute care needs were addressed. For example, case managers assisted participants in obtaining resources available in the community. They also transitioned eligible participants to an LCC for a longer-duration care plan.	Yes. Although not a part of initial program implementation, CareFirst began recruiting case managers in March 2014. The position was approved as part of the Year 1 carry-over funding request.
Technical assistance	Program consultant	CareFirst hired program consultants, who informed PCPs' behavior by providing them with provider- and panel-level data reports to identify key cost drivers, quality metrics, and potential gaps in care. Program consultants tended to focus on a panel's entire attributed population.	No

Sources: Interviews and document review.

LCC = local care coordination nurse; PCP = primary care provider.

B. Implementation effectiveness

In this section, we examine the evidence on implementation effectiveness—that is, we analyze measures of the intervention delivered and, when possible, compare those measures with the services the awardee intended to deliver. We assess the evidence on implementation effectiveness in five areas: (1) program enrollment, (2) service delivery, (3) staffing, (4) training, and (5) implementation timeliness. To conduct this analysis, we used data from interviews with program administrators and frontline staff, self-reported metrics included in CareFirst's self-monitoring and measurement reports to CMMI, and data from CareFirst on patients it enrolled in care coordination. We often report metrics through July 2015—one month after the end of the original award period—because our impact evaluation runs through July 2015. (Including July 2015 in the impact evaluation permits impact analyses through eight intervention quarters, with the final quarter—May through July 2015—largely occurring before the original award period ended.)

1. Program enrollment

CareFirst attributed about 35,000 Medicare FFS beneficiaries to the 14 panels in the first intervention quarter (starting August 2013), more than CareFirst's initial target of 25,000 beneficiaries. All 14 panels participated throughout the entire intervention. The number of Medicare FFS beneficiaries cumulatively attributed to the 14 panels increased to more than 40,000 from January to October 2015.

2. Service-related measures

Care coordination for high-risk beneficiaries. Although all attributed Medicare FFS beneficiaries (except those dually eligible for Medicaid) were eligible to receive care coordination services, the panels focused these services on high-risk patients. By July 2015, the program had developed 3,276 care plans for 3,152 unique Mediare beneficiaries. A few beneficiaries received more than one care plan, as LCCs reopened care plans at the request of either the beneficiary or the PCP, often coinciding with a change in the beneficiary's condition. The number of participants receiving care coordination services, as indicated by active care plans, increased throughout the award (Figure III.1). CareFirst exceeded its target number of care plans for high-risk beneficiaries, despite a delay of almost a year in the start of the program (Section III.B.5 discusses program timeliness). For example, the program had 1,895 active care plans in December 2014, more than its initial target of 1,350 active care plans (Figure III.1).

As shown in Figure III.2, most Medicare beneficiaries who received a care plan were considered to be at high risk based on CareFirst's illness burden scores. About 67 percent of all care plans went to beneficiaries in CareFirst's top illness burden score band (among five bands), although beneficiaries in this band accounted for only 25 percent of all attributed Medicare beneficiaries. The vast majority (91 percent) of care plans were provided to beneficiaries in the top two bands, whereas beneficiaries in those two bands accounted for 62 percent of all attributed members.

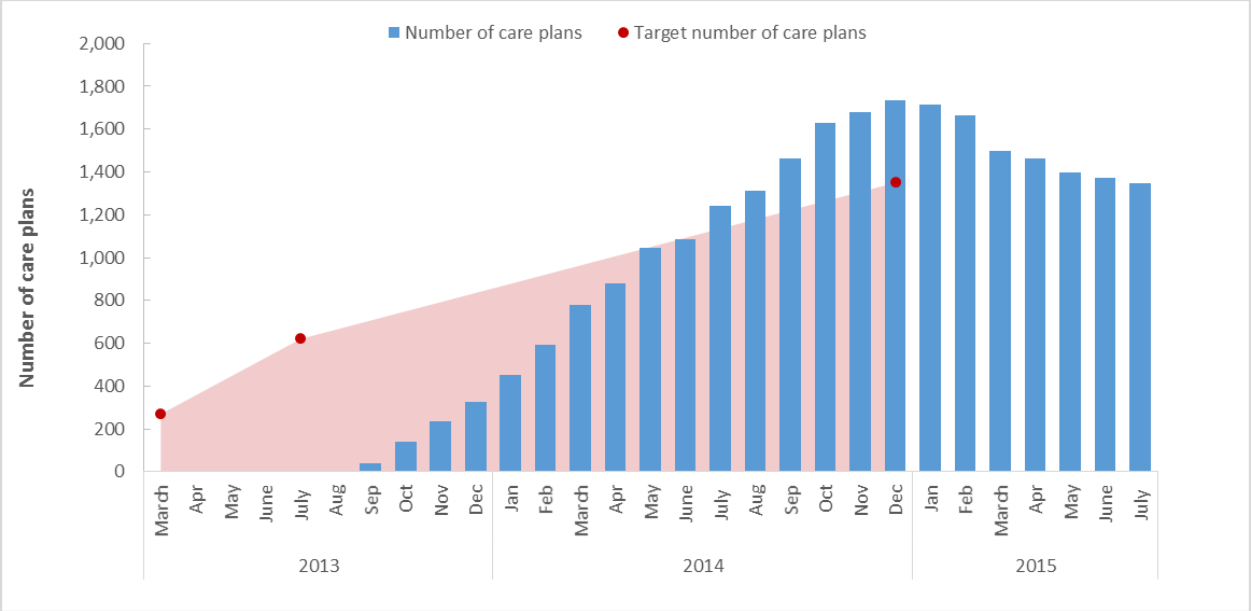
Although care plans went preferentially to patients in the top two bands, only a modest fraction of Medicare beneficiaries in these two bands received care plans, in part due to the large

number of people in these bands. We estimate that, at most, 22 percent of all beneficiaries in the top band of illness burden scores received a care plan and 6 percent of those in second band received a care plan. Less than 4 percent of beneficiaries in each of the three healthiest tiers received a care plan.

On average, LCCs successfully contacted participants in active care plans 0.75 times per week, nearly achieving CareFirst's goal of at least once per week (Table III.2). Most of these contacts (89 percent) were by telephone (data not shown), with the rest in person at the PCP's offices. LCCs routinely conducted medication reconciliations (1.6 times, on average, per active care plan) and referred patients to ancillary services (home health care, telemonitoring, and behavioral health) as appropriate.

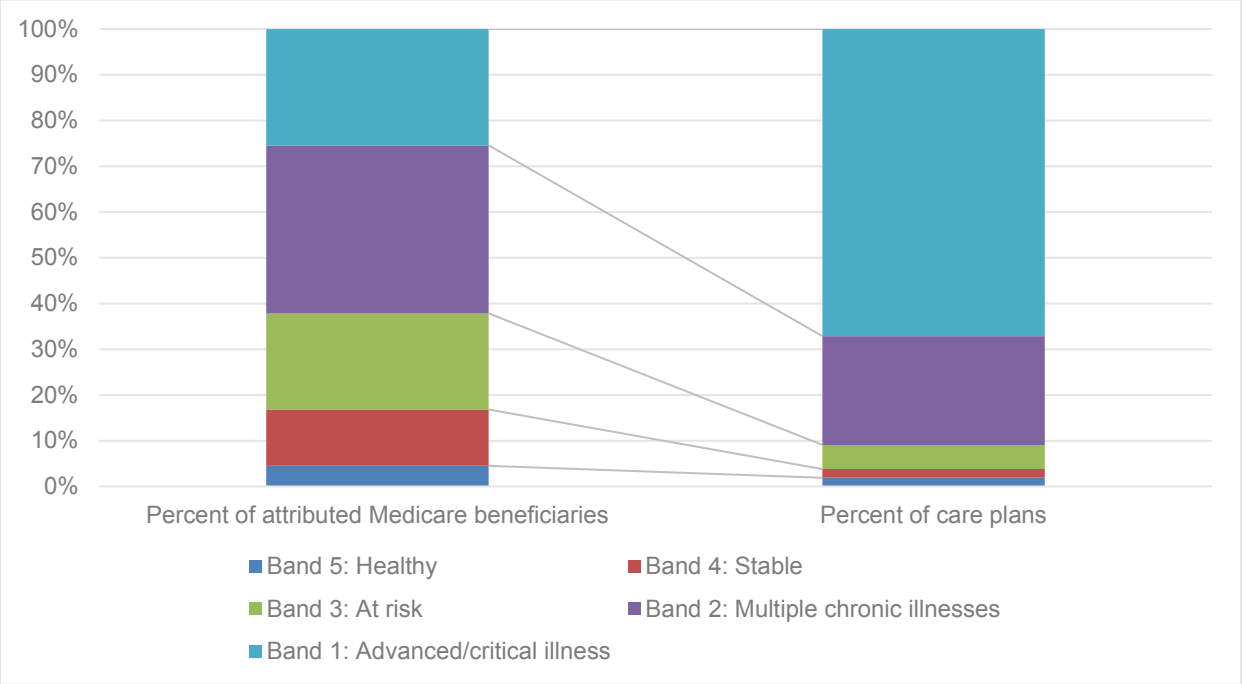
The awardee intended for care plans to remain active for varying lengths of time, depending on whether the participant achieved his or her care plans goals and continued to engage in the care planning process. As of July 2015, 91 percent all of initiated care plans had been closed. The closed plans were, on average, active for 260 days (Table III.2).

Figure III.1. Number of active care plans and targets, by month



Sources: Analysis of CareFirst's HCIA quarterly reports, December 2012 through July 2015; CareFirst's operational plan; and personal communication with CareFirst, May 2016.

Figure III.2. Percentage of attributed beneficiaries and care plan participants, by band of illness burden score



Source: Data provided through personal communication with CareFirst, May 2016. Data are through October 2015.

Table III.2. Service metrics (and targets, if applicable), by program component

Service metrics	Awardee target	Actual	Target met?
Program component: Care coordination for high-risk beneficiaries			
Number of care plans for Medicare FFS beneficiaries	1,350 active care plans in December 2014[a]	1,895 active care plans in December 2014 [3,656 cumulative care plans as of December 2015]	Yes
Average care plan duration	Care plans could be active for varying lengths of time, depending on whether the participant achieved his or her care plans goals and continued to engage in the care planning process	By July 2015, 91 percent all of care plans initiated had closed. Closed plans had been active, on average, 260 days. • Minimum: 4 days • 25th percentile: 130 days • 50th percentile: 224 days • 75th percentile: 359 days • Maximum: 898 days	NA
LCC successful contact with patients in active care plans	1 successful contact per week per beneficiary in an active care plan	0.75 successful contacts per week per beneficiary in an active care plan (this does not include unsuccessful attempts to contact a beneficiary)	Nearly
Medication reconciliation for those with care plans	CareFirst intended for LCCs to conduct a medication reconciliation at the start of each care plan and then again during each maintenance visit with the PCP	LCCs completed 1.6 medication reconciliations per care plan, on average	Yes
Referrals to three support services for those with care plans: • Electronic symptom monitoring • Home health care • Behavioral health care	Used as needed	Percentage of beneficiaries who were in active care plans who received: • Electronic symptom monitoring: 12% • Home health care: 16% • Magellan behavioral health services: 7%	NA
Program component: Financial incentives			
Calculation and, if applicable, payment of OIAs to panels	Calculate and pay OIAs for each of 3 performance years (2013-2015)	Calculated OIA for 2013 and 2014 • 2013: 5 panels earned OIAs ranging from $8,000-$116,000. • 2014: 12 panels earned OIAs ranging from $3,000-$494,000[b] CareFirst expects to complete the OIA for 2015 when claims data are ready	Yes

Table III.2 *(continued)*

Service metrics	Awardee target	Actual	Target met?
Program component: Technical assistance to panels			
Program consultant meetings with panels to discuss trends and opportunities to reduce spending and improve quality	At least quarterly	Program consultants held 130 to 170 meetings with panels each quarter in 2015 (or, on average, 9 to 12 times per panel per quarter)	Yes

Sources: Analysis of CareFirst's HCIA quarterly reports, December 2012 through July 2015, and personal communication with CareFirst, May 2015.

[a] We used December 2014 as the target month because CareFirst's operational plan, which first stated the program's goals, established goals through that date.

[b] CareFirst calculated that the panels reduced Medicare spending by $26 million in 2014 and, as a result, the panels earned $4 million in OIAs. However, CareFirst paid out only $1.5 million (with the same proportional reduction across panels) so that total CareFirst spending could stay within the amount approved in the HCIA budget.

FFS = fee-for-service; HCIA = Health Care Innovation Award; LCC = local care coordination nurse; OIA = Outcome Incentive Award; PCP = primary care provider.

Financial incentives. CareFirst paid PCPs $200 for each care plan they initiated and $100 for each existing care plan that PCPs reviewed and updated. In addition, CareFirst paid two rounds of OIAs to panels that kept the total cost of care for their attributed Medicare beneficiaries below a specified target (specific to each panel, based on patients' risk) and that met certain quality standards. In July 2014, CareFirst paid the first round of OIAs for services provided from the program start date on August 1, 2013, to the end of 2013; only five panels received OIAs, ranging from $7,843 to $116,045. In July 2015, CareFirst paid the second round of OIAs for services provided in calendar year 2014, the first full year that the program was implemented. Of 14 panels, 12 received OIAs in 2015, ranging from $2,868 to $494,132. CareFirst also planned to pay OIAs in July 2016 for 2015 performance (data on these payments were not available at the time of writing).

Technical assistance to panels. Program consultants met with panels, on average, 9 to 12 times per quarter in 2015, well above the initial target of once per quarter. During these meetings, consultants reviewed and discussed the cost and quality data for the panels' patients. Program consultants focused increasingly over time on developing strategies to improve panel referral patterns to more cost-effective specialists and settings of care.

3. Staffing measures

By July 2015, CareFirst engaged 149 PCPs across the 14 panels to participate in the HCIA-funded intervention. CareFirst directly funded 44 registered nurse LCCs, five program consultants, and four case managers. The 44 LCCs exceeded CareFirst's initial target of 27 LCCs. CareFirst hired these 17 additional LCCs because the panels served more Medicare FFS beneficiaries than initially projected—increasing the need for care coordination services—and because they could fund the positions with carry-over funds from Year 1, when the intervention was delayed (Section III.B.5). CareFirst also hired five program consultants, instead of its initial target of one. Lastly, though not originally part of CareFirst's core program design, CareFirst

hired four case managers to provide care transitions support to recently hospitalized patients. Case managers provided services only from March 2014 through September 2015.

4. HCIA-funded training

CareFirst implemented training to help LCCs provide care coordination services for high-risk patients. All LCCs completed an initial four-week training class (160 hours), complemented by hands-on experience in the field before beginning their work with panels.

To assess perspectives of HCIA-funded staff who received this training, we administered the HCIA Primary Care Redesign Trainee Survey between January and March 2015 (17-19 months after the start of implementation). Of the 43 LCCs who participated in CareFirst's HCIA-funded program at the time of the survey, 26 responded to the trainee survey (a response rate of 60 percent).

Almost all of the 26 LCC respondents (92 percent) reported receiving formal training (data not shown). Of the 24 who reported receiving training (formal or informal), all received new hire training. Most also reported receiving training for developing and implementing care plans, including training in writing care plans (96 percent) and motivational interviewing (71 percent). Half reported that they received training in cost-effective options for providing treatment (for example, emergency room use versus urgent care for the elderly) (data not shown).

Of the 24 LCCs who reported receiving training, three-quarters (18 LCCs) thought their training had a positive effect on the quality of care they provided and nearly as many (17) thought their training had a positive effect on the patient-centeredness of care they provided. Roughly two-thirds of the 24 LCC respondents reported that their training had a positive effect on their ability to explain information to patients (71 percent) and relay relevant information to care teams (75 percent), as well as to help patients control their own care (62 percent)—all key elements of implementing care coordination services (Table III.3).

The survey data also confirmed that LCCs routinely managed patients' care through activities related to developing and implementing care plans for high-risk patients, which is consistent with CareFirst's intervention design (Table III.4). For example, all 26 LCC respondents reported routinely helping to manage patients' care in the following ways: calling patients to check on medications and symptoms; coordinating care between visits; educating patients about managing their own care; counseling patients on how to exercise, receive good nutrition, and stay healthy; and engaging in patient coaching. Most LCC respondents (88 percent) also reported that they had attended medical appointments with patients. In addition, more than half of LCC respondents reported routinely assisting patients with accessing nonmedical services such as housing, job training, supplemental nutrition services (58 percent) and providing follow-up services for recently discharged beneficiaries (69 percent). Because both of these services are provided on an as-needed basis (depending on whether someone had an inpatient stay or a need for nonmedical services), these activities were a less common part of the program.

Table III.3. LCCs' perceptions of the effects of training on their care, from the trainee survey

Survey question		Percentage of respondents (and number out of 24[a]) who reported the training had a positive effect on this dimension of their care
Please indicate the impact you believe the training you received for the expansion of CareFirst's PCMH program to the Medicare population has had on the following aspects of care you provide to patients enrolled in CareFirst or Healthways	1. Quality of care	75% (18)
	2. Ability to respond in a timely way to patients' needs	NA[b]
	3. Efficiency/cost-effectiveness of care	63% (15)
	4. Patient-centeredness	71% (17)
	5. Equity	58% (14)
Please indicate whether the training you received has had a positive or negative effect on your ability to …	1. Explain information about patient care to patients and their families in lay terms	71% (17)
	2. Relay relevant information to the care team	75% (18)
	3. Work with diverse set of patients	63% (15)
	4. Access the care they need	75% (18)
	5. Help patients access nonmedical services	50% (12)
	6. Help patients take control of their own care	62% (16)
	7. Use data to evaluate my performance to improve the services I provide to patients	73% (19)

Source: Mathematica's analysis of trainee survey.

[a] The denominator includes all trainees who reported they received some training (formal or informal) for the expansion of CareFirst's PCMH program to Medicare beneficiaries.

[b] Not reported because fewer than 11 respondents reported yes.

LCC = local care coordination nurse; NA = not available; PCMH = patient-centered medical home.

Table III.4. LCCs' care management activities, as reported in the trainee survey

Activity	Percentage (and number) of 26 LCCs who reported that they	
	Personally help to manage patients' care through this activity *routinely*	Spend more than 2 hours on this activity on a typical work day
Call patients to check on medications, symptoms, or help coordinate care between visits	100% (26)	96% (25)
Execute standing orders for medication refills, ordering tests, or delivering routine preventive care	--[a]	--[a]
Educate patients about managing their own care	100% (26)	81% (21)
Counsel patients on exercise, nutrition, and how to stay healthy	100% (26)	73% (19)
Assist patients with accessing nonmedical services such as housing, job training, supplemental nutrition services (for example, SNAP benefits)	58% (15)	--[a]
Attend medical appointment with patients	88% (23)	--[a]
Follow up on care transitions	69% (18)	--[a]
Coaching patients	100% (26)	77% (20)

Source: Mathematica's analysis of trainee survey.

[a] Not reported because fewer than 11 respondents reported yes.

LCC = local care coordination nurse; SNAP = Supplemental Nutrition Assistance Program.

5. Program timeline

CareFirst experienced initial implementation delays due to problems obtaining complete Medicare claims data for patient attribution. CareFirst acquired the necessary data in June 2013 and officially launched the HCIA-funded initiative in August 2013, 13 months later than planned in its initial program application to CMMI. The Centers for Medicare & Medicaid Services (CMS) granted CareFirst a no-cost extension that enabled CareFirst to use its remaining HCIA funds to pay for program services through December 2015 and to use an additional six months (through June 2016) to calculate and pay OIAs for the final program year. CareFirst continued to pay for program services after December 2015 using its own funds while program administrators discussed with CMS options for sustaining the program.

C. Summary of facilitators of and barriers to implementation

Several factors facilitated implementation of CareFirst's HCIA-funded intervention, but others hindered implementation. We described those factors in detail in the second annual report (Geonnotti et al. 2015). Here we summarize key facilitators and barriers, along with any new information since the second annual report that supports those facilitators or barriers (Table III.5).

Table III.5. Summary of key facilitators and barriers to the implementation of CareFirst's HCIA-funded initiative

Item	Description based on findings in the second annual report	Additional supporting data not available in the second annual report, if applicable
	Facilitators (domain)	
Prior experience with a similar commercial PCMH program (internal factor)	PCPs reported that it would have been more difficult to implement the HCIA-funded intervention if they had not previously been involved with CareFirst's commercial PCMH program. All panels that participated in the HCIA-funded initiative were already established and functioning, with three years of operating experience in CareFirst's commercial PCMH program. CareFirst purposefully minimized the differences between its commercial and Medicare PCMH programs; the commercial program features remained largely intact with minimal modifications to the Medicare program. PCPs reported that it was helpful to build on their knowledge of the commercial program, making the transition rather seamless to extend services to Medicare beneficiaries.	
LCCs as a new resource for panels (implementation process)	PCPs reported that the addition of LCCs was a welcomed resource, as the PCPs would not have had time to focus as intensely on high-risk beneficiaries without HCIA funding to integrate LCCs into the care-planning process. Factors that appeared to help build a functional relationship between PCPs and LCCs included LCCs having a presence in the practice(s) as much as possible, their own space to work in the practice, access to the EHR, and a mechanism to educate practice staff about the role of the LCC.	PCPs continued to show support for staffing resources. PCPs who responded to the clinician survey reported that the availability of personnel had a positive impact on the implementation of the HCIA initiative at their practice location (74.0 percent of PCPs said it had a positive impact).
PCP engagement (implementation process)	PCP engagement was key to successfully integrating LCCs into their primary care practices and delivering care coordination services to participants. CareFirst believes PCPs must be willing to have an LCC based in their practices and engage in the care-planning process.	CareFirst's internal engagement score, based on the degree to which PCPs have engaged with the HCIA-funded initiative, increased from 53 percent at the start of the program to 86 percent by the end June 2015. LCCs submit engagement scores for participating PCPs, which regional care coordinator supervisors review and verify. The engagement score is a composite of the following questions that the LCC answers based on her or his experience with the PCPs in a panel: (1) the PCP helps create an environment in his or her practice that is conducive to conducting the program; (2) the PCP actively seeks to work with the LCC to identify and schedule members appropriate for care plans; (3) the PCP clearly and effectively explains the program to care plan-eligible members; (4) the PCP facilitates and guides other PCPs in the practice toward program goals; and (5) overall, PCPs are seen as active, willing partners in achieving program goals and facilitating cohesive panel performance.

Table III.5 *(continued)*

Item	Description based on findings in the second annual report	Additional supporting data not available in the second annual report, if applicable
CRISP data system to facilitate identifying eligible participants (external factor)	CRISP is Maryland's statewide health information exchange that provides real-time notifications to PCPs when their participants are in the hospital or ED. Several panels elected to participate in this initiative, which enabled PCPs to improve transitions of care and identify unstable participants who might benefit from a care plan. Given data lags in CareFirst's methods for identifying potentially eligible participants for care plans, CRISP has become an increasingly important tool for LCCs and PCPs to identify—in real time—those who could benefit from a care plan.	
	Barriers (domain)	
Challenges identifying who would benefit the most from a care plan (external factors)	Although PCPs and LCCs have freedom to select the participants who are most appropriate for care plans, CareFirst has refined the process to better target clinically unstable participants, who it considers the major drivers of health care costs. Throughout implementation, CareFirst learned that some participants can have high illness burden scores, but are not actually clinically unstable. Rather, their high illness burden scores might reflect a recent hospitalization for an acute, nonchronic event.	
Medical complexity of Medicare patients compared with commercial patients (external factor)	Staff reported that it is more difficult and time-consuming to develop care plans for Medicare beneficiaries because they generally have higher rates of chronic disease, are on more medications, and are treated by more specialists.	

Note: We reviewed four domains associated with implementation experience: (1) program characteristics, (2) implementation process, (3) internal factors, and (4) external environment. Implementation research suggests that barriers and facilitators within these domains are important determinants of implementation effectiveness.

CRISP = Chesapeake Regional Information System for our Patients; ED = emergency department; EHR = electronic health record; HCIA = Health Care Innovation Award; LCC = local care coordination nurse; PCMH = patient-centered medical home; PCP = primary care provider.

Three factors were particularly important in facilitating program implementation, and two factors were barriers. First, PCPs' experience in the commercial PCMH program facilitated implementation because program staff could build on their existing knowledge and tools to extend the program to Medicare beneficiaries. All 14 panels had been part of the PCMH program since its start in 2011, and CareFirst intentionally minimized the number of differences between the commercial program and the HCIA intervention for Medicare beneficiaries. Second, adding LCCs as a new resource for panels facilitated implementation because PCPs would not have had time to focus as intensely on high-risk patients without the support of LCCs. Third, PCPs were highly engaged in the program, which facilitated successful integration of LCCs into primary care practices and delivery of care coordination services to participants. Two important barriers to implementation included (1) challenges identifying who would benefit the most from a care plan; and (2) the medical complexity of Medicare patients compared with commercial patients, which sometimes made it difficult to adapt care coordination strategies developed for commercial patients to best meet the needs of Medicare patients.

D. Conclusions about the extent to which the program, as implemented, reflects core design

Despite a 13-month delay, CareFirst implemented its HCIA-funded intervention largely as planned. As noted previously, the 14 participating panels exceeded CareFirst's targets for the number of beneficiaries attributed to the intervention and for the number of care plans implemented. CareFirst delivered services for each of its three planned intervention components. The organization hired more LCCs than initially envisioned, and, in surveys, LCCs reported both receiving planned trainings to facilitate care coordination and routinely engaging in activities consistent with the planned design of the care coordination component, including contacting patients in active care plans roughly as intended, reconciling medications, working with engaged PCPs to improve clinical care, and referring appropriate patients to support services as needed. CareFirst also reimbursed PCPs for producing and updating care plans and paid out financial incentives in the form of OIAs for 2013 and 2014, as planned. Finally, throughout the intervention, program consultants provided technical assistance to panels to try to alter PCP referral patterns, encouraging use of more cost-effective specialists and sites of care. Our estimates for program impacts (Section V) account for the 13-month delay in program implementation by setting the start of the intervention period to when the intervention actually began in August 2013, not when it was originally planned to begin in July 2012.

Although the intervention was implemented largely as planned, two key implementation barriers might have limited success of the care coordination component. First, CareFirst learned that the process for identifying the highest-risk patients for care coordination services could have been limited by the extent to which LCCs and PCPs relied on illness burden scores. Some patients could have had have high illness burden scores but they were not actually clinically unstable; rather, their high illness burden scores reflected a recent hospitalization for an acute, nonchronic event, making them less likely to benefit from a care plan targeting complex, chronic conditions. Some PCPs and LCCs have refined their process for identifying candidates for care plans, relying on clinical judgement to supplement illness burden scores. Second, although the program was modeled on an existing program targeting commercial members, the complexity of

Medicare patients relative to the commercial patients made implementing the care coordination component challenging.

CareFirst also provided one program element beyond the intervention's core design. CareFirst added case managers to the program in March 2014 for about 18 months to provide additional support to LCCs for patients (not just those in care plans) with a higher level of acuity who often had also been recently hospitalized. The impact evaluation captures any effect of the services provided by case managers, together with the effects of the other three intervention components that constituted the intervention's core design.

IV. CLINICIANS' PERCEPTIONS OF PROGRAM EFFECTS ON THE CARE THEY PROVIDE TO PATIENTS

This section describes the available evidence on the extent to which CareFirst's intervention had its intended effects on changing PCPs' behavior as a way to achieve desired impacts on patients' outcomes. As described in Section III.A.3, the program's theory of action required that PCPs (1) engage in care coordination for high-risk Medicare beneficiaries and (2) change referral patterns. We use data from two rounds of the HCIA Primary Care Redesign Clinician Survey and from CareFirst's self-monitoring metrics on PCP engagement in care coordination to assess changes in providers' behavior and conclude whether the anticipated changes occurred. Both surveys rely on self-reported responses and reflect clinicians' perceptions of the program, rather than measuring quantitatively direct program effects on the care they provide.

A. Clinician survey

Survey methods. We administered the clinician survey in two rounds (fall 2014 and summer 2015). We sent the survey to PCPs working in the 14 panels at the time of each survey. A total of 86 and 80 clinicians participating in CareFirst's HCIA program responded to the survey during the first and second rounds, respectively (a response rate of 68 percent in round 1 and 63 percent in round 2).

Survey results. Almost all respondents to the clinician survey reported being somewhat or very familiar with the HCIA program (88 percent in round 1 and 91 percent in round 2). As shown in Table IV.1, the program appears to have had its intended effects for most providers familiar with the program on dimensions related to care coordination. Specifically, 68 to 75 percent of respondent said they thought the HCIA program improved the quality, timeliness, and patient-centeredness of care they provided to patients in their practices in the past year. The remaining respondents thought the program had no effect on those dimensions of their care or that it was too soon to tell (we did not separate the respondents into these two categories because the cell sizes were often smaller than the required 11 for reporting). In contrast to the generally positive perceived effects on quality, timeliness, and safety, only 29 to 49 percent of respondents said the program improved the efficiency or equity of care, or information available for clinical decision making, with the remaining respondents reporting the program had no effect on these dimensions of care or it was too early to tell. Clinician's perceptions of program effects were similar across the two survey rounds, although a modestly higher percentage in round 2 reported that the program improved the quality of their care (68 versus 78 percent) or improved the safety

of their care (47 versus 68 percent) (Table IV.1). Given that PCPs and LCCs were expected to work closely to establish and maintain care plans, it is not surprising that a large majority of clinicians in the sample (about 83 percent in round 2) reported working as part of a care team (data not shown). Most clinicians working in a care team agreed that members of the team relayed information in a timely manner (87 percent in round 2) and had sufficient time for participants to ask questions (89 percent in round 2) (data not shown; round 2 results closely matched those of round 1).

Table IV.1. PCPs' perceptions of the effects of the program on their care, from the clinician surveys (both rounds)

	Percentage (and number) of PCPs reporting that the HCIA had the following effect on the care they provided to patients enrolled in their practices in the past year			
	First round of survey (13 to 15 months after program implementation) N = 76		Second round of survey (21 to 23 months after program implementation) N = 73	
Dimension of care	Positive impact	No impact or too soon to tell	Positive impact	No impact or too soon to tell
Quality	68% (52)	28% (22)	78% (57)	21% (15)
Ability to respond in a timely way to patients' needs	71% (54)	28% (21)	67% (49)	32% (23)
Efficiency	37% (28)	54% (41)	49% (36)	41% (30)
Safety	47% (36)	47% (36)	68% (47)	34% (25)
Patient-centeredness	75% (57)	22% (17)	74% (54)	25% (18)
Equity	38% (29)	57% (43)	38% (28)	55% (40)
Information available for clinical decision making	NA	NA	49% (36)	49% (36)

Source: HCIA Primary Care Redesign Clinician Survey: Round 1 (field period 9/2014 – 11/2014), Round 2 (field period 5/2015 – 7/2015).

Note: The number (and percentages) are limited to PCPs who reported that they were at least somewhat familiar with the HCIA program.

HCIA = Health Care Innovation Award; PCP = primary care provider.

NA = not available.

B. CareFirst data on clinician behavior

According to data obtained from CareFirst, 90 percent of PCPs in participating panels opened at least one care plan. Of PCPs who opened a care plan, the average number of care plans was 24 (results not shown). Although some providers opened many care plans (the maximum was 107), most providers had at least 8 patients in a care plan (that is, 75 percent of PCPs with at least one care plan had 8 or more care plans). In addition, CareFirst's internal engagement score, based on LCCs' assessment on the degree to which PCPs have engaged with the HCIA-funded initiative, increased from 53 percent at the start of the intervention to 86 percent by the end of

June 2015. The engagement score for each PCP was based on the LCC's responses to the following statements: (1) the PCP helps create an environment in his or her practice that is conducive to conducting the program; (2) the PCP actively seeks to work with the LCC to identify and schedule members appropriate for care plans; (3) the PCP clearly and effectively explains the program to care plan-eligible members; (4) the PCP facilitates and guides other PCPs in the practice toward program goals; and (5) overall, PCPs are seen as active, willing partners in achieving program goals and facilitating cohesive panel performance. CareFirst assigned a number of points to each statement and then calculated a score (as a percentage) for each PCP by summing the number of points earned and dividing it by the total number of possible points.

C. Conclusions about intermediate program effects on clinicians' behavior

Based on available information, the HCIA-funded initiative appears generally to have had its intended effects on how PCPs provide care. Virtually all PCPs surveyed were aware of the program, and most believed the HCIA-funded initiative improved the quality, patient-centeredness, and timeliness of care. CareFirst's self-monitoring data also support these self-reported responses. However, about a quarter of PCPs thought the HCIA-funded initiative had no effect on these key dimensions of care or that it was too early to tell, suggesting that some PCPs were not as fully engaged as CareFirst might have hoped. Further, we do not have any direct evidence to assess whether the intervention changed PCPs' referral patterns, an important activity to achieve intended reductions in Medicare spending.

V. PROGRAM IMPACTS ON PATIENTS' OUTCOMES

This section of the report draws conclusions, based on available evidence, about the impacts of CareFirst's HCIA program on patients' outcomes in four domains: quality-of-care processes, quality-of-care outcomes, service use, and spending. We first describe the methods for estimating impacts (Section V.A) and then the characteristics of the 14 HCIA treatment panels at the start of the intervention (Section V.B). We next demonstrate that the treatment panels were similar at the start of the intervention to the panels we selected as a comparison group, which is important for limiting potential bias in impact estimates (Section V.C). Finally, in Section V.D, we describe the quantitative impact estimates, their plausibility given implementation findings, and our conclusions about program impacts in each domain. Our conclusions in this report are preliminary because the analyses do not yet include the six months that CareFirst's intervention was extended beyond the original award period. The findings in this report update the impact results from the Second Annual Report for CareFirst (Geonnotti et al. 2015), extending the outcome period by 6 months and adding new outcomes.

A. Methods

1. Overview

We estimated program impacts on patients' outcomes as the difference in outcomes for Medicare FFS patients served by the 14 treatment panels and those served by 42 matched comparison panels, adjusting for any differences in outcomes between these groups during the year before the intervention began. We prespecified primary tests, describing the evidence we

would need to conclude that the program was effective, and the awardee and CMMI reviewed these. Each test specified a population, outcome, time period, expected direction of effect, and threshold that we count as substantively important. The purpose of these primary tests was to focus the impact evaluation on hypotheses that would provide the most robust evidence about program effectiveness. We used the results from the primary and secondary tests (robustness checks) to draw conclusions about program impacts in each of the four evaluation domains. The remaining subsections describe each component of the impact evaluation in more detail.

2. Treatment group definition

The treatment group consists of Medicare FFS patients served by the 14 treatment panels in four baseline quarters before the intervention began (August 1, 2012, to July 31, 2013) and eight intervention quarters (August 1, 2013, to July 31, 2015).

We constructed the treatment group in three steps.

1. First, we used CareFirst's own decision rules to attribute Medicare FFS patients in each baseline and intervention month to the 14 treatment panels. Specifically, we attributed a patient each month to the PCP who, based on Medicare FFS claims, provided the plurality of primary care services in the past 12 months. If the beneficiary did not have any primary care services in the past 12 months, we attributed him or her to the PCP who provided the plurality of care in the past 24 months. If there was a tie, we attributed to the PCP who provided the most recent service. Then, in each month, we attributed the beneficiary to the treatment panel for which the PCP worked that month. CareFirst provided data on providers who worked in the 14 treatment panels, and when.
2. Second, in each baseline and intervention period, we assigned each patient to the first treatment panel he or she was attributed to in that period, and continued to assign him or her to that panel for all quarters in the period. This assignment rule—which is distinct from the attribution method—ensures that, during the intervention period, patients did not exit the treatment group solely because the intervention succeeded in reducing their service use (including visits at treatment panels). The definition for the baseline period corresponds to that of the intervention period so that, across the two periods, interpretation of the population changes over time should be comparable.
3. Third, we applied additional restrictions to refine the analysis sample in each quarter. A patient assigned to a treatment panel in a quarter was included in the analysis sample for that quarter if he or she (1) had observable outcomes for at least one day in the quarter; (2) lived in Maryland or surrounding states (Delaware, Pennsylvania, or Virginia) or Washington, D.C., for at least one day of the quarter; and (3) was not enrolled in Medicaid at any time during the quarter (because CareFirst excludes Medicare–Medicaid dual enrollees from its intervention). For this sample, outcomes were observable for beneficiaries who were enrolled in Medicare FFS (Part A and B), were alive, and had Medicare as their primary payer.

In addition to this full treatment sample, we defined a subset of patients who were at high risk of hospitalizations and other expensive medical care. This high-risk subgroup enabled us to

conduct secondary tests or robustness checks (Section V.A.7), examining whether any observed effects were concentrated among high-risk members. This would be expected from the program theory of action, given that CareFirst targets its care coordination services to high-risk beneficiaries. In each baseline quarter, we defined the evaluation's high-risk subgroup to consist of beneficiaries with a Hierarchical Condition Category (HCC) score in the top third among all treatment group members with observable outcomes at the start of the baseline period. The HCC score, developed by CMS, is a continuous variable that predicts a beneficiary's Medicare spending in the following year relative to the national average, with 1.0 indicating that the predicted spending is at the national average and 2.0 indicating that it is twice that average. The HCC score is similar to, but not exactly the same as, the illness burden scores that CareFirst calculated and used to help identify beneficiaries who would benefit from intensive care coordination services. In each intervention quarter, we defined the high-risk population to consist of beneficiaries whose HCC scores were in the top third among all observable Medicare beneficiaries assigned to the treatment panels at the start of the intervention period.

3. Comparison group definition

The comparison group consists of Medicare FFS beneficiaries whom we assigned to 42 matched comparison panels in each of the baseline and intervention quarters. Through our definition of the potential comparison panels, and then through statistical matching techniques to further refine this list to a set of final comparison panels, we selected comparison panels that were similar to the treatment panels during the baseline period on factors that can influence patients' outcomes, especially those factors that CareFirst used when deciding which panels to recruit for the intervention. This section describes how we constructed the matched comparison group; Section V.C shows the balance we achieved between the two groups on the matching variables.

We identified the 42 comparison panels in four steps:

1. First, at our request, CareFirst provided a list of all 149 panels (of 438) in the commercial program that met the following criteria that all 14 treatment panels also met: (1) located in Maryland, (2) joined the commercial PCMH program when it began in 2011, and (3) served at least 1,000 CareFirst members in 2012.
2. Second, we developed matching variables, defined at the start of the intervention (August 1, 2013), for all treatment and potential comparison panels. These variables included characteristics of the panel overall (for example, the number of PCPs in the panel and the panel's quality and financial performance in the commercial PCMH program); characteristics of all Medicare FFS beneficiaries assigned to the panels (for example, mean HCC score and utilization in the baseline period); and characteristics of high-risk beneficiaries assigned to the panels. We did not include measures of quality-of-care processes in the matching because, when we completed matching (spring 2015), these measures were not yet available. When assigning Medicare beneficiaries to the panels, we used the same attribution and panel assignment logic that we used for the treatment panels, as described previously. Section V.C describes the matching variables and their data sources in detail.

3. Third, we narrowed the pool of 149 to 101 potential comparison panels that, like the treatment panels, (1) had an average of at least 500 assigned Medicare FFS beneficiaries during the four baseline quarters, (2) had at least five PCPs at the start of the intervention, and (3) were located in urban areas.
4. Finally, we used propensity-score methods to select 42 comparison panels from the pool of 101 that were similar to the 14 treatment panels on the matching variables. The propensity score is the predicted probability, based on all of a panel's matching variables, that a given panel was selected for treatment (Stuart 2010). It collapses all of the matching variables into a single number for each panel that can be used to assess how similar panels are to one another. By matching each treatment panel to one or more comparison panels with similar propensity scores, we generated a comparison group that is similar, on average, to the comparison group on the matching variables. The approach, however, does not ensure that each comparison panel matches exactly to its treatment panel on all matching variables. We prioritized one matching variable—whether a panel is virtual—by requiring that a virtual treatment panel could match only to a virtual comparison panel, and a nonvirtual treatment panel could match only to a nonvirtual comparison panel. Such panels were likely to have fewer resources, and greater coordination challenges, than the nonvirtual panels, which were part or all of a single, larger practice.

We required each treatment panel to match to at least one, but no more than seven, comparison panels and that the overall ratio of comparison to treatment panels be 3:1. This matching ratio increases the statistical certainty in the impact estimates (relative to a 1:1 overall matching ratio), because it creates a more stable comparison group against which to compare the treatment group's experiences.

After completing the matching, we assigned Medicare FFS beneficiaries to the comparison practices in each intervention quarter using the same rules we used for the treatment group (Section V.A.2). We also defined a high-risk subgroup of the comparison group using the same rules as for the treatment group. That is, a beneficiary was in the high-risk group in the intervention quarter if his or her HCC score at the start of the intervention period was in the top third among all observable Medicare beneficiaries assigned to the treatment panels at the start of the intervention period.

Our decision to select comparison panels from the pool of CareFirst panels not participating in the intervention, rather than panels or practices external to CareFirst, reflects CMS's goal of estimating the marginal effect of HCIA funding on patients' outcomes. That is, we aimed to estimate the impacts of expanding CareFirst's PCMH program to Medicare FFS, not CareFirst's PCMH program as a whole. It is possible that, before the start of the HCIA program, the commercial program had some positive spillover effects for Medicare patients. For example, if PCPs developed more cost-effective referral patterns, this might have reduced the total cost of care for all of their patients, not only commercial patients. However, any such spillover does not contaminate our impact estimates because we intended to estimate the marginal impact of HCIA funding, separate from any positive spillover effects that might have existed without HCIA funding. Further, we anticipate any such spillover to be small, because the largest intervention

component was care coordination for high-risk beneficiaries, and providing care coordination for one participant was likely to have little influence on care for other patients the panel served.

4. Construction of outcomes and covariates

We used Medicare claims from August 1, 2009, to July 31, 2015, for beneficiaries assigned to the treatment and comparison panels to develop two types of variables: (1) outcomes, defined for each beneficiary in each baseline or intervention quarter; and (2) covariates, which describe a beneficiary's characteristics at the start of the baseline and intervention periods and are used in the regression models for estimating impacts to adjust for beneficiaries' characteristics before the period began. We used covariates defined at the start of each period, without updating them each quarter, to avoid controlling in each intervention quarter for previous quarters' program effects, as this would bias the effect estimates away from detecting true impacts. Appendix 1 provides details on the methods we used to construct these variables.

Outcomes. For each beneficiary, we calculated eight outcomes that we grouped into four domains:

1. Domain: Quality-of-care processes
    a. Diabetes quality-of-care composite (binary variable for each beneficiary); calculated as whether a beneficiary with diabetes had had all four recommended tests—lipid profile, hemoglobin A1c test, dilated eye exam, and nephropathy screening—during the previous 12 months
    b. IVD lipid profile (binary variable for each beneficiary); calculated as whether a beneficiary with IVD had a complete lipid profile during the previous 12 months
    c. Ambulatory-care follow-up visit within 14 days of a hospital discharge (binary variable for each beneficiary); calculated as whether all of an individual's discharges in a quarter were followed by an ambulatory visit with a primary care or specialist physician within 14 days of the discharge
2. Domain: Quality-of-care outcomes
    a. Inpatient admissions (number/quarter) for ambulatory care-sensitive conditions (ACSCs)
    b. Number of inpatient admissions followed by an unplanned readmission within 30 days (number/quarter)
3. Domain: Service use
    a. All-cause inpatient admissions (number/quarter)
    b. Outpatient ED visit rate (number/quarter); outpatient ED visits are defined as ED visits or observational stays that do not end in a hospital admission
4. Domain: Spending
    a. Total Medicare Part A and B spending (dollars/month)

Four of these outcomes—all but ACSC admissions and the three quality-of-care process measures—are outcomes that CMMI has specified as "core" for the evaluations of all HCIA programs. Our definition of the readmission measure, however, differs from CMMI's standard definition. CMMI typically defines readmissions as the proportion of inpatient admissions that end in an unplanned readmission. Instead, we analyzed impacts on the *number* of these unplanned readmissions across all beneficiaries per quarter, because this enables us to look at the total impact on readmissions across the treatment group, rather than readmissions contingent on an inpatient admission. We made this decision, in consultation with CMMI, because the intervention might also affect the number of and type of admissions.

All outcomes are quarter-specific—meaning that we calculated them for each baseline and intervention quarter separately—except for the two quality-of-care process measures for IVD and diabetes. Because these two measures assess whether a beneficiary received recommended preventive care services over a year-long period, we calculated these measures over full years rather than quarters: for example, over the baseline year (that is, the period corresponding to the four baseline quarters), over the first year of the intervention period (corresponding to the first four intervention quarters), and so on. We avoided calculating these measures for overlapping periods, meaning that no measurement year included services provided in another measurement year.

Finally, we defined all outcomes for all treatment and comparison group members, except for the three measures of quality-of-care processes. We calculated the measure of 14-day follow-up after discharge among only those patients with at least one hospital discharge in the relevant quarter. We calculated the diabetes composite measure among beneficiaries ages 18 to 75 with diabetes at the beginning of the period (baseline or intervention period), and calculated the measure of lipid screening among beneficiaries ages 18 or older with IVD at the beginning of the period.

Covariates. The covariates include (1) 18 indicators for whether a patient has each of the following chronic conditions: heart failure, chronic obstructive pulmonary disease, chronic kidney disease, diabetes, Alzheimer's and related dementia, depression, ischemic heart disease, cancer, asthma, hypertension, atrial fibrillation, stroke, hyperlipidemia, hip fracture, osteoporosis, rheumatoid arthritis, bipolar disorder, and schizophrenia; (2) HCC score; (3) demographics (age, gender, and race or ethnicity); and (4) original reason for Medicare entitlement (old age, disability, or end-stage renal disease). We defined all covariates as of the start of the relevant period (baseline or intervention).

5. Regression model

We used a regression model to implement the difference-in-differences design for estimating impacts. For each outcome, the model estimates the relationship between the outcome and a series of predictor variables, assuming that each of the predictor variables has a linear (additive) relationship with the outcome. The predictor variables include the patient-level covariates (defined in Section V.A.4); whether the patient is assigned to a treatment or a comparison panel; an indicator for each panel (which accounts for differences between panels in their patients' outcomes at baseline); indicators for each post-intervention quarter (or, for the diabetes and IVD

measures, for the final post-intervention quarter of the year-long measurement period); and an interaction of a beneficiary's treatment status with each post-intervention quarter (or, for the diabetes and IVD measures, the final post-intervention quarter of the year-long measurement period).

The estimated relationship between the interaction term and the outcome in a given quarter is the impact estimate for that quarter (or, for the diabetes and IVD measures, for the year ending with that quarter). It measures the average difference between outcomes for beneficiaries assigned to the treatment and comparison panels during that period, subtracting out any differences between these groups during the four baseline quarters. By providing separate impact estimates for each intervention quarter (or year, for the diabetes and IVD measures), the model enables the program's impacts to change the longer the panels are enrolled in the program. We can also test impacts over discrete sets of quarters or years, which is needed to implement the primary tests discussed in the next section. Finally, the model quantifies the uncertainty in the impact estimates, allowing for statistical tests that determine whether observed differences in outcomes between the treatment and comparison groups are likely due to chance. The model uses robust standard errors to account for clustering of outcomes across quarters for the same beneficiary and a dummy variable for each panel (fixed effects) to account for clustering of outcomes for beneficiaries assigned to the same panel. Appendix 2 provides details on the regression methods, including descriptions of the weights each beneficiary receives in the model.

6. Primary tests

Table V.1 shows the primary tests for CareFirst, by domain. Each test specifies a population, outcome, time period, expected direction of effect, and threshold that we count as substantively important. The purpose of these primary tests is to focus the impact evaluation on hypotheses that will provide the most robust evidence about program effectiveness (see Appendix 3 for detail and a description of how we selected each test). We provided both the awardee and CMMI an opportunity to comment on the primary tests.

Because the third annual report is designed to assess impacts during the original award period only (through June 2015), we conducted the primary tests only partially in this report. Specifically, we estimated impacts for the fifth through eighth intervention quarters (August 2014 through July 2015), and did not include the 9th and 10th intervention quarters. We will present final results, including the final quarters of the intervention period, in a future addendum to this report..

Our rationale for selecting these primary tests is as follows:

- **Outcomes.** CareFirst's central goal was to reduce hospitalizations, ED visits, and Medicare Part A and B spending, so our primary tests address these three outcomes. In addition, the primary tests address two quality-of-care outcomes the intervention is expected to affect: ACSC admissions and 30-day unplanned hospital readmissions. Finally, we include three quality-of-care process measures that, based on CareFirst's theory of action, we think the program could improve: (1) a composite measure for whether a beneficiary with diabetes received all of four recommended processes of care during the year (HbA1c test, lipid

profile, dilated eye exam, and nephropathy screening); (2) receipt of a complete lipid profile for people with IVD; and (3) receipt of a follow-up ambulatory care visit with a primary care or specialist provider within 14 days of hospital discharge. Although CareFirst did not set explicit targets for these particular quality-of-care process measures, the OIAs incentivized improvements in processes of care for chronic illnesses and the care transitions intervention could be expected to improve 14-day follow-up rates.

Table V.1. Specification of the primary tests for CareFirst Blue Cross Blue Shield

Domain (number of tests in the domain)[a]	Outcome (units)	Time period for impacts (controlling for baseline differences)[b]	Population	Substantive threshold (expected direction of effect)[c]
Quality-of-care processes (3)	Received all four recommended diabetes processes of care in the year (binary [yes or no]/beneficiary/year)	Second intervention year (corresponding to intervention quarters 5 through 8)[d]	Medicare FFS beneficiaries ages 18 to 75 with diabetes and assigned to treatment panels	15.0% (+)
	Received lipid profile in the year (binary [yes or no]/beneficiary/year)	Second intervention year (corresponding to intervention quarters 5 through 8)[d]	Medicare FFS beneficiaries aged 18 or older with IVD and assigned to treatment panels	15.0% (+)
	All inpatient admissions within a quarter were followed by an ambulatory care visit with a primary care or specialist provider within 14 days (binary [yes or no]/beneficiary/year)	Average over intervention quarters 5 through 10[d]	Medicare FFS beneficiaries with at least one hospital stay in the quarter and assigned to treatment panels	15.0% (+)
Quality-of-care outcomes (2)	Inpatient admissions for ambulatory care-sensitive conditions (#/beneficiary/quarter)	Average over intervention quarters 5 through 10[d]	Medicare FFS beneficiaries assigned to treatment panels	5.0% (-)
	30-day unplanned hospital readmissions (#/beneficiary/quarter)	Average over intervention quarters 5 through 10[d]	Medicare FFS beneficiaries assigned to treatment panels	5.0% (-)
Service use (2)	All-cause inpatient admissions (#/beneficiary/quarter)	Average over intervention quarters 5 through 10[d]	Medicare FFS beneficiaries assigned to treatment panels	5.0% (-)
	Outpatient ED visit rate (#/beneficiary/quarter)	Average over intervention quarters 5 through 10[d]	Medicare FFS beneficiaries assigned to treatment panels	5.0% (-)
Spending (1)	Medicare Part A and B spending ($/beneficiary/month)	Average over intervention quarters 5 through 10[d]	Medicare FFS beneficiaries assigned to treatment panels	4.0% (-)

[a] We will adjust the *p*-values from the primary test results for the multiple comparisons made within each domain, but not across domains.

[b] The regression models for estimating program impacts will control for differences in outcomes between the pre-intervention treatment and comparison groups.

[c] The substantive threshold is the impact as a percentage of the counterfactual. The counterfactual is the outcome the treatment group would have had in the absence of the HCIA-funded intervention.

[d] For all but two of the measures, we will take the average across 6 quarterly impact estimates (one for each intervention quarter from 5 through 10). For the diabetes and IVD process of care measures, we will use a single impact estimate—those for the second program year (corresponding to intervention quarters 5 through 8).

ED = emergency department; FFS = fee-for-service; HCIA = Health Care Innovation Award; IVD = ischemic vascular disease.

- **Time period.** CareFirst expected participating panels to show substantial impacts by their second year in the program. For this reason, our primary tests cover the 1.5 years from August 2014 through January 2016 (or the 5th through 10th intervention quarters [I5 through I10]), a period that began one year after the program started in August 2013. The final impact analysis will include one month beyond the intervention end date (December 31, 2015) so that we can include outcomes for the quarter that runs from November 2015 to January 2016, most of which falls during the program's operational period. Most of the measures are defined quarterly, so our impact estimates represent averages across relevant quarters. In contrast, because the quality-of-care process measures for IVD and diabetes are defined over a year, our primary tests assess impacts during the second full year of program operations (a period corresponding to I5 through I8).
- **Population.** For all but the three quality-of-care process measures, the population includes all Medicare FFS beneficiaries (excluding those dually eligible for Medicaid) assigned to the 14 treatment panels. This corresponds to CareFirst's definition of its target population. For the diabetes and IVD quality-of-care process measures, we limit the population to beneficiaries ages 18 to 75 with diabetes or ages 18 and older with IVD, respectively, and who were observable in FFS claims for all 12 months of the measurement year. For the 14-day follow-up measure, we limit the sample in each quarter to those who had at least one qualifying hospitalization during the quarter for which we could observe whether the beneficiary had a 14-day follow-up visit.
- **Direction (sign) of the impact estimate.** For the quality-of-care process measures, we expect the impact estimate to be positive, signaling an increase in the percentage of people receiving recommended care. For all other outcomes, we expect the impact estimates to be negative, indicating a reduction in service use or overall expenditures.
- **Substantive thresholds.** Some impact estimates could be large enough to be policy relevant (to CMMI and other stakeholders) even if they are not statistically significant; for this reason, we have prespecified thresholds for what we call substantive importance. We express the threshold as a percentage change from the counterfactual—that is, the outcomes that beneficiaries in the treatment group would have had if they had not received the HCIA-funded intervention. For all but the quality-of-care process measures, the 4 to 5 percent thresholds we chose (depending on the outcome) are 75 percent of CareFirst's expected effects during the primary test period (I5 through I10). (We use 75 percent recognizing that CareFirst could still be considered successful if it approached, but did not fully achieve, its anticipated effects.) The 15 percent threshold for the quality-of-care process measures is extrapolated from the literature (Peikes et al. 2011; Rosenthal et al. 2016) because CareFirst did not specify by how much it expected to improve these outcomes.

7. Secondary tests (robustness checks)

We also conducted secondary quantitative tests to help corroborate the findings from the primary tests. This is important because some of the differences observed between the treatment and comparison groups in the primary test results could reflect limitations of the non-experimental impact evaluation design or random fluctuations in the data. We have greater

confidence in the primary results if they are generally consistent with the expected broader pattern of results from the secondary tests.

We conducted three sets of secondary tests for CareFirst.

1. First, we estimated the program's impacts on all-cause admissions and total Medicare spending during two intervention periods in addition to those specified in the primary tests: (1) the first 6 months after the panels joined the intervention (I1 and I2), and (2) months 7 to 12 after the panels joined the intervention (I3 and I4). Because we and CareFirst expected program impacts to increase over time, with little or no impacts in the first few months of the program, the following pattern would be highly consistent with an effective intervention: little to no measured effects in the first two quarters, growing effects in quarters 3 and 4, and the largest impacts in quarters 5 through 10. (The primary tests conducted in this report cover I5 through I8). In contrast, if we found very large differences in outcomes (favorable or unfavorable) in the first 6 intervention months, this could suggest a limitation in the comparison group, not true intervention impacts.
2. Second, we reran all of the primary tests, limiting the sample only to high-risk Medicare FFS beneficiaries, defined by their HCC scores at baseline (Section V.A.2). The program's theory of action suggests that, if the intervention did have favorable impacts, these impacts should have been concentrated among high-risk beneficiaries because (1) they were more likely to have hospitalizations and other acute events that the program's services could help prevent and (2) CareFirst targeted high-risk beneficiaries for care coordination services. Therefore, if we were to find favorable impacts for the full population, we would expect these impacts to be larger for the higher-risk subset of the treatment group. Conversely, if we were to find substantively large *unfavorable* results for the full population, these secondary tests would enable us to assess whether the effects were similarly unfavorable in the high-risk population only. Evidence that outcomes were better for the high-risk group than the full population could signal that resources were diverted from lower-risk beneficiaries to serve the higher-risk group.
3. Third, we reestimated impacts on admissions and spending among the full Medicare FFS population (that is, not high-risk only), but limiting to beneficiaries assigned to the treatment and comparison groups by the start of the period, either baseline or intervention. This restriction prevents addition to the intervention sample over time. It is possible that differences in sample addition between the treatment and comparison groups could bias the impact results to some degree if the sample members added over time differ from earlier sample members (for example, they are younger and healthier); this could create differences in mean outcomes between the treatment and comparison groups that are unrelated to the HCIA intervention. We have explored this possibility because, as we will describe in Section V.D.1, the rate of net sample growth during the intervention period was slightly higher for the comparison group (growth of 22.7 percent from I1 to the I8) than for the treatment group (growth of 19.3 percent). We believe differences in sample addition drive the differences in net sample growth because the rate of sample loss was the same during the intervention period. That is, the percentage of beneficiaries assigned to treatment and comparison groups in the first quarter but lost to follow-up (due to death, movement into

managed care, or movement out of state) by the end of the intervention period is exactly the same (11.0 percent) in the treatment and comparison groups.

8. Synthesizing evidence to draw conclusions

Within each domain, we drew one of five conclusions about program effectiveness based on the primary test results, the results of secondary tests, and the plausibility of those findings given the implementation evidence:

1. Statistically significant favorable effect (the highest level of evidence)
2. Substantively important (but not statistically significant) favorable effect
3. Substantively important (but not statistically significant) unfavorable effect
4. No substantively large effect
5. Indeterminate effect

We cannot conclude that a program has a statistically significant unfavorable effect because, in consultation with CMMI, we decided to use one-sided statistical tests (which do not test for evidence of unfavorable effects). We used one-sided tests to increase the probability that, if a program truly did have impacts, we would be able to detect them.

Appendix 3 describes our decision rules for each of the five possible conclusions. In short, we concluded that a program had a statistically significant favorable effect in a domain if (1) at least one primary test result in the domain was favorable and statistically significant, after adjusting the statistical tests to account for multiple tests (if applicable) within a domain; or (2) the average impact estimate across all primary tests in the domain was favorable and statistically significant. In both cases, we also had to determine that the primary test results were plausible given the results of the secondary tests and implementation evidence. We concluded that a program had a substantively important favorable effect if the average impact estimate in the domain was substantively important but not statistically significant, and if the result was plausible given the secondary tests and implementation evidence. In contrast, if the average impact estimate was unfavorable (opposite the hypothesized direction), larger than the substantive threshold, and unfavorable effects were plausible given the other evidence, we concluded the program had a substantively important unfavorable effect. If the tests in a domain did not meet any of these criteria, we drew one of two conclusions. First, if the tests for at least one outcome in the domain (or all outcomes in the domain together) had sufficient statistical power to detect an impact of the size of the substantive threshold with at least 75 percent probability, we concluded there was not a substantively large effect because we are reasonably confident that we would have detected such an effect had there been one. Alternatively, if the power was not sufficient (less than 75 percent) to detect this type of impact, we concluded the impact in the domain was indeterminate. Indeterminate means either that the program truly did not have effects that were substantively large, or that it did, but our statistical tests were not able to detect them.

B. Characteristics of the treatment group at baseline

This section describes the characteristics of the treatment group at the start of the intervention (August 1, 2013). We also show this information in the second column of Table V.2. (Table V.2 serves a second purpose—to show the equivalence of the treatment and comparison panels at the start of the intervention—which we describe in Section V.C.)

Characteristics of the panels overall. At the start of the intervention, the 14 treatment panels, on average, consisted of nine PCPs each. Half of the panels were virtual, meaning they consisted of several small practices that joined contractually to participate in CareFirst's commercial PCMH program. This proportion is consistent with CareFirst's overall commercial program, in which about half of the 438 panels are virtual. Health systems owned 2 of the 14 treatment panels, again consistent with the proportion (15 percent) of panels that are of this type in the commercial program. Consistent with CareFirst's stated selection criteria for the panel, the treatment panels performed well in the commercial program in 2011 and 2012, achieving an average 4 percent savings against expected 2011–2012 care costs and an average quality score over those two years of 68 out of 100. In contrast, the average savings across the 101 panels in the potential comparison pool was 2 percent, and the mean quality score was 64 (comparable data are not available for all 438 panels in the commercial program). The treatment panels practiced in relatively affluent zip codes, where the median household income was almost $78,000 from 2008 to 2012 (compared with a national average of $53,046).

Characteristics of the panels' Medicare FFS beneficiaries. Treatment patients' characteristics were similar to the nationwide Medicare FFS averages. Among all Medicare beneficiaries not dually eligible for Medicaid and assigned to the treatment panels during the baseline period (August 1, 2012, through July 31, 2013), the HCC risk score was 1.1, close to the national average of 1.0. Patients in the treatment panels also had hospital admission rates, total Medicare spending, and 30-day readmission rates that were close to the national averages during the baseline period. The mean outpatient ED visit rate (81/1,000 people/quarter) was lower than the national average of 105, which could in part be due to the fact that the treatment group excludes those dually enrolled in Medicare and Medicaid, who often have high outpatient ED visit rates (Congressional Budget Office 2013).The high-risk Medicare FFS beneficiaries assigned to the treatment panels had substantially greater health care needs during the baseline period than the full treatment group. For example, their mean HCC risk score was about twice the mean for all treatment group members (2.0 versus 1.1).

Table V.2. Characteristics of treatment and comparison panels before the intervention start date (August 1, 2013)

Characteristic of panel	Treatment panels (N = 14)	Matched comparison panels (N = 42)	Absolute difference[a]	Standardized difference[b]	Medicare FFS national average
Exact match variable[c]					
Characteristics of the panel overall					
Panel type: Virtual (%)	50.0	50.0	0	0	n.a.
Propensity-matched variables[d]					
Characteristics of the panel overall					
Average quality score for the commercial program in 2011 and 2012[e]	68.1	66.4	1.64	0.239	n.a.
Average cost savings in the commercial program in 2011 and 2012 (%)[f]	3.9	3.2	0.7	0.190	n.a.
PCPs in panel who work in practices that are medical homes (%)	34.7	29.6	5.1	0.156	n.a.
Panel type: Health system (%)	14.3	8.2	6.1	0.237	n.a.
Number of PCPs	9.29	8.53	0.76	0.263	n.a.
Characteristics of a panel's practice(s) location(s)					
Median household income in zip code(s) where panel's practice(s) are located ($)	77,982	78,406	-424	-0.020	53,046[g]
Characteristics of all Medicare FFS, nondually eligible patients assigned to panels during the baseline year (August 1, 2012 – July 31, 2013)					
Number of beneficiaries	2,202	1352	850	1.208	n.a.
HCC risk score	1.08	1.07	0.01	0.082	1.0
All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	79.87	79.22	0.65	0.044	74[h]
Outpatient ED visit rate (#/1,000 beneficiaries/quarter)	81.33	82.66	-1.33	-0.082	105[i]
Medicare Part A and B spending ($/beneficiary/month)	998	988	10	0.073	860[j]
30-day unplanned hospital readmission rate (%)	15.4	15.7	-0.3	-0.108	16.0[k]
30-day unplanned hospital readmissions (#/beneficiary/quarter)[l]	10.96	10.81	0.16	0.047	n.a.
Inpatient admissions for ambulatory care-sensitive conditions (#/1,000 beneficiaries/quarter)[l]	13.28	12.92	0.36	0.094	11.8[m]
Disability as original reason for Medicare entitlement (%)	11.2	10.8	0.4	0.111	16.7[n]
Age (years)	73.84	73.87	-0.03	-0.022	71[o]
Female (%)	59.2	58.7	0.5	0.137	54.7[n]
Race: White (%)	85.1	82.0	3.2	0.207	81.8[n]

Table V.2 *(continued)*

Characteristic of panel	Treatment panels (N = 14)	Matched comparison group (N = 42)	Absolute difference[a]	Standardized difference[b]	Medicare FFS national average
Characteristics of high-risk Medicare FFS, nondually edligible patients assigned to panels during the baseline year (August 1, 2012 – July 31, 2013)					
Number of high-risk beneficiaries	693	427	266	1.043	n.a.
HCC risk score	2.00	2.00	0.01	0.084	1.0
All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	160.58	157.88	2.70	0.127	74
Outpatient ED visit rate (#/1,000 beneficiaries/quarter)	136.44	139.26	-2.82	-0.103	105
Medicare Part A and B spending ($/beneficiary/month)	1,843	1,832	11	0.050	860
30-day unplanned hospital readmission rate (%)	18.3	18.2	0.1	0.031	16.0
30-day unplanned hospital readmissions (#/beneficiary/quarter)[l]	25.96	25.16	0.80	0.110	NA
Inpatient admissions for ambulatory care-sensitive conditions (#/beneficiary/quarter)[l]	32.29	30.57	1.73	0.254	11.8
Variables not included in matching[p]					
Characteristics of Medicare FFS, nondually eligible patients assigned to panels during the baseline year who met diagnosis, age, and/or service use restrictions					
Receipt of all four recommended diabetes process of care measures, among those with diabetes ages 18 to 75 (%)	47.0	43.2	3.8	0.40	NA
Receipt of recommended lipid profile, among those with IVD ages 18 or older (%)	79.4	77.0	2.3	0.38	NA
Receipt of an ambulatory care visit within 14 days of all hospital discharges in the quarter, among those with at least one discharge in the quarter (%)	64.3	62.8	1.5	0.47	NA

Sources: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at CMS. Zip code household income data merged from the American Community Survey ZIP Code Characteristics. CareFirst provided data on characteristics of the panels, including quality scores and financial performance in the commercial program.

Notes: The comparison group means are weighted based on the number of matched comparison panels per treatment panel. For example, if four comparison panels are matched to one treatment panel, each of the four comparison panels has a matching weight of 0.25.

Absolute differences might not be exact due to rounding.

We did not audit or independently confirm the quality or financial performance scores that CareFirst reported for the panels in the commercial medical home program.

[a] The absolute difference is the difference in means between the treatment and matched comparison groups.

[b] The standardized difference is the difference in means between the treatment and matched comparison groups divided by the standard deviation of the variable. The standard deviation is calculated among the pooled treatment and matched comparison groups.

[c] Exact match means that a virtual treatment panel could be matched only to a virtual comparison panel, and a nonvirtual treatment panel could be matched only to a nonvirtual comparison panel.

[d] Variables that we matched on through a propensity score, which capture the relationship between a panel's characteristics and its likelihood of being in the treatment group.

[e] Average quality score for CareFirst's commercial program for 2011 and 2012. The quality score is out of 100 points.

Table V.2 *(continued)*

[f] Average financial performance in the commercial program is a function of credits (global projected care costs) minus debits (all services paid) for 2011 and 2012.

[g] U.S. Census Bureau, 2008–2012 American Community Survey, median household income.

[h] Health Indicators Warehouse (2014b).

[i] Gerhardt et al. (2014).

[j] Boards of Trustees (2013).

[k] Centers for Medicare & Medicaid Services (2014).

[l] These measure are included on the table for descriptive purposes but were not included in the matching model.

[m] This rate is for individuals ages 65 and above (Truven Health Analytics 2015).

[n] Chronic Conditions Data Warehouse (2016a, Table A.1).

[o] Health Indicators Warehouse (2014a).

[p] These baseline process of care measures were not available at the time we conducted matching.

CMS = Centers for Medicare & Medicaid Services; ED = emergency department; FFS = fee-for-service; HCC = Hierarchical Condition Category; PCP = primary care provider.

NA = not available.

n.a. = not applicable.

C. Equivalence of treatment and comparison groups at baseline

Demonstrating that the treatment and comparison groups are similar at the start of the intervention is important for the evaluation design. This similarity increases the credibility of a key assumption underlying difference-in-differences models—that the change over time in outcomes for the comparison group is the same change that would have happened for the treatment group, had the treatment group not received the intervention.

Table V.2 shows that the 14 treatment panels and the 42 selected comparison panels were similar at the start of the intervention on most matching variables. By construction, there were no differences between the two groups on the exact matching variable—whether the panel was virtual. There were some differences between treatment group beneficiaries and matched comparison group beneficiaries on the variables we matched through propensity scores, but the standardized differences across the propensity-score matching variables are almost all within our target of 0.25 standardized differences, and most were within 0.15 standardized differences (the 0.25 target is an industry standard; for example, see Institute of Education Sciences [2014]).

On average, the treatment panels had slightly more PCPs (by 0.76 providers) and considerably more attributed Medicare FFS beneficiaries, overall (by 850 beneficiaries) and for the high-risk participants (by 266). However, in discussion with CMMI, we determined that—although these two variables fell outside our preferred standard—it is reasonable to accept the selected comparison group for three reasons. First, we can account for differences in panel size through regression weights in our impact analyses. Second, there is no correlation between the number of attributed beneficiaries and the outcomes during the baseline period (results not shown), so differences in size within the observed range are unlikely to bias the impact results. Third, if there were any systematic differences in outcomes (that do not vary over time) that result from a different number of primary care providers or beneficiaries, the difference-in-differences model would account for them.

The treatment and comparison panels also differed in baseline performance on the three quality-of-care process measures, which—as described in Section V.A.3—we did not include in the propensity-score matching algorithm because the measures were not available at the time of matching. These three measures assess preventive care for those with diabetes, lipid testing for those with IVD, and 14-day follow-up ambulatory care visits for those with a recent hospital discharge. For all three measures, the differences between the treatment and comparison groups exceed our thresholds (with standardized differences ranging from 0.38 to 0.47) and all differences favor the treatment group. That is, the treatment panels had higher measure scores during the baseline period—reflecting higher quality of care—than the comparison panels. However, the absolute differences between the groups were not particularly large. For example, the absolute difference for 14-day follow-up visits was only 1.5 percentage points (64.3 percent for the treatment group and 62.8 for the comparison group). The difference-in-differences model used to estimate impacts assumes that these differences in baseline performance would persist into the intervention period in the absence of the intervention itself.

D. Beneficiaries' outcomes and intervention impacts

In this section, we first present sample sizes and mean outcomes, by quarter, for the treatment and comparison groups. These mean outcomes provide context for understanding the difference-in-differences estimates that follow; however, the differences in mean outcomes are not regression-adjusted and not impact estimates by themselves. Next, we present the results of the primary tests, by domain. Then, we present the results of the secondary tests (robustness checks) and assess whether the primary test results are plausible given the secondary test results and the implementation evidence. We end with conclusions about program impacts in each domain. These conclusions for CareFirst are preliminary because this report covers outcomes only through July 2015, one month after the end of the original award period, whereas CareFirst's HCIA intervention ran through December 2015, as described previously.

1. Sample sizes

The sample sizes for impact estimation differ depending on the outcome. We present sample sizes by domain.

Quality-of-care processes (Table V.3)

- The **diabetes preventive care composite measure** is defined among Medicare FFS beneficiaries with diabetes ages 18 to 75. The sample size for the treatment group and the weighted comparison group ranges from 3,977 to 4,563 across the baseline year and each of the two intervention years. This population accounts for about 15 percent of the total Medicare FFS sample in the treatment and comparison groups.
- The **lipid profile measure for people with IVD** is defined among Medicare FFS beneficiaries with IVD ages 18 or older. The sample size for the treatment group and the weighted comparison group ranges from 9,024 to 9,977 across the baseline year and each of the two intervention years. This population accounts for about 30 percent of the total Medicare FFS sample in the treatment and comparison groups. This percentage is higher than for the diabetes measure because (1) IVD (which is a broad disease category) is more

common than diabetes among the treatment and comparison beneficiaries and (2) the diabetes measure excludes beneficiaries older than 75 but the IVD measure does not.

- The **14-day follow-up measure** is defined among Medicare FFS beneficiaries who have at least one hospital stay in the quarter. For the treatment group, the sample size ranges from 1,711 to 2,140 beneficiaries across the baseline and intervention quarters (accounting for about 6 percent of all treatment beneficiaries in each quarter). For the comparison group, the sample ranges from 3,543 to 4,201 across the baseline and intervention quarters (accounting for a similar proportion of the total comparison group). After weighting the comparison group to account for the larger number of comparison panels than treatment panels and for the difference in panel size between treatment and comparison groups, the comparison group sample sizes are similar to those in the treatment group.

Quality-of-care outcomes, service use, and spending. The sample sizes for all outcomes in these three domains are the same. In the first baseline quarter (B1), the treatment group includes 29,409 beneficiaries assigned to the 14 participating panels and the comparison group includes 59,670 beneficiaries assigned to the 42 comparison panels (Table V.4). The sample sizes increase modestly during the four baseline quarters (by 11 percent from B1 to B4). This net increase indicates that sample addition (due to beneficiaries being newly attributed to the treatment or comparison practices) exceeds sample attrition (due to beneficiaries dying, switching from FFS Medicare to managed care, moving out of state, or enrolling in Medicaid in addition to Medicare). The sample sizes drop modestly from the last baseline quarter to the first intervention quarter, reflecting that the sample definition (Section V.A.2) retains sample members in successive baseline and intervention quarters, even if they are no longer attributed to the treatment or comparison panel, but not between the baseline and intervention periods. The sample increases modestly during the intervention period, again reflecting greater sample addition than attrition over time. The net sample increase during the intervention period is slightly smaller for the treatment group (19.3 percent from I1 to I8) than the comparison group (22.7 percent over the same time period).

Table V.3. Unadjusted mean outcomes (quality-of-care processes) observed among select Medicare FFS beneficiaries, by treatment status and quarter

		Number of Medicare FFS beneficiaries (panels)			Mean outcomes		
Period	Quarter(s)	T	C (not weighted)	C (weighted)	T	C	Difference (%)
Among those with diabetes and ages 18 to 75, the percentage who received all four recommended diabetes processes of care in the year (%/year)							
Baseline	B1–B4[a]	4,155 (14)	8,875 (42)	4,249	47.9	43.6	4.3 (10.0%)
Intervention	I1–I4[a]	4,347 (14)	9,404 (42)	4,563	44.5	44.5	0.1 (0.1%)
	I5–I8[a]	3,977 (14)	8,858 (42)	4,286	45.7	44.1	1.5 (3.5%)
Among those with ischemic vascular disease and ages 18 or older, the percentage who received complete lipid profile in the year (%/year)							
Baseline	B1–B4[a]	9,841 (14)	19,098 (42)	9,024	79.7	77.1	2.6 (3.4%)
Intervention	I1–I4[a]	9,977 (14)	19,498 (42)	9,402	79.2	76.7	2.5 (3.3%)
	I5–I8[a]	9,603 (14)	19,211 (42)	9,220	79.2	77.2	2.0 (2.6%)
Among beneficiaries with at least one inpatient admission in the quarter, the percentage of beneficiaries whose inpatient admissions in the quarter were all followed by an ambulatory care visit with a primary care or specialist provider within 14 days of discharge (%/quarter)							
Baseline	B1	1,711 (14)	3,543 (42)	1,703	63.8	61.3	2.5 (4.1%)
	B2	1,859 (14)	3,779 (42)	1,712	63.4	61.6	1.7 (2.8%)
	B2	1,997 (14)	3,911 (42)	1,932	64.2	64.6	-0.3 (-0.5%)
	B4	1,863 (14)	3,806 (42)	1,836	64.3	64.2	0.1 (0.1%)
Intervention	I1	1,765 (14)	3,323 (42)	1,654	65.1	64.0	1.1 (1.8%)
	I2	1,843 (14)	3,533 (42)	1,740	62.0	59.0	3.0 (5.1%)
	I3	1,865 (14)	3,612 (42)	1,757	64.6	61.8	2.8 (4.5%)
	I4	1,912 (14)	3,784 (42)	1,903	64.5	64.7	-0.2 (-0.2%)
	I5	1,910 (14)	3,828 (42)	1,840	65.0	65.7	-0.7 (-1.0%)
	I6	2,140 (14)	4,154 (42)	1,987	64.3	62.9	1.4 (2.2%)
	I7	1,966 (14)	4,201 (42)	1,947	67.8	65.4	2.4 (3.6%)
	I8	1,969 (14)	4,079 (42)	1,952	68.3	66.9	1.4 (2.1%)

Sources: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at the Centers for Medicare & Medicaid Services.

Table V.3 *(continued)*

Notes: The baseline quarters are measured relative to the start of the baseline period on August 1, 2012. For example, the first baseline quarter (B1) runs from August 1, 2012, to October 31, 2012. The intervention quarters are measured relative to the start of the intervention period on August 1, 2013. For example, the first intervention quarter (I1) runs from August 1, 2013, to October 31, 2013. In each period (baseline or intervention), the treatment group each quarter includes beneficiaries assigned to a treatment panel by the start of the quarter and who met other sample criteria—that is, they were enrolled in FFS Medicare; lived in Maryland or surrounding areas; were not enrolled in Medicaid; and met any restrictions laid out in the measure with respect to age, chronic conditions, or recent hospital admissions. In addition, for the measures of diabetes and ischemic vascular disease, we required beneficiaries to be observable for the full 12 months covered by the measure. In each period (baseline or intervention), the comparison group includes all beneficiaries assigned to a comparison panel by the start of the quarter and who met the other sample criteria. See text for details.

The outcome means were weighted such that (1) each treatment beneficiary gets a weight of 1; and (2) each comparison beneficiary gets a weight that is the product of two weights: (a) a matching weight, equal to the reciprocal of the total number of comparison panels matched to the same treatment panel as the beneficiary's assigned panel, and (b) a practice size weight, which equals the average number of beneficiaries assigned to the matched treatment panel during the four baseline quarters divided by the average number of beneficiaries assigned to the beneficiary's comparison panel over those quarters. The difference between the treatment and comparison groups in a quarter is calculated by subtracting the mean outcome for the comparison group from the mean outcome for the treatment group. The percentage difference equals that difference divided by the mean outcome for the comparison group.

[a] The quality-of-care process measures were calculated over year-long periods, corresponding to the baseline and intervention quarters shown in the table.

C = control; FFS = fee-for-service; T = treatment.

Table V.4. Unadjusted mean outcomes (quality-of-care outcomes, service use, and spending) measured for all Medicare FFS beneficiaries, by treatment status and quarter

	Number of Medicare FFS beneficiaries (panels)			Inpatient admissions for ambulatory care-sensitive conditions (#/1,000/quarter)			30-day unplanned hospital readmissions (#/1,000/quarter)			All-cause inpatient admissions (#/1,000/quarter)			Outpatient ED visit rate (#/1,000/quarter)			Medicare Part A and B spending ($/month)		
Q	T	C (no wgt)	C (wgt)	T	C	Diff (%)	T	C	Diff (%)	T	C	Diff (%)	T	C	Diff (%)	T	C	Diff (%)
Baseline period (August 1, 2012 – July 31, 2013)																		
B1	29,409 (14)	59,670 (42)	29,458	12.5	13.6	-1.0 (-7.7%)	10.7	11.4	-0.7 (-6.5%)	78.7	78.1	0.7 (0.8%)	82.7	82.6	0.1 (0.1%)	$997	$960	$37 (3.8%)
B2	30,613 (14)	62,558 (42)	30,882	13.9	13.0	1.0 (7.6%)	10.1	10.7	-0.6 (-5.9%)	79.6	79.9	-0.4 (-0.5%)	75.5	80.1	-4.6 (-5.8%)	$956	$973	$-16 (-1.7%)
B3	32,132 (14)	64,124 (42)	31,709	14.0	14.1	-0.0 (-0.2%)	10.7	10.9	-0.2 (-1.7%)	81.9	81.5	0.4 (0.6%)	74.7	73.8	1.0 (1.3%)	$1,003	$985	$18 (1.8%)
B4	32,846 (14)	65,894 (42)	32,951	11.5	10.8	0.7 (6.5%)	10.7	9.5	1.2 (12.2%)	77.5	74.8	2.8 (3.7%)	83.8	86.9	-3.1 (-3.5%)	$1,001	$985	$15 (1.5%)
Intervention period (August 1, 2013 – July 31, 2015)																		
I1	31,500 (14)	61,236 (42)	31,140	10.9	11.3	-0.4 (-3.3%)	11.0	8.9	2.1 (23.2%)	77.1	70.9	6.1 (8.6%)	80.4	81.3	-0.9 (-1.1%)	$1,015	$948	$67 (7.1%)
I2	32,855 (14)	64,325 (42)	32,652	11.7	11.8	-0.1 (-1.0%)	10.6	10.3	0.2 (2.4%)	74.6	74.4	0.2 (0.2%)	77.1	74.4	2.7 (3.6%)	$952	$949	$4 (0.4%)
I3	33,525 (14)	66,308 (42)	33,568	11.7	11.7	0.0 (0.0%)	9.8	7.6	2.2 (28.9%)	75.6	70.7	4.9 (6.9%)	78.4	77.5	0.8 (1.1%)	$998	$923	$75 (8.1%)
I4	34,592 (14)	68,675 (42)	34,618	11.7	10.5	1.2 (11.8%)	9.9	7.7	2.2 (28.7%)	74.3	71.7	2.6 (3.6%)	88.8	89.3	-0.5 (-0.5%)	$985	$978	$6 (0.7%)
I5	35,536 (14)	71,730 (42)	35,823	10.9	10.3	0.6 (6.2%)	10.5	8.3	2.2 (26.2%)	72.9	70.7	2.2 (3.2%)	83.6	87.4	-3.8 (-4.3%)	$1,003	$1,024	$-21 (-2.0%)
I6	36,512 (14)	73,699 (42)	36,941	12.2	12.9	-0.7 (-5.6%)	11.0	9.3	1.7 (18.2%)	78.6	72.3	6.3 (8.7%)	83.0	84.6	-1.6 (-1.9%)	$989	$939	$50 (5.4%)
I7	36,943 (14)	74,364 (42)	37,404	11.6	11.1	0.5 (4.7%)	9.3	7.3	2.0 (27.7%)	70.6	68.0	2.6 (3.8%)	79.0	85.8	-6.9 (-8.0%)	$1,037	$971	$66 (6.8%)
I8	37,593 (14)	75,940 (42)	38,218	11.5	9.8	1.8 (18.4%)	8.8	8.8	-0.1 (-1.0%)	69.1	67.4	1.6 (2.4%)	86.2	94.2	-8.0 (-8.5%)	$1,025	$1,002	$23 (2.3%)

Sources: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at the Centers for Medicare & Medicaid Services.

Notes: The baseline quarters are measured relative to the start of the baseline period on August 1, 2012. For example, the first baseline quarter (B1) runs from August 1, 2012, to October 31, 2012. The intervention quarters are measured relative to the start of the intervention period on August 1, 2013. For example, the first intervention quarter (I1) runs from August 1, 2013, to October 31, 2013. In each period (baseline or intervention), the treatment group each quarter includes all beneficiaries assigned to a treatment panel by the start of the quarter and who met other sample criteria—that is, they were enrolled in FFS Medicare, lived in Maryland or surrounding areas, and were not enrolled in Medicaid. In each period, the comparison group includes all beneficiaries who were assigned to a comparison panel by the start of the quarter and who met the other sample criteria. See text for details.

Table V.4 *(continued)*

The outcome means were weighted such that (1) each treatment beneficiary gets a weight of 1; and (2) each comparison beneficiary gets a weight that is the product of two weights: (a) a matching weight, equal to the reciprocal of the total number of comparison panels matched to the same treatment panel as the beneficiary's assigned panel, and (b) a practice size weight, which equals the average number of beneficiaries assigned to the matched treatment panel during the four baseline quarters divided by the average number of beneficiaries assigned to the beneficiary's comparison panel over those quarters. The difference between the treatment and comparison groups in a quarter is calculated by subtracting the mean outcome for the comparison group from the mean outcome for the treatment group. The percentage difference equals that difference divided by the mean outcome for the comparison group.

B = baseline; C = comparison; Diff = difference; ED = emergency department; FFS = fee-for-service; I = intervention; Q = quarter; T = treatment; no wgt = unweighted; wgt = weighted.

2. Mean outcomes for the treatment and comparison groups, by domain and quarter

Quality-of-care processes. For both the treatment and comparison groups, 61.0 to 65.0 percent of beneficiaries who had any hospital stay in a baseline quarter had all of those stays followed by an ambulatory care visit within 14 days of discharge. This percentage increased modestly during the intervention period, so that by I8 the value was 66.9 percent for the comparison group and 68.3 for the treatment group.

During the baseline year, 47.9 percent of treatment and 43.6 percent of comparison beneficiaries with diabetes and ages 18 to 75 received all four recommended processes of care. This percentage increased slightly to 44.1 in the second program year for the comparison group, and it declined to 45.7 for the treatment group.

During the baseline year, 79.7 and 77.1 percent of the treatment and comparison beneficiaries, respectively, ages 18 or older with IVD received the recommended lipid test—and these percentages remained essentially constant in the two program years.

Quality-of-care outcomes. For both the treatment and comparison groups, the number of ACSC admissions dropped from about 13 per 1,000 beneficiaries in B2 to about 11 in B3. For both groups, the rates remained close to 11 for all subsequent baseline and intervention quarters.

For the 30-day unplanned readmissions measure, the rates steadily declined during the baseline and intervention periods for the comparison group (from 11.4 per 1,000 beneficiaries in B1 to 8.8 in I8). The rates did not decline as steadily for the treatment group and, as a result, the treatment group had substantially (18 to 28 percent) higher rates of readmissions in six of the eight intervention quarters.

Service use. All-cause inpatient admissions generally declined for both the treatment and comparison groups from B3 to I8 (by 13 to 15 percent). However, there was no decline in all-cause admissions during the intervention period for the robustness check that prevented sample addition (data not shown). This suggests that the decline during the intervention period for the full sample was driven by a change in population composition over time—that is, relatively healthy beneficiaries with lower hospitalization rates entering the sample, as opposed to reductions in hospitalization rates among the initial population. Inpatient admissions were modestly higher (0.2 to 8.7 percent higher) for the treatment group than the comparison group in all but one quarter, without any consistent trend of increasing or decreasing differences.

The outpatient ED visit rates fluctuated over time and were generally similar between the treatment and comparison groups, with the treatment group having moderately lower rates (by 8.0 to 8.5 percent) in the last two intervention quarters.

Spending. There was no clear trend in the differences in mean Medicare Part A and B spending over time for the comparison group compared with the treatment group. The difference was -2.0 to +7.1 percent in all baseline and intervention quarters.

3. Results for primary tests, by domain

Overview. The primary tests conducted for this report cover the full primary test period for two quality-of-care process measures (for these measures, the test period is the second program year). For all other measures, the estimates presented in this report are considered preliminary because they reflect four (I5 through I8) of the six planned quarters (I5 through I10) for the final primary tests. An addendum to this report will present results from the full primary test period.

For three of the study domains—quality-of-care processes, service use, and spending—the regression-adjusted differences between the treatment and comparison groups were small (Table V.5). None of these differences were statistically significant or larger than the substantive thresholds in either a favorable or an unfavorable direction. In contrast, in the quality-of-care outcomes domain, we found substantively large and *unfavorable* differences between the treatment and comparison groups. The large standard errors for the estimates in this domain (relative to the point estimates), however, mean that the unfavorable impact is estimated imprecisely and thus may be due to chance alone.

Quality-of-care processes. The likelihood of receiving recommended processes of care for diabetes or IVD was 5.7 and 1.0 percent lower, respectively, for the treatment group (an unfavorable estimate) than the estimated counterfactual. (Our estimated counterfactual—the outcome the treatment group members would have had in the absence of the HCIA intervention—is the treatment group mean minus the difference-in-differences estimate.) We do not consider these unfavorable point estimates to be substantively large because both are smaller than the substantive threshold for these outcomes of 15 percent. We cannot conclude whether these unfavorable results are statistically significant because our one-sided statistical tests are designed only to assess improvements in outcomes. The likelihood of receiving an ambulatory care visit within 14 days of hospital discharge was 0.4 percent higher in the treatment group than its estimated counterfactual, a (favorable) difference that was neither substantively large nor statistically significant. The combined estimate across the three measures in the quality-of-care processes domain was -2.1 percent, an unfavorable point estimate that was not substantively large. The statistical power to detect substantively large effects was good (more than 99 percent) for all three quality-of-care process measures individually and, in addition, combined across the measures.

Quality-of-care outcomes. The rate of ACSC admissions for the treatment group during the primary test period was 3.7 percent higher than our estimate of the counterfactual, and the rate of unplanned readmissions was 16.3 percent higher. These higher rates for the treatment group are in the unfavorable direction (indicating an increase in ACSC admissions and readmissions). The difference is not substantively large for ACSC admissions, but it is for 30-day readmissions (the threshold for each measure is 5 percent). After combining results across the two outcomes in this domain, the combined effect was 10 percent, larger than the substantive threshold of 5 percent and in the unfavorable direction.

Table V.5. Results of primary tests for CareFirst

Primary test definition					Statistical power to detect an effect that is[a]		Results			
Domain (# of tests in domain)	Outcome (units)	Time period for impacts	Population	Substantive threshold (expected direction of effect)[b]	Size of the substantive threshold	Twice the substantive threshold[c]	Treatment group mean	Regression-adjusted difference between the treatment and estimated counterfactual[b] (standard error)	Percentage difference[d]	*p*-value[e]
Quality-of-care processes (3)	Received all four recommended diabetes processes of care in the year (binary [yes or no]/beneficiary/year)	Second intervention year (August 1, 2014, to July 2015)[f]	Medicare FFS beneficiaries ages 18 to 75 with diabetes assigned to treatment panels	15.0% (+)	> 99.0%	> 99.0%	45.7	-2.8% (1.5)	-5.7%	0.90
	Received complete lipid profile in the year (binary [yes or no]/beneficiary/year)	Second intervention year (August 1, 2014, to July 2015)[f]	Medicare FFS beneficiaries ages 18 or older with ischemic vascular disease assigned to treatment panels	15.0% (+)	> 99.0%	> 99.0%	79.2	-0.8% (0.9)	-1.0%	0.63
	All inpatient admissions within a quarter were followed by an ambulatory care visit with a primary care or specialist provider within 14 days (binary [yes or no]/beneficiary/year)	Average over intervention quarters 5–8 (August 1, 2014, to July 2015)[f]	Medicare FFS beneficiaries with at least one hospital stay in the quarter assigned to treatment panels	15.0% (+)	> 99.0%	> 99.0%	66.3	+0.3% (1.3)	+0.4%	0.50
	Combined	Varies by outcome	Varies by outcome	15.0% (+)	> 99.0%	> 99.0%	n.a.	n.a.	-2.1%	0.95
Quality-of-care outcomes (2)	Inpatient admissions for ambulatory care-sensitive conditions (#/1,000 beneficiaries/quarter)	Average over intervention quarters 5–8 (August 1, 2014, to July 2015)[f]	All Medicare FFS beneficiaries assigned to treatment panels	5.0% (-)	28.1%	55.0%	11.6	0.4 (0.8)	+3.7%	0.58

Table V.5 *(continued)*

Primary test definition					Statistical power to detect an effect that is[a]		Results			
Domain (# of tests in domain)	Outcome (units)	Time period for impacts	Population	Substantive threshold (expected direction of effect)[b]	Size of the substantive threshold	Twice the substantive threshold[c]	Treatment group mean	Regression-adjusted difference between the treatment and estimated counterfactual[b] (standard error)	Percentage difference[d]	*p*-value[e]
	30-day unplanned readmissions (#/1,000 beneficiaries/quarter)	Average over intervention quarters 5–8 (August 1, 2014, to July 2015)[f]	All Medicare FFS beneficiaries assigned to treatment panels	5.0% (-)	21.3%	37.8%	9.9	1.3 (0.9)	+16.3%	0.90
	Combined (%)	Average over intervention quarters 5–8 (August 1, 2014, to July 2015)[f]	All Medicare FFS beneficiaries assigned to treatment panels	5.0% (-)	25.5%	48.7%	n.a.	n.a.	+10.0%	0.89
Service use (2)	All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	Average over intervention quarters 5–8 (August 1, 2014, to July 2015)[f]	All Medicare FFS beneficiaries assigned to treatment panels	5.0% (-)	60.1%	96.4%	72.8	1.9 (2.3)	+2.6%	0.67
	Outpatient ED visits (#/1,000 beneficiaries/quarter)	Average over intervention quarters 5–8 (August 1, 2014, to July 2015)[f]	All Medicare FFS beneficiaries assigned to treatment panels	5.0% (-)	69.3%	98.9%	82.9	-2.6 (2.4)	-3.1%	0.23
	Combined (%)	Average over intervention quarters 5–8 (August 1, 2014, to July 2015)[f]	All Medicare FFS beneficiaries assigned to treatment panels	5.0% (-)	79.2%	99.8%	n.a.	n.a.	-0.2%	0.46
Spending (1)	Medicare Part A and B spending ($/beneficiary/month)	Average over intervention quarters 5–8 (August 1, 2014, to July 2015)[f]	All Medicare FFS beneficiaries assigned to treatment panels	4.0% (-)	60.5%	96.5%	$1,014	+$9 ($26.0)	+0.9%	0.64

Table V.5 *(continued)*

Sources: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at the Centers for Medicare & Medicaid Services.

Note: The results for each outcome are based on a difference-in-differences regression model, as described in the text. Estimates are calculated for Medicare beneficiaries who are observable in the relevant time period: that is, beneficiaries who are enrolled in Medicare FFS (Part A and B), are alive, and have Medicare as their primary payer. Additional sample restrictions apply to the quality-of-care process measures; see text for details.

[a] The power calculation is based on actual standard errors from the analysis. For example, in the last row, a 4.0 percent effect on Medicare Part A and B spending (from the counterfactual of $1,014 + $9 = $1,023) would be a change of $41. Given the standard error of $26 from the regression model, we would be able to detect a statistically significant result 60.5 percent of the time if the impact was truly -$41, assuming a one-sided statistical test at the $p = 0.10$ significance level.

[b]The substantive threshold is the impact as a percentage of the counterfactual. The counterfactual is the outcome the treatment group would have had in the absence of the HCIA-funded intervention.Our estimate of the counterfactual is the treatment group mean minus the regression-adjusted difference-in-differences estimate.

[c] We show statistical power to detect a very large effect (twice the size of the substantive threshold) because this provides additional information about the likelihood that we will find effects if the program is indeed effective. If power to detect effects is less than 75 percent even for a very large effect, then the evaluation is extremely poorly powered for that outcome.

[d] Percentage difference is calculated as the regression-adjusted difference between the treatment and comparison group, divided by the adjusted comparison group mean.

[e] *p*-values test the null hypothesis that the regression-adjusted difference-in-differences estimate is less than or equal to zero for outcomes in the quality-of-care processes domain, or greater than or equal to zero in all other domains (a one-sided test). Because it is a one-sided test, as the difference-in-differences estimate approaches infinity in an unfavorable direction (negative for process of care measures and positive for all other measures), the *p*-value approaches 1, whereas it would approach 0 in a two-sided test. We adjusted the *p*-values for the multiple (three) comparisons made within the quality-of-care processes domain, and (separately) for the two comparisons made within the quality-of-care outcomes domain, and for the two comparisons made within the service use domain.

[f] We estimated impacts as the average across intervention quarters 5 through 8 for all outcomes but two: namely, the quality-of-care process measures for diabetes and ischemic vascular disease. For those two measures, we calculated outcomes instead over year-long periods (rather than quarters). The impact estimates apply to the same time period—that is, the year that corresponds to intervention quarters 5 through 8—but the estimate is not an average of quarterly estimates.

ED = emergency department; FFS = fee-for-service; HCIA = Health Care Innovation Award.

n.a. = not applicable.

The statistical power to detect effects the size of the substantive threshold was poor for both ACSC admissions (28.1 percent) and 30-day unplanned readmissions (21.3 percent). Power was also poor (25.5 percent) for the combined effect in the domain.

Service use. The treatment group's admission rate was 2.6 percent higher, and the outpatient ED visit rate was 2.6 percent lower, than the estimated counterfactuals. Neither of these differences was statistically significant or substantively large. After combining results across the two outcomes in this domain, the outcomes for the treatment group were almost identical (0.2 percent lower) to the estimated counterfactual. Power to detect effects that were the size of the substantive thresholds was marginal for the admissions and outpatient ED visit measures individually (60.1 and 69.3, respectively) but good (75.2 percent) for the two outcomes combined.

Spending. The treatment group averaged $1,014 per beneficiary per month in Part A and B spending during the fifth through eighth intervention quarters, a value 0.9 percent (or $9) higher than the estimated counterfactual. This difference was much smaller than the substantive threshold of 4 percent. Statistical power to detect an effect the size of the substantive threshold was marginal (60.5 percent).

Aggregate estimates for CMMI's core measures. The estimates presented for the CMMI core outcomes—that is, for 30-day unplanned readmissions, all-cause inpatient admissions, the outpatient ED visit rate, and Medicare Part A and B spending—have so far been expressed per 1,000 beneficiaries per quarter (or, for spending, per beneficiary per month). Table V.6 translates these rates or per-beneficiary-month estimates into estimates of aggregate impacts during the year-long primary test period presented in this report. We calculated these aggregate impacts by multiplying the point estimates by the average number of Medicare beneficiaries in the treatment group and by the number of quarters or months during the primary test year. Although the point estimates are small for most of these measures, the aggregate estimates are fairly large because they are scaled to the full Medicare population of roughly 35,000 beneficiaries and to the full year of the primary test period. For example, the results in Table V.5 show the intervention was associated with an increase in Medicare Part A and B spending of $9 per beneficiary per month, or 0.9 percent relative to the estimated counterfactual. However, across roughly 35,000 beneficiaries and 12 months, this small spending increase per beneficiary per month translates into an aggregate cost of the program of roughly $4.1 million. These large point estimates should be interpreted with caution because the estimates are not statistically significant for any of the outcomes (the *p*-values for these aggregate estimates are the same as they are for the main results shown in Table V.5).

Table V.6. Results for primary tests for CMMI's core outcomes expressed as aggregate effects for all Medicare FFS beneficiaries in the treatment group

Outcome (units)	Aggregate impact estimate during the primary test year (August 1, 2014, through July 31, 2015)	*p*-value
30-day unplanned readmissions (#)	+193	0.90
All-cause inpatient admissions (#)	+282	0.67
Outpatient ED visits (#)	-396	0.23
Medicare Part A and B spending ($)	+$4,135,000	0.64

Sources: Authors' calculation, based on analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at the Centers for Medicare & Medicaid Services.

Note: To estimate the aggregate impact during the primary test year (intervention quarters 5 through 8) we (1) multiplied the per beneficiary per quarter (or month) estimate from Table V.5 by the average number of Medicare FFS beneficiaries in the treatment group during the four primary test quarters, then (2) scaled the estimate to a year by multiplying the resulting product by 4 (or 12). The *p*-values are taken from Table V.5 and are therefore one-sided (testing that the program improved outcomes) and adjusted for multiple comparisons conducted within each outcome domain.

CMMI = Center for Medicare & Medicaid Innvation; ED = emergency department; FFS = fee-for-service.

4. Results for secondary tests

Estimates during the first intervention year (August 1, 2013, to July 31, 2014). As shown in Table V.7, the differences in hospitalizations and spending for the treatment group and its estimated counterfactual were small (2.5 percent or less) and not statistically significant during the two secondary test periods: the first six months of the intervention (I1 and I2) and the next six months (I3 and I4). These results help support the credibility of the comparison group because we do not see large differences (favorable or unfavorable) during the first year of panel participation, a period during which we and CareFirst did not expect to see large program effects. This increased confidence in the comparison group, in turn, gives us greater confidence in the primary test results and, eventually, the conclusions of the impact evaluation.

Estimates for high-risk beneficiaries. The primary test results suggest unfavorable impacts in the quality-of-care outcomes domain, so we reestimated impacts for the two outcomes in this domain among high-risk beneficiaries only. Table V.7 shows that these impact estimates are not substantively unfavorable for the high-risk beneficiaries, even though the estimate for 30-day unplanned readmissions, in particular, was substantively large and unfavorable among the full population. In fact, the point estimate for 30-day readmissions is slightly favorable for the high-risk group. We might expect to see findings such as these if the intervention diverted attention from lower- to higher-risk beneficiaries.

Estimates limiting the sample to prevent sample addition. The secondary test results limited to those beneficiaries attributed at the start of the baseline or intervention period are consistent with the primary test results. They show no statistically significant or substantively large difference between the treatment group and its estimated counterfactual for inpatient admissions or Medicare spending (the only two outcomes assessed in these secondary tests; Section V.A.7).

Table V.7. Results of secondary tests for CareFirst

Secondary test definition				Results			
Domain	Outcome (units)	Time period for impacts	Population	Treatment group mean	Regression-adjusted difference between treatment and the estimated counterfactual (standard error)	Percentage difference[a]	*p*-value[b]
Estimates during the first intervention year (August 1, 2013 – July 31, 2014)							
Service use	All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	Intervention quarters 1, 2	All Medicare FFS beneficiaries assigned to treatment panels	75.8	1.0 (2.7)	1.4%	0.65
	All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	Intervention quarters 3, 4	All Medicare FFS beneficiaries assigned to treatment panels	74.9	1.8 (2.7)	2.5%	0.75
Spending	Medicare Part A and B spending ($/beneficiary/month)	Intervention quarters 1, 2	All Medicare FFS beneficiaries assigned to treatment panels	$984	$9 ($27)	1.0%	0.63
	Medicare Part A and B spending ($/beneficiary/month)	Intervention quarters 3, 4	All Medicare FFS beneficiaries assigned to treatment panels	$991	$17 ($29)	1.7%	0.72
Estimates for high-risk beneficiaries for the one domain (quality-of-care outcomes) in which the primary tests indicate substantively large effects							
Quality-of-care outcomes	30-day unplanned readmissions (#/1,000 beneficiaries/quarter)	Intervention quarters 5-8	High-risk Medicare FFS beneficiaries assigned to treatment panels	22.4	-0.04 (2.5)	-1.8%	0.49
	Inpatient admissions for ambulatory care-sensitive conditions (#/1,000 beneficiaries/quarter)	Intervention quarters 5-8	High-risk Medicare FFS beneficiaries assigned to treatment panels	28.8	0.8 (2.3)	2.9%	0.63
	Combined	Intervention quarters 5-8	High-risk Medicare FFS beneficiaries assigned to treatment panels	n.a.	n.a.	1.3%	0.56
Estimates limiting the sample to prevent sample addition after the first baseline or intervention quarter							
Service use	All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	Intervention quarters 5–8	Medicare FFS beneficiaries assigned to treatment panels in the first baseline or first intervention quarter	75.7	1.4 (2.5)	1.8%	0.71

Table V.7 *(continued)*

Secondary test definition				Results			
Domain	Outcome (units)	Time period for impacts	Population	Treatment group mean	Regression-adjusted difference between treatment and the estimated counterfactual (standard error)	Percentage difference[a]	*p*-value[b]
Spending	Medicare Part A and B spending ($/beneficiary/month)	Intervention quarters 5–8	Medicare FFS beneficiaries assigned to treatment panels in the first baseline or first intervention quarter	$1,034	$5 ($29)	0.5%	0.57

Sources: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at the Centers for Medicare & Medicaid Services.

Notes: The results for each outcome are based on a difference-in-differences regression model, as described in the text. Estimates are calculated for Medicare beneficiaries who are observable in the relevant time period: that is, beneficiaries who are enrolled in Medicare FFS (Part A and B), are alive, and have Medicare as their primary payer. We defined high-risk beneficiaries as those with a Hierarchical Condition Category score in the top third among all treatment group members at the beginning of the baseline period (for outcomes in the baseline period) or intervention period (for outcomes in the intervention period).

[a] Percentage difference is calculated as the regression-adjusted difference between the treatment and comparison group, divided by the adjusted comparison group mean.

[b] The *p*-values from the secondary test results were not adjusted for multiple comparisons within or across domains.

FFS = fee-for-service.

n.a. = not applicable.

5. Consistency of impact estimates with implementation findings

The impact estimates in the primary tests are plausible given the implementation findings. The primary test results did not show any favorable effects during the second year of the program (that is, the first four quarters of the six-quarter primary test period) that were statistically significant or substantively important. The implementation evidence shows the program was active during the year. For example, as described in Section III.B.2, LCCs provided care coordination services to 1,300 to 1,800 high-risk Medicare beneficiaries during this period. However, even with a well-implemented intervention, it is possible that the program was not able to change beneficiaries' or providers' behaviors in ways that would affect impact outcomes during the primary test period covered in this report.

The substantively large unfavorable impact estimate for quality-of-care outcomes is surprising, but not implausible. By showing the impact estimates were not substantively unfavorable for the high-risk beneficiaries, the secondary test results imply that unfavorable impact estimates for lower-risk beneficiaries drove the overall unfavorable results. It is possible the intervention could have had unfavorable impacts for the lower-risk patients if participation in the HCIA intervention (1) diverted PCPs' attention away from lower-risk to higher-risk beneficiaries in ways that meaningfully detracted from clinical care for lower-risk beneficiaries and/or (2) prevented the panels from participating in other quality improvement activities that could reduce readmission rates for all beneficiaries.

6. Preliminary conclusions about program impacts, by domain

Based on all evidence currently available, we have drawn the following preliminary conclusions about program impacts during the first 12 of the planned 18 months of the primary test period. Table V.8 summarizes these preliminary conclusions and their support.

- **The program did not have a substantively large impact on quality-of-care processes or service use.** For all outcomes in these domains, the primary test results were neither substantively large nor statistically significant. The statistical power to detect effects in these domains was good (more than 75 percent). Specifically, in the quality-of-care processes domain, power was good for each of the measures in the domain. In the service use domain, power was good for the combined impact estimate across two outcomes in the domain. The secondary test results support these primary test results by (1) showing no impacts in the first program year (when none were expected) and (2) demonstrating that differential sample addition over time between the treatment and comparison groups did not drive results. These conclusions are also consistent with implementation findings because, although the program was implemented reasonably well, it is plausible the program did not have its intended effects.
- **The program had a substantively large *unfavorable* effect on quality-of-care outcomes.** The primary test results showed a substantively large unfavorable estimate for quality-of-care outcomes, driven by a large unfavorable estimate for 30-day unplanned readmissions, in particular. However, the standard errors were large for both this estimate and the estimated combined effect in the domain. Therefore, we have low confidence in the conclusion of substantively unfavorable impacts. It is possible that the large observed point

estimates were due to chance rather than true unfavorable impacts. However, there is a potentially plausible explanation for how the program could have worsened quality-of-care outcomes. The secondary test results suggest the program had unfavorable effects for the lower-risk beneficiaries (Section V.D.4). Though we have no direct evidence to suggest this is happening, it is possible these results could occur (1) if PCPs diverted attention from lower- to higher-risk patients (for example, to focus time on care coordination for high-risk patients); or (2) if participation in the intervention prevented treatment panels from undertaking other quality initiatives to reduce readmission rates for their full Medicare population that comparison panels might have. Table V.4 shows the comparison group did experience reductions in readmission rates among Medicare beneficiaries, which were not matched in the treatment group.

- **The program had an indeterminate effect on Medicare spending.** The primary test results were neither substantively large nor statistically significant. However, the statistical power was marginal (60.5 percent) to detect effects the size of the substantive threshold. As a result, null findings from the primary test in this domain could be due to (1) the program truly not having a substantively large effect or (2) the program having a substantively large effect but our tests failing to detect it. The fact that we observed no declines in service use (which the awardee anticipated would drive reductions in spending)—and that primary tests for service use were well powered—suggests that lack of effects on spending is the more likely explanation.

Table V.8. Preliminary conclusions about the impacts of CareFirst's HCIA program on patients' outcomes, by domain

		Evidence supporting conclusion		
Domain	Preliminary conclusion	Primary test result(s) that supported conclusion	Primary test result(s) plausible given secondary tests?	Primary test result(s) plausible given implementation evidence?
Quality-of-care processes	No substantively large effect	• No substantively large or statistically significant effects; well powered to detect effects on all outcomes in the domain	Yes	Yes
Quality-of-care outcomes	Substantively large unfavorable effect	• Combined effect across the two measures in domain was unfavorable and larger than the substantive threshold	Yes	Yes
Service use	No substantively large effect	• No substantively large or statistically significant effects; well powered to detect a substantively large effect on the combined outcome in the domain	Yes	Yes
Spending	Indeterminate effect	• No statistically significant or substantively important effect; power was marginal to detect an effect on the single outcome in the domain	Yes	Yes

Sources: Tables V.5 and V.7

VI. DISCUSSION AND CONCLUSIONS

CareFirst used its $20 million in HCIA funds to extend a PCMH program designed for its commercial members to Medicare FFS beneficiaries in Maryland. Building on infrastructure developed for the commercial program, the intervention had three main components: (1) care coordination for high-risk Medicare beneficiaries; (2) financial incentives to PCPs to reduce total Medicare spending for their patients while meeting quality targets; and (3) technical assistance to panels to identify cost-saving opportunities, primarily through referring patients to more cost-effective providers and settings. Through these three intervention components, CareFirst aimed to improve quality of care for the participating panels' Medicare FFS beneficiaries; reduce the need for expensive hospitalizations and ED visits, particularly among high-risk beneficiaries; and reduce total Medicare spending.

The results from our impact evaluation suggest CareFirst did not meet these goals during the original three-year award period. Outcomes for Medicare FFS patients served by the 14 treatment panels were not statistically or substantively better than those for Medicare patients served by 42 matched comparison panels in any of the four evaluation domains: quality-of-care processes, quality-of-care outcomes, service use, and spending. The evaluation was well powered to detect substantively large impacts on quality-of-care processes and service use, but not quality-of-care outcomes or spending.

The lack of favorable effects does not appear to be due to major problems implementing the intervention as planned. After an initial 13-month implementation delay due to delays in receiving complete Medicare FFS claims data, CareFirst delivered a complex intervention consistent with its core design. CareFirst's experience and infrastructure from its commercial PCMH program enabled it to quickly ramp up the intervention after the initial delay. Several measures capture the generally successful implementation:

- CareFirst hired 44 LCCs, more than originally anticipated, to assist PCPs in developing and implementing care plans, and paid PCPs for their time coordinating the care of high-risk patients.
- The 14 participating panels provided care coordination services for 3,276 patients, exceeding original targets.
- In each performance year completed by the end of the original award period (2013 and 2014), CareFirst calculated and paid OIAs for panels that reduced spending relative to targets. (CareFirst could have concluded that a panel generated savings, even though our evaluation found no overall impacts on spending, due to differences in methods. CareFirst calculated savings by comparing actual spending for a panel's Medicare patients with a target level of spending that assumed spending would grow by 2.5 percent per year absent the intervention. Our evaluation used a difference-in-differences framework with a matched comparison group to estimate impacts. We found that actual spending growth in both the treatment and comparison groups was much less than 2.5 percent per year.)

- Program consultants met with panels more often than initially planned (at least monthly) to analyze claims data and identify opportunities to lower total spending while maintaining quality of care for their panels' Medicare patients.

Further, the lack of effects appears not to be due to an inability to engage PCPs as planned. PCPs are central to the awardee's theory of action because they, jointly with LCCs, had to provide care coordination services to high-risk beneficiaries. A large majority (90 percent) of the 149 PCPs in panels participated in care coordination, as gauged by developing at least one care plan (most PCPs developed 10 or more care plans). PCPs are also the clinicians with medical authority to change referrals patterns. Further, the primary care clinician survey results indicate that most PCPs thought the program improved the quality, safety, and timeliness of the care they provide. However, we have no evidence to assess whether the program changed PCPs' referral patterns, another important element of the awardee's theory of action.

These findings suggest the lack of measured effects might be due to one of three factors. First, although the program was generally implemented as planned (after the 13-month delay), a few key implementation barriers might have limited the effectiveness of care coordination services. The process for identifying high-risk patients for care coordination might not have consistently identified those who could benefit most from care plans. Although PCPs and LCCs identified those who were at high risk, often as indicated by high illness burden scores, those patients might not necessarily have been clinically unstable (rather, their high burden scores might have reflected a recent acute event not tied to a chronic illness). Further, LCCs might not have been able to sufficiently adapt the care planning process from the commercial program to be successful for Medicare patients. As several respondents noted in interviews, it was difficult to adjust the care coordination process—originally designed for commercial patients—to the Medicare population due to the generally higher complexity of patients' needs.

Second, there might have been limitations in the core design of the intervention itself. Our evaluation was not designed to identify specific limitations in program design that could account for lack of effects. However, a quantitative analysis of CareFirst's data suggests program impacts for those receiving care coordination would have to be very large—perhaps unrealistically large—to drive the targeted reduction in overall spending. Specifically, based on the percentage of the panels' Medicare patients who enrolled in care coordination (8 percent) before or during the primary test period and their total spending relative to the average beneficiary's spending in the treatment group (about 2:1), we estimate the program would have had to reduce spending for those receiving care coordination services by 25 percent to achieve the intended full-panel reduction in spending of 4 percent. Such large reduction could be difficult to achieve, particularly given the challenges noted previously in systematically identifying those who could benefit most from care coordination services and in adapting care coordination strategies from the commercial to the Medicare population.

Finally, it is possible that impacts take time to accrue and will grow larger when the final six months of program operations are included in the impact evaluation. The cumulative number of Medicare beneficiaries who received care coordination services continued to grow in the final six months (data not shown), so it is possible that impacts will be largest during this period.

However, the impacts will have to be very large during these final six months to generate favorable results during the full 18-month primary test period.

The results presented in this study also highlight the importance of assessing the likelihood of, and examining evidence for, possible unintended consequences of CareFirst's HCIA-funded intervention. The results suggest the program might have worsened quality-of-care outcomes—particularly 30-day unplanned readmission rates—for lower-risk patients (although this seemingly unfavorable impact estimate could also be due to chance events, given low statistical power to detect true impacts in the domain). An unfavorable impact on quality-of-care outcomes could have happened if the program diverted important clinical attention away from lower-risk patients to higher-risk ones, diverted panels' attention away from broad-based quality improvement efforts that could improve outcomes for the panels' full Medicare population, or both.

This study estimated the marginal effect of extending the commercial PCMH program to Medicare beneficiaries because that is the most policy-relevant analysis for CMS. CMS might decide whether to join CareFirst in its existing PCMH initiative on an ongoing basis and/or in other regions. Estimates of the the likely marginal impact of joining this effort on Medicare beneficiaries can help inform CMS decisions. However, it is possible that CareFirst's commercial program alone has some positive spillover effects for Medicare beneficiaries. For example, if the PCPs in a panel—responding to incentives and technical assistance in the commercial program—change their referral patterns for *all* of their patients (not only their commercial patients), this could reduce costs of specialty care for Medicare patients. The impact estimates in this report would not capture such positive spillover because the estimates compare outcomes for Medicare FFS beneficiaries served by the 14 treatment panels to comparison panels already participating in the commercial program. We anticipate positive spillover, if any, to be minimal because the primary intervention in the commercial program—as in the Medicare expansion of the program—is care coordination for high-risk patients. This care coordination is likely to benefit only those who actually receive the services.

The next step for the evaluation is to add the final six months of program operations to the study period, completing the evaluation. CareFirst received a no-cost extension to continue its HCIA-funded intervention beyond the initial award period (which ended June 2015) through December 2015. As a result, we will update the implementation metrics in this report with the number of care plans created and number of beneficiaries CareFirst attributed through December 2015. We will also (1) generate claims-based outcomes to cover the final six months of the primary test period (August 1, 2015, to January 31, 2016); (2) conduct the final primary tests incorporating these outcomes; and (3) update our conclusions if necessary. We will report final evaluation results in a future addendum to this report.

This page has been left blank for double-sided copying.

REFERENCES

Boards of Trustees, Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds. "2013 Annual Report of the Boards of Trustees of the Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds." Table V.D1. Washington, DC: Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds, 2013. Available at http://downloads.cms.gov/files/TR2013.pdf. Accessed August 13, 2014.

Centers for Medicare & Medicaid Services. "CSV Flat Files—Revised: Readmissions Complications and Deaths—National.csv." Baltimore, MD: CMS, 2014. Available at https://data.medicare.gov/data/hospital-compare. Accessed August 14, 2014.

Chronic Conditions Data Warehouse. "Table A.1.a. Medicare Beneficiary Counts for 2005 – 2014." Baltimore, MD: Centers for Medicare & Medicaid Services. Available at https://www.ccwdata.org/web/guest/medicare-tables-reports. Accessed June 29, 2016.

Chronic Conditions Data Warehouse. "Table B.2.a Medicare Beneficiary Prevalence for Chronic Conditions for 2005 Through 2014." Baltimore, MD: Centers for Medicare & Medicaid Services. Available at https://www.ccwdata.org/web/guest/medicare-tables-reports. Accessed June 29, 2016.

Congressional Budget Office. "Dual-Eligible Beneficiaries of Medicare and Medicaid: Characteristics, Health Care Spending, and Evolving Policies." Washington DC: 2013.

Geonnotti, Kristin, Greg Peterson, Lauren Hula, Boyd Gilman, Catherine DesRoches, Sandi Nelson, Laura Blue, Keith Kranker, Kate Stewart, Frank Yoon, and Lorenzo Moreno. "Findings CareFirst BlueCross BlueShield." In Moreno, Lorenzo, Boyd Gilman, Greg Peterson, Catherine DesRoches, Sheila Hoag, Linda Barterian, Laura Blue, Katherine Bradley, Emily Ehrlich, Kristin Geonnotti, Lauren Hula, Keith Kranker, Rumin Sarwar, KeriAnn Wells, Joseph Zickafoose, Sandi Nelson, Kate Stewart, and Frank Yoon. "Evaluation of Health Care Innovation Awards (HCIA): Primary Care Redesign Programs." Second annual report to CMS. Volume II: Individual Program Summaries. Princeton, NJ: Mathematica Policy Research, December 11, 2015.

Gerhardt, Geoffrey, Alshadye Yemane, Keri Apostle, Allison Oelschlaeger, Eric Rollins, and Niall Brennan. "Evaluating Whether Changes in Utilization of Hospital Outpatient Services Contributed to Lower Medicare Readmission Rate." *Medicare & Medicaid Research Review,* vol. 4, no. 1, 2014, pp. E1–E13.

Gilman, Boyd, Sheila Hoag, Lorenzo Moreno, Greg Peterson, Linda Barterian, Laura Blue, Kristin Geonnotti, Tricia Higgins, Mynti Hossain, Lauren Hula, Rosalind Keith, Jennifer Lyons, Brenda Natzke, Brenna Rabel, Rumin Sarwar, Rachel Shapiro, Victoria Peebles, Cara Stepanczuk, KeriAnn Wells, and Joseph Zickafoose. "Evaluation of the Health Care Innovation Awards (HCIA): Primary Care Redesign Programs. First Annual Report, Volumes I and II." Princeton, NJ: Mathematica Policy Research, November 14, 2014.

Health Indicators Warehouse. "Average Age of Medicare Beneficiaries (mean)." Hyattsville, MD: National Center for Health Statistics, HIW, 2014a. Available at http://www.healthindicators.gov/Indicators/Average-age-of-Medicare-beneficiaries-mean_308/Profile/ClassicData. Accessed November 19, 2014.

Health Indicators Warehouse. "Hospital Inpatient Medicare Admissions (per 1,000 beneficiaries)." Hyattsville, MD: National Center for Health Statistics, HIW, 2014b. Available at http://www.healthindicators.gov/Indicators/Hospital-inpatient-Medicare-admissions-per-1000-beneficiaries_2001/Profile/ClassicData. Accessed August 13, 2014.

Institute of Education Sciences. "What Works Clearinghouse: Procedures and Standards Handbook, Version 3.0." Washington, DC: U.S. Department of Education, IES, 2014. Available at http://ies.ed.gov/ncee/wwc/DocumentSum.aspx?sid=19. Accessed September 15, 2014.

Peikes, Deborah, Stacy Dale, Eric Lundquist, Janice Genevro, and David Myers. "Building the Evidence Base for the Medical Home: What Sample and Sample Size Do Studies Need? White Paper." AHRQ Publication No.11-0100-EF. Rockville, MD: Agency for Healthcare Research and Quality, October 2011.

Rosenthal, MB, S. Alidina, M. Friedberg, S. Singer, D. Eastman, Z. Li, E. Schneider. A Difference-in-Differences Analysis of Changes in Quality, Utilization, and Cost Following the Colorado Multi-Payer Patient-Centered Medical Home Pilot. *Journal of General Internal Medicine,* 2016, vol. 31, no. 3, pp. 289-296, March 2016.

Stuart, Elizabeth A. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science,* vol. 25, no. 1, 2010, pp. 1–21.

Truven Health Analytics. "AHRQ Quality Indicators, Prevention Quality Indicators v5.0 Benchmark Data Tables." Prepared for the Agency for Healthcare Research and Quality, U.S. Department of Health and Human Services. Santa Barbara, CA: Truven Health Analytics, March 2015. Available at http://www.qualityindicators.ahrq.gov/Downloads/Modules/PQI/V50/Version_50_Benchmark_Tables_PQI.pdf. Accessed August 18, 2015.

U.S. Census Bureau. "2008–2012 American Community Survey, Median Household Income." Washington, DC: U.S. Census Bureau, 2012.

CHAPTER 3

DENVER HEALTH AND HOSPITAL AUTHORITY

Laura Blue, Lauren Hula, Tricia Higgins, Greg Peterson, Boyd Gilman, Keith Kranker, Kate Stewart, Sheila Hoag, and Lorenzo Moreno

This page has been left blank for double-sided copying.

DENVER HEALTH AND HOSPITAL AUTHORITY

CHAPTER SUMMARY

Introduction. Denver Health and Hospital Authority (Denver Health) used its $19.8 million Health Care Innovation Award (HCIA) to develop 21st Century Care, a program to transform Denver Health's primary care delivery system to more effectively meet its patients' medical, behavioral, and social needs. Under 21st Century Care, Denver Health developed risk-stratification methodology to sort its patient population into risk tiers, and then allocated resources—including newly hired clinic support staff, such as patient navigators, and newly created high-risk clinics—based on need. The intervention targeted all Denver Health primary care users, members of its managed care plan, and frequent users of the Denver Health hospital and emergency department (ED)—a combined population that exceeded 100,000 people in any given month. With 21st Century Care, Denver Health hoped to (1) improve patients' health outcomes by 5.0 percent, based on an internal composite quality metric; (2) increase patients' satisfaction with between-visit care by 5.0 percent, without decreasing visit-based care satisfaction; and (3) decrease total cost of care by 2.5 percent, relative to an inflation-adjusted baseline.

Objectives. This report has three main objectives: (1) to describe the design and implementation of Denver Health's HCIA-funded intervention, including the role of primary care providers in the intervention and the extent to which anticipated changes in providers' behavior occurred; (2) to assess impacts of the intervention on Medicare fee-for-service (FFS) outcomes and Medicare Part A and B spending during the three years of the award funding; and (3) to use both implementation and impact findings to identify possible explanations for the observed impacts.

Methods. We reviewed Denver Health's program documents and self-monitoring metrics, conducted site visits and interviews with Denver Health leadership and program staff, and surveyed participating clinicians. To estimate program impacts, we assessed outcomes for roughly 3,700 Medicare FFS patients served by Denver Health's eight federally qualified health centers (FQHCs)—a small subset of the intervention's target population—and compared these with outcomes for roughly 7,000 Medicare patients served by 15 (unmatched) comparison FQHCs located in urban regions of Colorado, adjusting for measured differences in patient and FQHC characteristics (including patients' outcomes) between the two groups during an 18-month baseline period.

Program design and implementation. The intervention had four components: (1) stratify patients based on risk to more efficiently allocate resources, (2) leverage health information technology (IT) to provide between-visit support, (3) redesign Denver Health's primary care delivery teams, and (4) create high-risk clinics to provide individualized care to patients with complex care needs. Denver Health implemented 21st Century Care largely as planned and well enough to be a reasonable test of the program's core design.

Clinicians' perceptions of intervention effects on the care they provide. In surveys we administered, clinicians reported that they perceived the program as effective and most believed

the HCIA-funded initiative improved the quality, patient-centeredness, and timeliness of the care they provided to patients. However, we have little evidence to assess whether clinicians working at Denver Health before the intervention changed the way they delivered clinical services, as our survey did not ask detailed questions about changes to clinicians' daily activities.

Impacts on patients' outcomes. The impact estimates indicated largely indeterminate effects of the intervention on Medicare FFS beneficiaries. We found no evidence of statistically significant or substantively large differences between the treatment and comparison groups in three of four evaluation domains: quality-of-care processes, quality-of-care outcomes, and spending. However, for two of these domains—quality-of-care outcomes and spending—we had poor statistical power to detect effects. This means we cannot be sure whether the intervention truly had no effects in these domains, or whether it did have effects and our evaluation failed to detect them. In the fourth evaluation domain, service use, we estimated a substantively large *unfavorable* effect, driven by an estimated increase of 14.2 percent (relative to the comparison group) in the outpatient ED visit rate. This increase does not seem plausible given the implementation evidence, and could be due to statistical noise.

Conclusion. The evaluation yielded largely indeterminate evidence about the impact of Denver Health's HCIA-funded intervention on Medicare FFS beneficiaries who used Denver Health FQHCs. However, this group comprised only a small proportion (less than 5 percent) of the 21st Century Care target population. Denver Health did implement its program on schedule and as planned. We therefore see three likely explanations for the indeterminate impact results. First, it is possible the program was effective for Medicare beneficiaries at Denver Health's FQHCs but that we failed to detect program impacts on quality-of-care outcomes or spending due to poor statistical power. Second, it is possible 21st Century Care had effects for some of its target population, but not for Medicare FFS primary care users in particular. This is possible because the treatment group for this evaluation excluded many people that might have experienced large program impacts, including those without Medicare coverage or without a Denver Health primary care provider at the start of the HCIA period. Finally, it is possible we did not observe substantively large or statistically significant favorable effects because the program truly did not have these effects. This would mean the program, although implemented well, failed to reduce patients' needs for acute care and, in turn, reduce spending.

Summary of intervention and impact results for Denver Health

Intervention description			
Awardee description		Integrated safety-net health system; largest provider to Medicaid and uninsured patients in Colorado	
Award amount ($ millions)		$19.8 million	
Award extended beyond June 2015?		No	
Location		Denver, Colorado (urban)	
Target population		All patients (about 250,000) meeting one of the following criteria: • Served by Denver Health's 8 FQHCs • In Denver Health's managed care plan • Used Denver Health's hospital or ED at least 3 times in one year, or at least twice if the patient also had a serious mental health diagnosis	
Interventions		Stratified patients into 4 risk tiers and, within those tiers, into clinically similar groups, to triage to other intervention services • Text message reminders about appointments • Enhanced primary care teams in 8 FQHCs, incorporating 23 HCIA-funded patient navigators and 3 clinical pharmacists • High-risk clinics that offered longer and more comprehensive appointments than typically covered by insurance	
Metrics of intervention delivered		• 79,400 contacts with patient navigators • 19,000 contacts with clinical pharmacists • Text message reminders to 28,000 patients, with an average of 8 messages per person • High-risk clinics at capacity at intervention end	
Impact evaluation methods			
Core design		Difference-in-differences model with comparison group (unmatched)[a]	
Treatment group	Definition	Medicare FFS beneficiaries attributed to one of 8 FQHCs by the intervention start	
	# of beneficiaries during primary test period[b]	2,317 to 3,133	
Comparison group definition		Medicare FFS beneficiaries attributed to 15 comparison FQHCs by the start of the intervention	
Impact results: Quality-of-care processes domain			
Ambulatory care visit within 14 days of discharge (% of beneficiaries/quarter)		Comparison mean[c]	50.5%
		Impact estimate (% difference)	+0.6 pp (+1.1%)
Impact conclusion[d]		No substantively large effect	
Impact results: Quality-of-care outcomes domain			
30-day unplanned hospital readmissions (#/1,000 beneficiaries/quarter)		Comparison mean[c]	15.1
		Impact estimate (% difference)	+0.9 (+6.1%)
Inpatient admissions for ACSC conditions (#/1,000 beneficiaries/quarter)		Comparison mean[c]	9.3
		Impact estimate (% difference)	+0.3 (+3.3%)
Combined impact estimate[e]		+4.7%	
Impact conclusion[d]		Indeterminate effect	
Impact results: Spending domain			
Medicare Part A and B spending ($/beneficiary/month)		Comparison mean[c]	$948
		Impact estimate (% difference)	+$8 (+0.9%)
Combined impact estimate[e,f]		+0.4%[d,e]	
Impact conclusion[d]		Indeterminate effect	

Note: See the Denver Health chapter for details on the intervention, impact methods, and impact results. As explained in the chapter, we did not draw impact conclusions in the service use domain.

[a] The comparison group was unmatched because statistical matching did not meaningfully improve balance on prespecified matching variables relative to the full pool of potential comparison practices. We relied on the difference-in-differences model to account for any differences in outcomes that stemmed from persistent (time-invariant) differences between the treatment and comparison practices.

[b] For some outcome measures the sample is limited to a relevant subset of beneficiaries.

[c] The comparison mean is the estimate of the outcome the treatment group beneficiaries would have had if they had not received the intervention. It is equal to the mean for the treatment group over the intervention quarters (in the primary test period) minus the impact estimate.

Summary of intervention and impact results for Denver Health *(continued)*

[d] We drew conclusions at the domain level based on the results of prespecified primary tests, secondary tests (robustness checks), and consistency with implementation evidence. For each domain, we could draw one of five conclusions: (1) Statistically significant favorable effect (the highest level of evidence), (2) Substantively important (but not statistically significant) favorable effect, (3) Substantively important (but not statistically significant) unfavorable effect, (4), No substantively large effect, and (5) Indeterminate effect. Section V.A.8 of this report describes the decision rules we used to reach each of these possible conclusions.

[e] The combined estimate is the average across all the individual estimates in each domain, in which the impact estimate for each individual outcome is expressed as percentage change relative to the comparison group.

[f] The combined measure is the average of point estimates from two overlapping time periods specified in the primary tests (that is, the 5th through 11th intervention quarters and the 8th through 11th intervention quarters) as described in Denver Health chapter.

*Significantly different from zero at the .10 level, one-tailed test.

**Significantly different from zero at the .05 level, one-tailed test.

***Significantly different from zero at the .01 level, one-tailed test.

ACSC = ambulatory care-sensitive condition; FFS = fee-for-service; FQHC = federally qualified health center; HCIA = Health Care Innovation Award; pp = percentage point.

I. INTRODUCTION

This report presents findings from the evaluation of the Denver Health and Hospital Authority (Denver Health) Health Care Innovation Award (HCIA), with a focus on program impacts on patients' outcomes during the three-year award period (June 2012 through June 2015). Section II provides an overview of Denver Health's HCIA-funded intervention and the design of the impact evaluation. Section III describes the design and implementation of the intervention, including how the program could be expected to affect evaluation outcomes through changes in patients' and providers' behavior. In Section IV, we assess providers' perceptions of program effects on their patients' care, and whether there is evidence to assess changes in clinicians' behavior during the award period. Section V describes our methods for, and results and conclusions from, estimating program impacts on patients' outcomes in four domains: quality-of-care processes, quality-of-care outcomes, service use, and spending. Section VI draws conclusions by synthesizing the impact and implementation findings from the evaluation.

II. OVERVIEW OF DENVER HEALTH'S HCIA-FUNDED INTERVENTION AND THE IMPACT EVALUATION

A. Denver Health's HCIA-funded intervention

Denver Health received a three-year, $19.8 million award to implement 21st Century Care, a program designed to transform the primary care delivery system to more effectively meet its patients' medical, behavioral, and social needs (Table II.1). Denver Health is an integrated safety-net system in Denver, Colorado. It is the largest provider of health care to Medicaid beneficiaries and uninsured patients in the state. Its facilities include eight federally qualified health centers (FQHCs), as well as urgent care facilities, an acute care facility with inpatient and emergency department (ED) services, and a managed care plan.

Denver Health began implementing 21st Century Care in late October 2012 in its eight FQHCs. The intervention included categorizing the target population into risk tiers, providing low-cost text messaging appointment and preventive care reminders to those in the lowest-risk tiers and more involved care coordination and care transitions services to those in higher-risk tiers. As part of the HCIA-funded program, Denver Health also created three new clinics to serve patients with the greatest health care needs.

Through 21st Century Care, Denver Health hoped to (1) improve patients' health outcomes by 5.0 percent, based on an internal composite quality metric; (2) increase patients' satisfaction with between-visit care by 5.0 percent, without decreasing visit-based care satisfaction; and (3) decrease total cost of care by 2.5 percent, which reflects reductions of 0.7 percent in the first year of the program, 3.0 percent in the second year, and 3.4 percent in the third year on a per-person-per-year basis, relative to an inflation-adjusted baseline. Although Denver Health estimated these effects across the entire intervention population, program administrators expected to achieve the most significant cost reductions among its highest-risk patients through decreased use of expensive services, such as inpatient and ED care.

Denver Health's HCIA program ended in June 2015. The awardee received a no-cost extension to continue its reporting and self-monitoring work through September 2015 using HCIA funding, although no HCIA funds were used to fund program services after June. Denver Health continues to operate many 21st Century Care program components using internal funding.

Table II.1. Summary of Denver Health's PCR program and our evaluation for estimating its impacts on patients' outcomes

Program description	
Award amount	$19,789,999
Award start date	June 2012
Implementation date	October 29, 2012
Award end date	June 30, 2015
Awardee description	Denver Health is an integrated safety net system in Denver, Colorado, and the largest provider of health care to Medicaid beneficiaries and uninsured patients in the state.
Intervention overview	The goal of the Denver Health program—called 21st Century Care—was to transform Denver Health's primary care delivery system to more effectively meet its patients' medical, behavioral, and social needs. This was planned largely through risk-stratification of the patient population and targeted resources for those patients deemed highest risk.
Intervention components	1. **Stratify patients based on risk to more efficiently allocate resources.** Denver Health used clinical risk grouping software, augmented with in-house administrative and clinical data, to assign each patient in the target population to one of four risk-stratification tiers, with Tier 1 representing the lowest-risk patients and Tier 4 representing the highest-risk patients. 2. **Leverage health IT to provide between-visit support.** Denver Health invested in health IT to send patients text message reminders. Patients in all risk groups were eligible for services provided through this program component. 3. **Redesign Denver Health's primary care delivery teams.** Enhanced primary care delivery teams included clinical pharmacists, registered nurses, behavioral health consultants, licensed clinical social workers, and patient navigators, who joined all eight FQHCs at Denver Health. Patient navigators are nonclinical staff who help patients with appointment scheduling, transportation, and other nonclinical needs. Patients in the mid-level risk groups (Tiers 2 and 3), along with the highest-risk patients (Tier 4), were eligible to receive services from enhanced primary care delivery teams. 4. **Create high-risk clinics to provide individualized care to patients with complex care needs.** Denver Health created three high-risk clinics, each with a different care model and target population. The IOC was a primary care clinic that focused on high-risk adults with a primary physical diagnosis and multiple comorbidities. The IOC's enhanced primary care team coordinated to provide individualized care to address patients' physical and behavioral health care needs, and to connect patients to social services. The second high-risk clinic, co-located at MHCD, offered community-based case management services in addition to enhanced primary care to adult patients with severe mental health conditions and two or more hospitalizations in the previous year. The third high-risk clinic, for children with special health care needs, worked with children with multiple chronic needs. Like the IOC, Denver Health designed the CSHCN clinic to provide patients with access to the enhanced primary care team during each office visit. Only patients in the highest-risk group (Tier 4) were eligible to receive care in one of the three specialized high-risk clinics created using HCIA funds.
Target population	The target population for 21st Century Care included the following three groups: 1. All primary care patients at Denver Health (defined by the awardee as any person who had a primary care visit in the previous 18 months) 2. All patients enrolled in Denver Health's managed care plan 3. Frequent users[a] of Denver Health hospital, ED, and urgent-care services who did not fall into the previous two categories This target population included nearly 250,000 distinct patients over the course of the award period (roughly 130,000 at any given time).

Table II.1 *(continued)*

Target impacts on patients' outcomes	• 5.0 percent improvement in patients' health outcomes, as measured on an internal composite measure • 5.0 percent increase in patients' satisfaction with between-visit care • 2.5 percent decrease in the total cost of care
Workforce development	Added new staffing positions to expand the capacity of Denver Health's FQHCs and to create three new high-risk clinics for patients with complex care needs
Location	Denver, Colorado (urban area)
Impact evaluation	
Core design	Difference-in-differences with comparison group (unmatched) at the FQHC level
Treatment group	Medicare FFS beneficiaries, including those dually eligible for Medicaid, who were assigned to one of Denver Health's eight FQHCs by the start of the period (either baseline or intervention)
Comparison group	Medicare FFS beneficiaries, including those dually eligible for Medicaid, who were assigned to one of 15 (unmatched) comparison FQHCs by the start of the period (either baseline or intervention). The comparison FQHCs were drawn from urban regions of Colorado.
Intervention component(s) included in impact evaluation	All—but only for Medicare FFS beneficiaries attributed by the start of the intervention period. Some aspects of some program components do not affect Medicare beneficiaries (such as the creation of the high-risk CSHCN clinic), so, in practice, our impact estimates do not reflect these aspects.
Extent to which the treatment group reflects the awardee's target population (for the component(s) evaluated)	**Low.** We anticipate that our evaluation treatment group covers no more than 5 percent of the awardee's total target population. Denver Health is the largest provider of health care to Medicaid and uninsured patients in Colorado, but our impact evaluation covers only Medicare FFS beneficiaries. Moreover, one of Denver Health's intervention goals was to bring new patients into primary care (especially patients who were frequent users of the Denver Health hospital and ED, and not otherwise receiving primary care services), but—to limit bias from changing population composition—our impact evaluation covers only beneficiaries who received primary care at Denver Health before the intervention began. Denver Health expected program impacts to be greater than average among Medicare FFS beneficiaries.
Study outcomes, by domain	1. **Quality-of-care processes.** Receipt of a follow-up ambulatory care visit within 14 days of hospital discharge 2. **Quality-of-care outcomes.** 30-day unplanned readmissions and select inpatient admissions for ambulatory care-sensitive conditions[b] 3. **Service use.** All-cause inpatient admissions and outpatient ED visits 4. **Spending.** Medicare Part A and B spending
Intervention type, based on component(s) evaluated	Practice transformation

Source: Review of Denver Health reports, including its original application, operational plan, and 13 quarterly narrative reports to CMS.

[a] Denver Health defined frequent users as (1) people with three or more urgent care visits, ED visits, or hospital admissions (including inpatient and observational stays) in the past 12 months; or (2) people with two or more hospital admissions, along with a serious mental health diagnosis. Qualifying mental health conditions included schizophrenic disorders and select affective and personality disorders, among others.

[b] The select ambulatory care-sensitive conditions include heart failure, hypertension, angina, diabetes long-term complications, uncontrolled diabetes, lower extremity amputation, chronic obstructive pulmonary disease or asthma in older adults (ages 40 and older), perforated appendix, and dehydration. We do not include the following because Denver Health told us that, under 21st Century Care, staff were not monitoring admissions for these conditions in particular: bacterial pneumonia, urinary tract infection, and asthma among younger adults (ages 18 to 39).

CMS = Centers for Medicare & Medicaid Services; CSHCN = children with special health care needs; ED = emergency department; FFS = fee-for-service; FQHC = federally qualified health center; HCIA = Health Care Innovation Award; IOC = intensive outpatient clinic; IT = information technology; MHCD = Mental Health Center of Denver; PCR = primary care redesign.

B. Overview of impact evaluation

To estimate program impacts on patients' outcomes, we compared outcomes for Medicare fee-for-service (FFS) beneficiaries served by the 8 Denver Health FQHCs (treatment FQHCs) with outcomes for beneficiaries served by 15 other urban FQHCs in Colorado (comparison FQHCs), adjusting for differences in patient and FQHC characteristics (including patients' outcomes) between these two groups before the intervention began. The bottom panel of Table II.1 summarizes our impact evaluation design. Although Denver Health served a largely Medicaid and uninsured population with its HCIA program, due to limitations in available data we analyzed outcomes only for the Medicare FFS population (including those who were dually eligible for Medicare and Medicaid). Results might not be generalizable to the full population that Denver Health's program served.

We selected the 15 comparison FQHCs for the evaluation from the pool of all FQHCs in the 17 counties of Colorado's Front Range urban corridor, the most populous part of the state, ranging from the Wyoming border in the north to Pueblo County in the south. The comparison group is not a matched comparison group because matching did not substantially improve the similarity of treatment and comparison FQHCs relative to not matching. Instead, to select comparison FQHCs, we used the full pool of all FQHCs in the region, but eliminated those that appeared to be obviously poor matches due to characteristics not shared by any of the treatment practices. Specifically, we excluded from the comparison group those FQHCs that (1) served a narrow target population or an otherwise restricted population (for example, women's clinics or Indian Health Centers); (2) offered limited services (for example, mobile or school-based clinics); (3) did not have at least 30 attributed Medicare FFS beneficiaries in every year for which we needed data for this evaluation (2009 to 2015, inclusive); (4) were staffed largely by volunteers; (5) were associated in billing data with multiple street addresses, suggesting that more than one clinic operated under an umbrella organization's identifier for billing; or (6) had unusually few Medicare FFS beneficiaries transitioning from FFS into managed care (for reasons described in detail in Section V.A.2). This left 15 comparison FQHCs.

We estimated impacts on outcomes, as measured in Medicare FFS claims data, which we grouped into four domains: (1) quality-of-care processes, (2) quality-of-care outcomes, (3) service use, and (4) spending. Across the HCIA awardees in primary care redesign (PCR), we designed our impact evaluations to identify promising interventions or intervention components—consistent with the evaluation goals of the Center for Medicare & Medicaid Innovation (CMMI) to find programs that could be scaled or retested in the future. Before conducting analyses, we specified a series of primary tests, describing the evidence we would need to conclude that the program was effective, and the awardee and CMMI reviewed these tests. Each test specified a population, outcome, period, expected direction of effect, and threshold that we count as substantively important. The purpose of these primary tests was to focus the impact evaluation on hypotheses that would provide the most robust evidence about program effectiveness. We used the results from the primary tests and robustness checks, along with implementation evidence, to draw conclusions about program impacts in each of the four evaluation domains. Because we sought to identify promise, rather than only those programs with unequivocally demonstrated success, we conducted one-sided statistical tests (that is, testing

only for program benefits) and used a threshold for statistical significance of 0.1, which is not as strict as the conventional standard of 0.05.

Our impact evaluation design reflects the effects of all 21st Century Care intervention components—as well as any contemporaneous changes the awardee made to its service delivery—but only among Medicare FFS beneficiaries who received Denver Health primary care services before the intervention began. This means that, first, our impact estimates do not reflect the marginal impact of the HCIA-funded intervention alone because our impact methods cannot distinguish effects of the HCIA-funded program from those of other quality-improvement initiatives at the same time. For example, three of Denver Health's eight FQHCs also participated in CMMI's FQHC Advanced Primary Care Practice Demonstration—a demonstration that ran from October 2011 to October 2014 and tested a patient-centered medical home model. (We control for this demonstration participation in our regression models, described later in this report.) Second, as noted previously, the evaluation does not cover patients without Medicare FFS coverage (such as uninsured patients, Medicaid beneficiaries, or Medicare Advantage beneficiaries), but these excluded patients are the large majority of the Denver Health target population. To limit bias due to changes in the composition of the treatment group during the intervention period, the evaluation also excludes Medicare FFS beneficiaries who received 21st Century Care services but did not have a Denver Health primary care provider before the intervention began. These exclusions limit the generalizability of the impact findings.

III. PROGRAM IMPLEMENTATION

This section provides a detailed description of Denver Health's HCIA-funded intervention, highlighting how it evolved over time and its theory of action. Second, the section assesses the extent to which the intervention was implemented as planned based on measures of program enrollment, service delivery, staffing, training, and timeliness. Third, the section summarizes the facilitators and barriers associated with implementation effectiveness.

We based our evaluation of Denver Health's program implementation on a review of the awardee's quarterly reports to CMMI and self-monitoring program metrics, telephone discussions and follow-up communications with program administrators, and information collected during site visit interviews with frontline staff conducted in April 2014 and April 2015. We did not verify the quality of the performance data reported by the awardee in its self-measurement and monitoring reports.

A. Program design and adaptation

1. Target population and patient identification, recruitment, and enrollment

Denver Health targeted everyone who received primary care from its clinics as well as those it deemed *should* be receiving primary care from the clinics even if they were not. Denver Health operationalized this definition in the following way:

1. All primary care patients at Denver Health (defined by the awardee as any person who had a primary care visit in the previous 18 months)

2. All patients enrolled in Denver Health's managed care plan
3. Frequent users of Denver Health services who did not fall into the previous two categories

Denver Health defined frequent users as (1) people with three or more urgent-care visits, ED visits, or hospital admissions (including inpatient and observational stays) in the past 12 months; or (2) people with two or more hospital admissions, along with a serious mental health diagnosis. Qualifying mental health conditions included schizophrenic disorders and select affective and personality disorders, among others.

The target population covered roughly 130,000 people at any given time. Patients were not actively recruited and did not necessarily know about the HCIA intervention or their inclusion in the target population. Rather, as described in Sections III.A.2 and III.A.3, individual program components enrolled patients to receive specific program services.

2. Intervention components

The 21st Century Care program included four key intervention components: risk stratification, health information technology (IT) for between-visit support, enhanced primary care delivery teams within its FQHCs, and new high-risk clinics for patients with complex care needs. To refine each of these four program components over time, Denver Health used Lean processes—a system of assessment and adaptation methods, originally developed by Toyota Motor Corporation to encourage continuous improvement

Risk stratification. Denver Health used commercially available clinical risk grouping (CRG) software, augmented with in-house administrative and clinical data, to assign each patient in the target population to one of four risk-stratification tiers. Tier 1 represented the lowest-risk patients and Tier 4 represented the highest-risk patients. The goals of the risk tiering were to (1) identify patients with different levels of current and predicted medical costs; and (2) within each of the tiers, identify clinically similar groups of patients who could benefit from particular interventions. Denver Health refined the risk-tiering process throughout the award period, increasing its clinical relevance with each iteration.

Health IT for between-visit support. Denver Health invested in health IT to send patients five types of text messages: (1) appointment reminders, (2) flu vaccine reminders, (3) well-child check-up reminders, (4) diet support messages to encourage healthy eating behaviors, and (5) tobacco cessation support. Denver Health sent the text messages using an automated system designed with HCIA funding. Patients from the target population in all risk groups were eligible for services provided through this component. Before this automated system, clinical staff had to reach out to patients to remind them of appointments and preventive care. The goal of the new IT systems was to provide a low-cost way to reach a large number of patients, freeing clinical staff to focus on activities requiring direct patient–staff interaction.

Enhanced primary care delivery teams. Denver Health used HCIA funding to expand care teams to include new support roles, including clinical pharmacists, behavioral health consultants, licensed clinical social workers, and patient navigators. Most of the new staff hired were patient navigators, nonclinical staff responsible for providing assistance with appointment scheduling,

access to community resources—including transportation, housing, and social services—and other nonclinical patient needs.

Patient navigators, who joined all eight of Denver Health FQHCs, worked with patients and other members of the care team in several different ways to improve care transitions and care coordination. For example, to support care transitions, navigators (1) reviewed adult hospital discharge reports, generated daily by the Denver Health hospital; (2) contacted patients two or three days after hospital discharge using a standardized protocol that assessed transition-to-home needs; and (3) assisted the patient with scheduling a follow-up appointment with his or her primary care provider (PCP). Navigators involved additional care team members as needed in assessing transition to home needs. For example, clinical pharmacists reviewed records for all recently discharged patients to identify opportunities for medication interventions.

To support care coordination, patient navigators used reports generated by Denver Health to identify and contact patients in Tiers 3 and 4 and offer enhanced care coordination services. For patients interested in these services, navigators consulted with PCPs and other members of the care team to develop care plans and organize case conferences, in which the care team came together to discuss the patient's needs. Patient navigators were prohibited from providing patients with medical education or medical advice, but referred patients as needed to others on the care team (such as nurses and clinical pharmacists) who could provide these services.

Denver Health added clinical pharmacists to provide medication therapy management services to high-risk patients, educate providers regarding evidence-based pharmacotherapeutic care for those high-risk patients, and improve medication adherence. Denver Health also expanded a previous primary care and behavioral health pilot program that embedded behavioral health consultants in some of Denver Health's FQHCs to work with patients in need of short-term mental health counseling or other behavioral health needs.

High-risk clinics for patients with complex care needs. Denver Health created three high-risk clinics, each with a different care model and target population. Only patients in the highest-risk group (Tier 4) were eligible to receive care in one of the three specialized high-risk clinics created using HCIA funds. The intensive outpatient clinic (IOC) was a primary care clinic that focused on high-risk adults with a primary physical diagnosis and multiple comorbidities. Compared with usual care, the IOC provided a wider range of services than a typical outpatient clinic—for example providing intravenous fluids and insulin to bring down high glucose levels in a patient with diabetes, rather than sending the patient to the ED. In addition, the IOC provided one-stop access to a multidisciplinary team, including physicians, nurse practitioners, an addiction counselor, behavioral health specialists, and a social worker, and—as a result—visits were longer (in some cases several hours) than typical primary care appointments. The second high-risk clinic, co-located at the Mental Health Center of Denver (MHCD), expanded community-based case management services to adult patients with severe mental health conditions and two or more hospitalizations in the previous year. The third clinic was for children with special health care needs (CSHCN). Because our impact estimates do not capture services provided by this clinic (that is, all Medicare FFS beneficiaries included in our impact evaluation were older than 18 [results not shown] and thus too old to qualify for the clinic's

services), we do not describe it in detail in this report. For more details on the CSHCN clinic, see Higgins et al. (2015).

3. Delivering program services to the target population.

Eligibility for intervention services depended on risk tier. Table III.1 describes how each program component recruited and enrolled participants, if applicable. Figure III.1 shows the number of patients qualifying for various program services at a given point in time (December 2014).

Table III.1. Target population and patient identification, recruitment, and enrollment by program component

Program component	Target population	Identification strategy	Recruitment/ enrollment strategy	Intervention protocol	Adaptations
			Health IT		
Automated text messaging	Adults, Tiers 1–4	Denver Health enrolled patients who met specific inclusion criteria. For example, adults with a cell phone and BMI greater than 30 who had a visit within the past six months were eligible for the diet support text message program.	Patients consented to participate in specific text messaging programs.	Denver Health sent patients text messages with appointment, flu vaccine, and well child check-up reminders; diet support to encourage healthy eating behaviors; and tobacco cessation support.	(1) Denver Health developed a mechanism to make it easier to obtain patients' consent. (2) Text messaging for tobacco and weight management ended after the pilot test period because participants found it confusing.
			Enhanced primary care delivery teams		
Care transitions	Adults, Tiers 2–4, who had been hospitalized at Denver Health	Denver Health provided this intervention to patients who received primary care at one of its FQHCs. Patient navigators identified patients using adult hospital discharge reports (run daily).	There was no formal enrollment process for this intervention. Patients could have received intervention services without knowing they were part of the HCIA program.	Patient navigators contacted patients two or three days after hospital discharge using a standardized protocol to assess transition to home needs. Navigators involved additional clinical staff as needed. Clinical pharmacists reviewed records for all discharged patients to identify opportunities for medication interventions.	Denver Health held several Lean events to develop and refine the transitions-of-care intervention.

Table III.1 *(continued)*

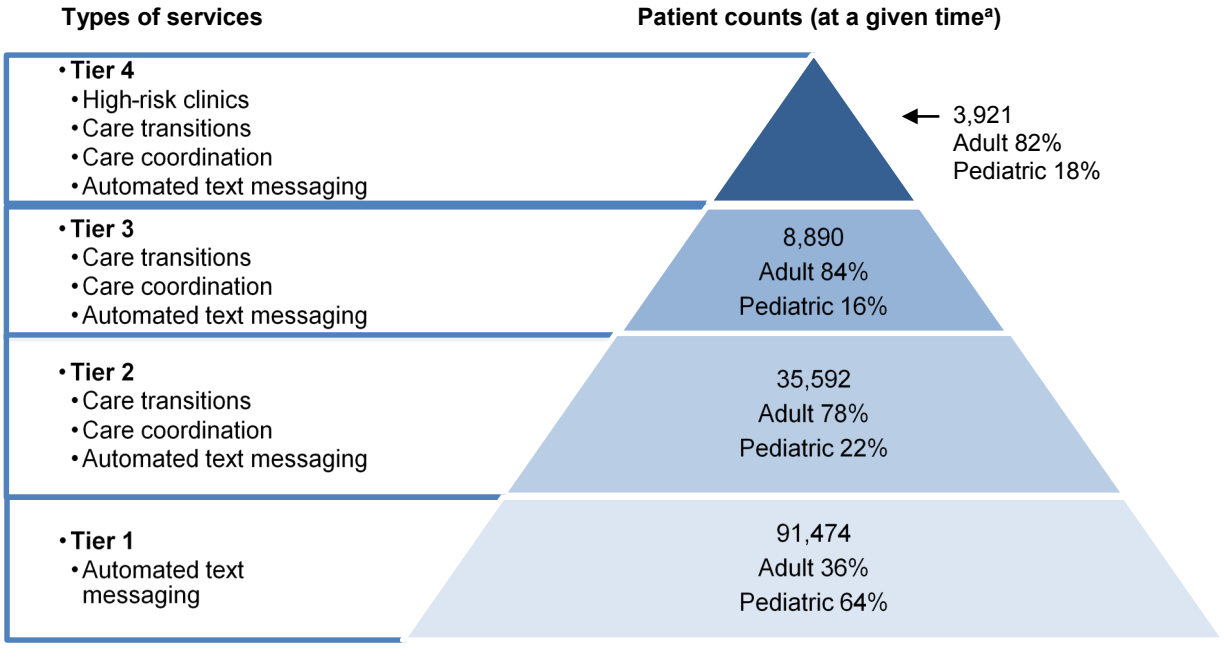
Program component	Target population	Identification strategy	Recruitment/ enrollment strategy	Intervention protocol	Adaptations
Care coordination	Adults, Tiers 3 and 4, who were assessed as high-risk and high-cost patients	Denver Health provided this intervention to patients who received primary care at one of its FQHCs. Patient navigators used Denver Health-generated reports to identify patients in Tiers 3 and 4.	Although patients had to agree to participate in care planning, there was no formal enrollment process for this intervention. Patients could have received intervention services without knowing they were part of the HCIA program.	Patient navigators contacted a list of high-risk, high-cost patients and, with patients' approval, completed adult care coordination forms. They consulted with the PCP and enhanced care team to develop a care coordination plan.	Denver Health held several Lean events to develop and refine the care coordination intervention.
			High-risk clinics[a]		
Intensive outpatient clinic (IOC)	Adults, Tier 4, with three hospital admissions within the past 12 months	Denver Health generated a daily list of patients eligible for care at this clinic.	Patient navigators contacted patients recently admitted to the hospital who qualified for the IOC. In addition, PCPs referred patients. Patients' consent was required.	The IOC was a primary care clinic for high-risk adults with a primary physical diagnosis and multiple comorbidities. The IOC allowed for longer-than-usual visits, walk-in visits, and a higher level of care team-to-patient contact, including extensive social work and nursing care management support. A multidisciplinary care team identified and addressed patients' needs.	The IOC added hospital rounding in which physicians visited IOC patients in the hospital to identify barriers to care, ensure IOC participation, and help decrease the length of stay. The IOC also added home visitation for eligible patients.
Mental Health Center of Denver (MHCD)	Adults, Tier 4, with a severe mental health condition and two or more hospitalizations in the previous year	Denver Health used a variety of methods to identify adult Tier 4 patients with a severe mental health condition and two or more hospitalizations in the previous year.	Denver Health contacted patients eligible for this clinic. In some cases, staff members visited eligible patients in the hospital to discuss the possibility of seeking follow-up care at the clinic. Patients' consent was required.	This clinic, co-located at MHCD, expanded existing community-based case management services to additional adult patients.	None

Sources: Interviews and document review.

[a] This table excludes the CSHCN clinic because we did not include the clinic's target population (children) in our impact evaluation.

BMI = body mass index; CSHCN = children with special health care needs; FQHC = federally qualified health center; HCIA = Health Care Innovation Award; IOC = intensive outpatient clinic; IT = information technology; Lean = Toyota Production System's Lean methodology; MHCD = Mental Health Center of Denver; PCP = primary care provider.

Figure III.1. Target population and types of services provided, by risk-stratification tiers



Source: Denver Health and Hospital Authority, as of December 3, 2014.

[a] Denver Health stratified patients daily. As a result, the number of patients per tier fluctuated slightly each day. This figure represents the target population at a point in time (December 3, 2014).

4. Theory of action

Based on extensive review of Denver Health's program activities and goals, we developed a theory of action to describe the mechanisms through which program administrators expected the program to improve the outcomes we selected for the impact evaluation (see Table II.1 or Section V.A.4 for a list of these outcomes). Denver Health expected 21st Century Care to improve study outcomes through more coordinated care and improved preventive care, delivered by enhanced primary care teams and new clinics for high-risk patients. The implementation process for 21st Century Care was iterative, as opposed to clearly sequential, because—as noted previously—Denver Health used Lean processes. These processes tested, assessed, refined, and then expanded program components (for example, piloting and assessing a program component in one clinic and, if found to be successful, implementing it in additional clinics). The expected theory of action had two distinct pathways:

Primary pathway to improved outcomes: Transforming the primary care delivery system to more effectively meet patients' medical, behavioral, and social needs. This pathway included the following mechanisms:

1. **Improved risk stratification leads to more efficient allocation of resources.** Denver Health develops and refines a risk-stratification algorithm, using CRGs and in-house

administrative and clinical data to assign each patient in the target population to one of four risk-stratification tiers. Denver Health makes decisions about appropriate allocation of resources by risk-stratification tier (see Figure III.1) at the health system level.

2. **Enhanced primary care teams in regular and high-risk clinics improve care coordination and care transitions support, leading to better management of chronic conditions, improved care for behavioral health issues, and better access to nonclinical services, addressing social determinants of health.**
3. **Better management of chronic conditions and better access to nonclinical services leads to reduced frequency of acute medical events among patients in Tiers 2, 3, and especially 4, resulting in fewer admissions for ambulatory care-sensitive conditions, outpatient ED visits, and all-cause admissions.**
4. **Patients in the regular and high-risk clinics who receive navigator support change their care-seeking behaviors, visiting their PCPs rather than the ED for non-urgent issues, further reducing outpatient ED visits.** Navigators assist patients with appointment scheduling and transportation and encourage them to call their PCPs before visiting the ED.
5. **Patient navigators identify patients with acute medical events and plan follow-up care, increasing rates of follow-up within 14 days of hospital discharge and decreasing 30-day unplanned readmissions.**
6. Reduced use of acute care services, especially ED visits and admissions, lead to reduced Medicare Part A and B spending.

Secondary pathway to improved outcomes: Using health IT to improve health system efficiency. Planned mechanisms of this pathway included the following.

1. **Denver Health invests in health IT, developing the capacity to send patients text message reminders between in-person appointments.** The new IT system identifies patients to receive the text message appointment reminders. Patients who wish to participate opt into the system.
2. **Text message appointment reminders reduce no-show rates and lead to more efficient scheduling for providers.**
3. **Better scheduling leads to increased capacity at Denver Health's primary care clinics, improving Denver Health's ability to deliver follow-up care for patients within 14 days of hospital discharge and reduce outpatient ED visits (due to increased availability of primary care appointments).** Increased capacity and efficiency gains also mean that Denver Health can (1) increase its total patient population, expanding primary care services to people it identifies as needing primary care due to frequent acute-care visits; and (2) reduce its own costs for managed care beneficiaries. However, these last two outcomes are not assessed in our impact evaluation.

Text box III.1. Example from Denver Health illustrating the program's theory of action

This following text is a quote from Denver Health's final progress report narrative to CMMI, illustrating the primary pathway of the theory of action. [Text in brackets are our additions]. Names have been changed to protect privacy:

"Patient story from the IOC [Intensive Outpatient Clinic]: After a sexual assault, Ms. Smith developed severe depression, requiring multiple hospitalizations. At the same time that she was struggling mentally, her physical health deteriorated. Because of a coagulation disorder, she developed a severe blood clot in her lung. The blood thinners required to treat that caused spontaneous intracranial hemorrhages (bleeding in her brain). Treatment for this has necessitated multiple neurosurgical procedures. As a result, she has developed both a seizure disorder requiring additional medications and has a chronic craniectomy. This "hole in her skull" cannot be repaired because she continues to need blood thinners to prevent future blood clots. Ms. Smith continues to experience a significant amount of anxiety related to her decline in health and function, and worries profoundly about recurrent bleeding. Since enrolling in our clinic, our social worker has been able to assist her in getting her housed in a group home, and our nurse has accompanied her to neurosurgery appointments. This has provided her with some of the additional support that she needs to manage her conditions. For example, during a recent visit with her Mental Health Center of Denver (MHCD) case manager, she complained of headache and was noted to have high blood pressure. The case manager facilitated a same day appointment [at the IOC] with a group of health care providers that knew her. She was quickly evaluated and was able to get an urgent CT [computed tomography] scan directly from the clinic, which was reassuringly negative. She was able to be discharged back to home from the clinic after little more than an hour, avoiding a lengthy emergency department visit and possible hospitalization. During the visit, her anxiety and fears over having another bleed were addressed. Our patient navigator was able to arrange her follow-up appointment including transportation, so she left with reassurance about when her next visit would occur. Working together with MHCD, we are hopeful that we can help see her through the repair of her craniectomy and achieve medical stabilization improving her health and her quality of life, while reducing unnecessary hospitalizations and emergency department visits."

5. Intervention staff and workforce development

Denver Health expanded the capacity of its primary care delivery system, adding new staff positions in its eight FQHCs (Table III.2). As discussed earlier, Denver Health also created three high-risk clinics, staffed by multidisciplinary teams designed to meet the needs of patients with the most complex conditions. Table III.2 does not include information on registered nurses or social workers hired with HCIA funds for the primary care clinics, as those staff served only pediatric patients—a population not covered by our impact evaluation. The table also excludes some staff at the high-risk clinics, such as an IOC social worker, because they were not funded by the HCIA for most or all of the award; we do not consider non-HCIA-funded staff to be part of the intervention workforce. Some HCIA workforce members had worked at Denver Health in a different capacity or clinic before the award; others, such as most of the patient navigators, were newly hired at the outset of the intervention period.

Denver Health offered nine training courses for new HCIA-funded staff, including new employee orientation; CMMI orientation, in which participants learned about the purpose, goals, and strategies of 21st Century Care; computer system training; and clinic orientation. Many patient navigators also attended a patient navigation training session at the University of Colorado.

Table III.2. Key details about intervention staff

Program components	Staff members	Staff responsibilities	Adaptations
Enhanced primary care delivery teams and high-risk clinics	Patient navigators	Denver Health placed patient navigators in all eight FQHCs. Patient navigators focused on providing between-visit care coordination and transitions of care for patients in risk-stratification Tiers 2–4. Patient navigators also worked in the high-risk clinics recruiting patients and providing care coordination. Patient navigators were not required to have clinical training and did not provide medical education.	Denver Health refined the role of patient navigators throughout the award to focus on transitions of care and high-risk care coordination.
Enhanced primary care delivery teams	Clinical pharmacists	Denver Health added clinical pharmacists to provide medication therapy management services to high-risk patients, educate providers regarding evidence-based pharmacotherapeutic care for those high-risk patients, and improve medication adherence.	None
	Behavioral health consultants	21st Century Care expanded a previous primary care/behavioral health pilot program that embedded behavioral health consultants in Denver Health's FQHCs. Behavioral health consultants worked with patients in need of short-term mental health counseling or other behavioral health needs.	None

Sources: Interviews and document review.

Note: This table refers only to staff funded by HCIA. Additional staff positions in the high-risk clinics and FQHCs were not funded by HCIA.

FQHC = federally qualified health center; HCIA = Health Care Innovation Award.

B. Implementation effectiveness

In this section, we examine the evidence on implementation effectiveness—that is, we analyze measures of the intervention delivered and, when possible, compare those measures with the services the awardee intended to deliver. We assess the evidence on implementation effectiveness in five areas: (1) success of the risk-stratification algorithm, (2) participants served and services provided, (3) staffing, (4) training and staff engagement, and (5) implementation timeliness. To conduct this analysis, we used data from interviews with program administrators and frontline staff, self-reported metrics included in Denver Health's self-monitoring and measurement reports to CMMI, and data directly from Denver Health.

1. Success of the risk-stratification algorithm

Identification of high-, medium-, and low-cost patients. Over the course of the HCIA, Denver Health went through three major iterations of its tiering algorithm, which it used daily to risk-stratify the patient population of roughly 130,000 people each month. Each algorithm iteration identified low-, medium-, and high-cost users. For example, based on Denver Health's second iteration of the tiering algorithm, implemented in May 2013, Tier 1 patients accounted for 34 percent of all adults patients and had a mean per person per month spending of $271; Tier 4 adult patients accounted for 7 percent of adults patients and had a mean per person per month spending of $4,350 (Johnson et al. 2015a). Denver Health implemented its third algorithm iteration in May 2014. With each iteration, the awardee integrated clinical perspectives into its

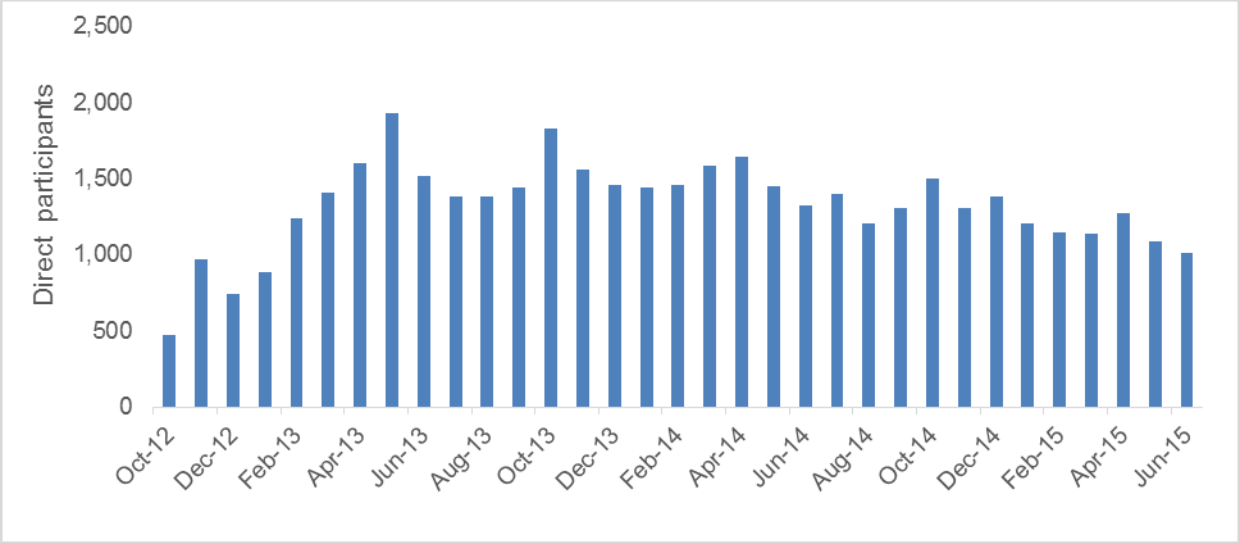
predictive modeling, with the goal of better identifying patients who could benefit from clinical intervention.

Challenges to identifying patients who might benefit from clinical intervention. Denver Health reported challenges in using utilization data alone to find patients at chronic high risk of acute care use—but identifying these patients was important for the success of 21st Century Care. That is, Denver Health assumed that 21st Century Care could reduce service use (such as hospitalizations and ED visits) by identifying patients with chronic care needs and then delivering preventive care to preempt higher-cost acute care later on. (See the theory of action in Section III.A.4.) Over the course of the award, however, Denver Health learned that many of its highest-cost patients were only temporarily high cost, suggesting that many of them would have returned to moderate- or low-cost status even without intervention. For example, under its risk stratification algorithm, Denver Health identified so-called super utilizers—all of whom were Tier 4—as people with three or more hospital admissions in a 12-month period, or two or more admissions and a mental health diagnosis. These people accounted for about 30 percent of adult facility costs. By analyzing pre-intervention data, however, research staff at Denver Health showed that, even without special intervention, fewer than half of these super utilizers at a single point in time were still in the category seven months later, and only 28 percent were in the category at the end of 12 months (Johnson et al. 2015b). Because of this challenge using utilization data alone to find chronic high-risk patients, Denver Health, as noted previously, added clinical information (in the form of both CRGs and clinical data such as lab results) to its second and third iterations of the risk-stratification algorithm (although lab results were later removed in subsequent algorithm iterations). Denver Health reported that each revision to the algorithm helped to identify patients who would benefit most from 21st Century Care's intensive services.

2. Participants served and services provided

Number of direct program participants. Denver Health defined direct participants as all patients who received services from a staff member funded by the HCIA. For example, if an HCIA-funded patient navigator contacted a patient to provide transitions-of-care support, Denver Health counted this as a direct contact because the patient navigator was an HCIA-funded position. Denver Health did not include text messaging in its direct patient counts. Over time, Denver Health transitioned many HCIA-funded staff to funding from its general operating budget. When this happened, Denver Health reclassified the services delivered by these staff members as indirect. From inception through the end of the program, the 21st Century Care program directly served 18,626 unique participants (Figure III.2). This was 93 percent of Denver Health's stated target of 20,000 direct participants.

Figure III.2. Number of unique direct participants by month, October 2012–June 2015



Source: Quarterly reports from the Health Care Innovation Award implementation contractor, the Lewin Group.

Note: Each bar represents the number of unique participants in that month. Summing two (or more) months would double-count those who participate in two (or more) months.

Number of patient encounters, by tier and type of staff member. Over the duration of the HCIA, patient navigators made 79,423 contacts with patients. These contacts included telephone and in-person conversations, letters, and other forms of communications. Over the same period, clinical pharmacists recorded 19,136 patient contacts. Behavioral health consultants served patients throughout the award period but stopped reporting data about encounters to CMMI when these staff positions were transferred from HCIA funds to other funding in 2014. For this reason, encounter information for the behavioral health consultants is not comparable to the data for other staff positions, and, for clarity, we have not included it in this report. In addition, we do not report encounter data for clinical social workers, as these staff either worked with pediatric patients only or were not funded by the HCIA for all or most of the award.

As planned, the number of patient encounters varied by tier, with higher-tier members receiving disproportionately more contacts (Table III.3). For example, Tier 4 patients comprised only about 3 percent of the patient population under the third and final iteration of the risk-stratification algorithm, but accounted for 32.8 percent of all encounters with patient navigators and clinical pharmacists during the award period. In contrast, although Tier 1 patients accounted for about 65 percent of all patients (including a large majority of the pediatric population, not covered by our impact evaluation), the people in Tier 1 accounted for only 6.4 percent of all encounters with patient navigators and clinical pharmacists. Tier 2 patients received the most patient encounters in total. However, Tier 2 also included more patients than Tiers 3 or 4, so that the rate of encounters per person per month was lower.

Table III.3. Patient navigator and clinical pharmacist contacts, by risk tier

Tier	Proportion of patient population[a]	Number of HCIA patient encounters (October 2012–June 2015)			Percentage of all patient navigator or clinical pharmacist encounters
		Patient navigators	Clinical pharmacists	Combined	
Tier 4	2.8%	17,406	19,136	36,542	32.8%
Tier 3	6.3%	18,408	6,035	24,443	21.9%
Tier 2	25.4%	35,904	6,407	42,311	38.0%
Tier 1	65.4%	6,776	405	7,181	6.4%
Not tiered	< 1.0%[a]	929	58	987	0.9%
Total	**100%**	**79,423**	**32,041**	**111,464**	**100%**

Source: Denver Health measuring and self-monitoring reports to the Center for Medicare & Medicaid Innovation.

[a] Denver Health applied its risk stratification algorithm daily, and the proportions shown here are for a single date for which Denver Health provided data: December 3, 2014. Using the version of the algorithm available on that date, no patients lacked an assigned risk tier ("Not tiered").

Mode and content of patient encounters, by type of HCIA-funded staff member. Patient navigators relied most heavily on telephone conversations (77 percent). In contrast, clinical pharmacists relied on telephone and in-person conversations, and other forms of communication, such as letters (Table III.4).

Table III.4. Patient encounters with HCIA-funded staff, by staff position

HCIA-funded position (# FTEs)	Total number of patient encounters	Most frequent mode of contact (% of all contacts)	Top reasons that encounters were initiated, according to Denver Health records	Primary action taken during encounter, according to Denver Health records
Patient navigator (23)	79,423	1. Telephone (75%) 2. In-person conversation (10%)	1. Hospital discharge 2. Appointment reminders	1. Appointment scheduled 2. Appointment reminder made 3. Coordinated patient's transition to home
Clinical pharmacist (2.5)	19,136	1. Telephone (53%) 2. In-person conversation (13%)	1. Pharmacy: diabetes management 2. Pharmacy: anticoagulation 3. Pharmacy: hypertension management	1. Patient education 2. Address adherence 3. Update medication list

Source: Denver Health measuring and self-monitoring reports to the Center for Medicare & Medicaid Innovation.
FTE = full-time equivalent; HCIA = Health Care Innovation Award.

The content of the encounters—including what prompted them and the resulting action—varied by staff members in ways that are consistent with the program design (Table III.4). According to Denver Health records, for patient navigators, the most commonly reported reasons for an encounter were that a patient was discharged from the hospital or was due for a primary

care visit. Patient navigators' most common actions were to remind patients of appointments, to schedule or reschedule appointments, and to coordinate return to home after discharge. For clinical pharmacists, the most commonly reported reasons for an encounter were to manage medications for particular conditions (diabetes, anticoagulation, and hypertension). The clinical pharmacists' most common actions were educating patients on medications, encouraging adherence to medications, and updating medications.

Text messages. Denver Health expected the text-messaging program component to be a low-cost, broad-based intervention. By June 30, 2015, program staff had invited 104,915 patients to participate in the text messaging intervention. Of these, 26 percent (27,671 patients) enrolled in the service. Denver Health had expected this modest enrollment, given that this was an opt-in service. On average, participants in this service received eight text messages over the course of the award. Most text messages were appointment reminders (51 percent), followed by flu vaccine reminders (37 percent), and well-child check reminders (11 percent—not relevant to the impact evaluation because they did not affect the Medicare FFS population). Denver Health also piloted text message programs for diet support and tobacco cessation, but discontinued these programs within about one month because they were found to be ineffective and not well received.

Enrollment in high-risk clinics. By the end of the program, Denver Health's high-risk clinics were at or near capacity. The awardee did not provide monthly enrollment counts. However, as of May 2015 the clinics had accomplished the following:

- The IOC had enrolled and treated 380 high-risk patients, nearly reaching its enrollment goal of 400 patients.
- The MHCD clinic had enrolled 85 patients, close to its capacity of 100 patients.

3. Staffing

Denver Health was largely successful in hiring intervention staff, achieving 164 percent of its cumulative new hire full-time equivalent (FTE) target. At the height of the program, or the time with the greatest number of HCIA-funded staff, Denver Health used HCIA funds to support 47.4 FTE staff positions. This included 23 FTE patient navigator positions, 2.5 FTE clinical pharmacist positions, and 1.5 FTE clinical social worker positions. (The HCIA-funded social workers served a pediatric population. Also, 2.8 FTE registered nurse positions worked exclusively in the CSHCN clinic.) Denver Health also used HCIA funding to hire 4.8 FTE clinical and clerical positions for the new high-risk clinics. Finally, Denver Health hired 12.8 FTE administrative, evaluation, and IT staffing positions for the program. The total number of positions supported by the award decreased to 22.6 FTEs in the third year of the award when the start-up health IT roles were eliminated and some staff—including behavioral health consultants in 2014 and clinical pharmacists and IOC staff in 2015—transitioned from HCIA to general operational funding.

Identifying and retaining people who were a good fit for the positions was a challenge for Denver Health. Administrators, clinicians, and staff reported a high degree of turnover among

patient navigators and added that many high-performing navigators used the position as a stepping stone to more formal training in the medical field. Denver Health administrators also reported problems finding qualified IT staff and vacant positions in the IT department were a persistent problem throughout the award.

At the end of the HCIA period, Denver Health moved all clinical, evaluation, and IT positions necessary to operate 21st Century Care into the general Denver Health operating budget. Denver Health's budget committee approved all clinical positions and funded 75 percent of the navigator program, which was sufficient to cover all filled positions at that time. These actions suggest Denver Health believed 21st Century Care was implemented effectively enough to sustain funding for these positions.

4. Training and staff engagement

Denver Health provided basic training to all staff who were funded by the intervention, as well as a training to non-HCIA-funded staff in all of the clinics involved in the intervention. According to CMMI's HCIA implementation contractor, over the course of the award, Denver Health provided training to 607 staff for a total of 9,420 hours of training from July 2012 to June 2015.

Denver Health provided additional specialized training to patient navigators. Twenty-one HCIA-funded patient navigators participated in a 32-hour patient navigation training certification course. Denver Health developed the course in collaboration with the University of Colorado Health Sciences Center; the School of Public Health at the University of Colorado administered the course. The curriculum covered knowledge and skills related to patient navigators' core competencies. To assess perspectives of HCIA-funded staff who received this and other training, we administered the HCIA Primary Care Redesign Trainee Survey from January to March 2015 (27 to 29 months after the start of implementation). A total of 17 patient navigators responded to the survey (with a response rate of 85 percent because we sent the survey to 20 patient navigators). In addition to patient navigators, a small number of staff in other roles responded to the survey; however, in this section we focus on patient navigators because they comprised the large majority of respondents, they received the most HCIA-related training, and the other respondent groups were too small for us to report results separately for them without jeopardizing respondents' confidentiality.

Almost all surveyed patient navigators (94.1 percent) reported receiving formal training (Table III.5). Most survey respondents ranked the training as excellent or good (93.8 percent) and said it was useful in their work (100 percent). Of the patient navigators who reported receiving training, most thought their training had a positive effect on the quality of care they provided (88.2 percent), patient-centeredness of care (82.4 percent), and equity (76.5 percent). In addition, a large majority of patient navigators (more than 80 percent) indicated their training positively affected various components of their jobs, including their ability to explain things to patients in lay terms, provide information to the care team, work with a diverse set of patients, help patients access medical and nonmedical services, and empower patients to take control of their care. A smaller majority (about 65 percent) reported their training helped them use data to evaluate their performance to improve the services they provided to patients. The survey data

also confirmed that patient navigators routinely managed patients' care through calling patients to check on medications and symptoms and providing follow-up services for recently discharged beneficiaries—activities the staff were expected to perform to make the program successful (Table III.6).

Table III.5. Patient navigators' perceptions of the effects of training on the care they provided to patients

Survey question		Percentage of respondents who reported the training had a positive effect on this dimension of the care they provided [a]
Please indicate how you would rate all of the training you received related to the CMMI award		93.8% (15)
Please indicate the impact you believe the training you received has had on the following aspects of care you provide to patients enrolled in 21st Century Care	1. Quality of care	88.2% (15)
	2. Ability to respond in a timely way to patients' needs	64.7% (11)
	3. Efficiency/cost-effectiveness of care	70.6% (12)
	4. Patient-centeredness	82.4% (14)
	5. Equity	76.5% (13)
Please indicate whether the training you received has had a positive or negative effect on your ability to …	1. Explain information about patient care to patients and their families in lay terms	82.4% (14)
	2. Relay relevant information to the care team	82.4% (14)
	3. Work with diverse set of patients	82.4% (14)
	4. Access the care they need	88.2% (15)
	5. Help patients access nonmedical services	88.2% (15)
	6. Help patients take control of their own care	82.4% (14)
	7. Use data to evaluate my performance to improve the services I provide to patients	64.7% (11)

Source: HCIA Primary Care Redesign Trainee Survey.

[a] The denominator for the first question only includes patient navigators who reported they received formal training from 21st Century Care (N=16). The denominator for the remaining questions includes all patient navigators who reported they received some training (formal or informal) from 21st Century Care (N=17).

CMMI = Center for Medicare & Medicaid Innovation; HCIA = Health Care Innovation Award.

Table III.6. Patient navigators' care management activities

Activity	Percentage (and number) of 17[a] patient navigators who reported that they helped to manage patients' care through this activity *routinely*
Call patients to check on medications, symptoms, or help coordinate care between visits	82.4% (14)
Educate patients about managing their own care[b]	64.7% (11)
Counsel patients on exercise, nutrition, and how to stay healthy	--[c]
Assist patients with accessing nonmedical services, such as housing, job training, and supplemental nutrition services (for example, SNAP benefits)	--[c]
Attend medical appointment with patients	--[c]
Follow up on care transitions	88.2% (15)

Source: HCIA Primary Care Redesign Trainee Survey.

[a] The denominator includes all patient navigators who reported they received some training (formal or informal) from 21st Century Care.

[b] Denver Health prohibited patient navigators from giving medical advice or education because, typically, patient navigators did not have clinical training. However, patient navigators could provide patients with approved reference materials and schedule appointments with clinicians to help patients better manage their own care.

[c] To protect respondents' confidentiality, we do not report this number because fewer than 11 respondents reported yes.

HCIA = Health Care Innovation Award; SNAP = Supplemental Nutrition Assistance Program.

Finally, in addition to formal training, Denver Health engaged its program staff by holding 114 Lean events throughout the award period to help implement 21st Century Care. At these events, staff developed new pilot projects, assessed existing pilots' successes, and planned for adaptation or scale-up of parts of the program—all with the goal of improving overall efficiency. In our interviews with program leadership and frontline staff, staff reported that Lean processes encouraged input from people in different roles throughout the Denver Health system and facilitated systemwide improvements in a more immediate way than otherwise might have been possible.

5. Implementation timeliness

Denver Health did not set timeline milestones for implementation, because it expected to experiment using Lean methods and then expand promising activities. However, within a few months of program launch, Denver Health had implemented key aspects of 21st Century Care, including developing the first version of the risk-stratification and tiering methodology (implemented in November 2012 and later refined throughout the award period), hiring and training new staff to redesign the primary care delivery teams, and creating three high-risk clinics—all of which were operating by April 2013. Denver Health launched its text messaging program in March 2013 and continued to pilot test and refine the program throughout the intervention. However, Denver Health offered its FQHCs considerable flexibility in determining

when to implement specific components of the 21st Century Care program, particularly the patient navigation activities for care transition and care coordination.

C. Summary of facilitators of and barriers to implementation

Several factors facilitated implementation of 21st Century Care, but others hindered implementation. We described those factors in detail in the second annual report (Higgins et al. 2015). Here we summarize key facilitators and barriers, along with any new information since the second annual report that supports those facilitators or barriers (Table III.7).

Table III.7. Summary of key facilitators and barriers to the implementation of Denver Health's program

Item	Description based on findings in the second annual report	Additional supporting data not available in the second annual report, if applicable
	Facilitators	
Empowered frontline staff could adapt implementation strategies and activities based on patients' needs	Denver Health empowered its frontline staff to adapt implementation strategies and activities based on the needs of their patients. For example, frontline staff in the IOC reported they expanded services to better meet patients' needs, including the addition of group visits for pain management and hospital rounding by physicians.	
System wide emphasis on self-monitoring and continuous quality improvement	Denver Health's systemwide emphasis on using Lean methods of self-monitoring and continuous quality improvement facilitated its implementation of 21st Century Care. The Lean methodology offers a process and management improvement system that relies on self-monitoring, continuous quality improvement, and the elimination of waste. Using the Lean methodology involved holding frequent rapid-improvement events with 21st Century Care team leaders and frontline staff to refine staff roles, improve processes, and redesign workflows. Staff reported the use of Lean processes encouraged input from people in different roles throughout the Denver Health system and facilitated systemwide improvements in a more immediate way than otherwise might have been possible.	Denver Health held about 114 21st Century Care Lean events over the three-year award. For example, Denver Health conducted several Lean working sessions to develop a streamlined process for referring patients to MHCD.
Buy-in among providers and staff to the integrated care team model	Widespread provider and staff buy-in into the integrated care team model facilitated implementation of 21st Century Care. PCPs and staff expressed support for continued involvement of patient navigators, behavioral health consultants, social workers, and clinical pharmacists as critical members of the care team.	

Table III.7 *(continued)*

Item	Description based on findings in the second annual report	Additional supporting data not available in the second annual report, if applicable
	Barriers	
Integration of patient navigators into care teams	Initially, Denver Health faced challenges integrating new staff, especially patient navigators, into care teams because existing staff had limited or no experience working with patient navigators. Moreover, the patient navigators' role within the care teams was not clearly defined at the beginning of the program.	
Challenges posed by small multidisciplinary teams, including personality clashes and unfilled positions	An important aspect of the new high-risk clinics was staff members' commitment to collaborate across multidisciplinary teams. This approach to care enabled patients to see multiple professionals during one visit. Because patients at high-risk clinics often had serious barriers to accessing care—including transportation and mental health issues—staff reported that one-stop shopping for medical, behavioral, and social services improved patients' overall care compliance. However, providers and staff noted that due to the small size of the teams, personality clashes or unfilled positions resulted in significant challenges in the high-risk clinics.	
Lack of an interoperable, systemwide EHR	Denver Health's lack of an interoperable, systemwide EHR was a barrier to implementation of 21st Century Care. Clinicians and staff reported that the use of multiple systems for tracking 21st Century Care activities created data and communication challenges for the integrated care teams. Denver Health is rolling out a new EHR system in 2016.	Denver Health cited the lack of an EHR as a major barrier that affected the implementation and operation of 21st Century Care.

Sources: Interviews with program staff and document review.

Note: We reviewed four domains associated with implementation experience: (1) program characteristics, (2) implementation process, (3) internal factors, and (4) external environment. Implementation research suggests that barriers and facilitators within these domains are important determinants of implementation effectiveness.

EHR = electronic health record; IOC = intensive outpatient clinic; Lean = Toyota Production System's Lean methodology; MHDC = Mental Health Center of Denver; PCP = primary care provider.

D. Conclusions about the extent to which the program, as implemented, reflects core design

21st Century Care was implemented as planned and well enough to be a reasonable test of the program's core design. Denver Health did not specify targets for most implementation metrics, so we cannot judge how close the award implementation was to the awardee's original goals. However, the awardee did deliver a meaningful intervention, as shown by its success in implementing all four program components:

Risk stratification. Denver Health developed a sophisticated risk-stratification algorithm, which it continuously refined and improved throughout the award period. Denver Health conducted the risk tiering, as planned, from the start of the intervention in October 2012 and throughout the award period. The tiering successfully organized patients into different cost categories, as planned, and over time became more sophisticated in also organizing patients (within tiers) into subgroups with similar clinical characteristics and intervention needs. It is unclear, however, how well risk stratification identified patients likely to benefit from intensive preventive care. Denver Health recognized that people with exceptionally high service use at one time did not necessarily continue to have exceptional service use in the future. Over the course of the award, Denver Health integrated clinical information into its risk-stratification algorithm to try to better identify patients who would benefit from intervention. The risk-tiering system formed the foundation for Denver Health's other 21st Century Care activities and will continue to serve Denver Health in the future.

Text messaging. Denver Health launched the text messaging service as planned. By the end of the intervention, Denver Health staff invited 104,915 patients to participate and, of these, 26 percent (27,671) enrolled. This enrollment rate is consistent with enrollment in other opt-in programs elsewhere (for example, see Laibson 2005). Denver Health eliminated several text messaging pilots based on its own assessments of their ineffectiveness, but its appointment reminder text messaging system persisted and staff considered it to be successful.

Enhanced primary care teams. All eight of Denver Health's primary care clinics incorporated new staff, including patient navigators, which were new roles in the clinics, fully funded by HCIA dollars. They exceeded their original targets for new staff hired with HCIA funds. As with other aspects of the program, Denver Health used Lean techniques to pilot test different activities for patient navigators before determining the most efficient and effective use of navigators' time: working on the care coordination and care transition duties described previously. Over the course of the award, 21st Century Care directly served 18,626 unique patients. The HCIA-funded staff directed their encounters disproportionately to patients in the higher-risk tiers, as planned. The patient navigators accounted for the bulk of the new FTEs and spent most of their time reminding patients about visits, scheduling visits, and helping to coordinate care from discharge.

High-risk clinics. All three of Denver Health's high-risk clinics were open within six months of the program starting (by April 2013), and Denver Health identified eligible patients for them based on early versions of its risk-tiering algorithm. By the end of the award period, the high-risk clinics were close to capacity.

IV. CLINICIANS' PERCEPTIONS OF THE INTERVENTION'S EFFECTS ON THE CARE THEY PROVIDED TO PATIENTS

This section describes the available evidence on the extent to which Denver Health's intervention had its intended effects on changing PCPs' behavior as a way to achieve desired impacts on patients' outcomes. Given the awardee's theory of action (Section III.A.4), we believe it was desirable but not essential for the program to change the way PCPs delivered care in the FQHCs. New staff in these primary care clinics were hired to provide *additional* support

services to augment, rather than replace, the clinical services provided by existing staff. Ideally, however, PCPs would be able to leverage the services and supports offered by new primary care support staff to spend more of their time on physician-level (or nurse practitioner-level) care.

We use data from two rounds of the HCIA Primary Care Redesign Clinician Survey to assess changes in clinicians' behavior and conclude whether such changes occurred. Both surveys rely on self-reported responses and reflect clinicians' perceptions of the program, rather than measuring quantitatively direct program effects.

A. Clinician survey

Survey methods. We administered the clinician survey in two rounds (fall 2014 and summer 2015). We sent the survey to PCPs working on 21st Century Care at the time of each survey. A total of 81 and 92 clinicians participating in Denver Health's HCIA program responded to the survey during the first and second rounds, respectively (a response rate of 67 percent in Round 1 and 65 percent in Round 2).

Survey results. Almost all respondents to the clinician survey reported being somewhat or very familiar with the HCIA program (83 percent [76 respondents] in Round 1 and 79 percent [73 respondents] in Round 2). As shown in Table IV.1, the program appeared to have intended effects for most providers; specifically, 79 to 82 percent of respondents said they thought the HCIA program improved the quality, timeliness, and patient-centeredness of care they provided to patients in their practices in the past year. The remaining respondents thought the program had no effect on those dimensions of their care or that it was too soon to tell (we did not separate the respondents into these two categories because the cell sizes were often smaller than the required 11 minimum needed for reporting). In contrast, only 15 to 49 percent of respondents said the program improved equity of care or information available for clinical decision making, with the remaining respondents reporting the program had no effect on these dimensions of care or it was too early to tell. Clinicians' perceptions of program effects were modestly higher in every domain in the second than first round of the survey.

More than 90 percent of PCPs reported engaging in team-based care—the planned model for 21st Century Care (results not shown). However, there is no clear evidence of a change in services provided to patients by the existing clinicians. In both survey waves, a majority of respondents said that 25 percent of their work or more could be done by someone with less training.

Table IV.1. PCPs' perceptions of the effects of the program on the care they provided to patients, from the clinician surveys (both rounds)

	Percentage (and number) of PCPs reporting that the HCIA had the following effect on the care they provided to patients enrolled in their practices in the past year			
	First round of survey (23 to 25 months after program implementation) N = 76		Second round of survey (31 to 33 months after program implementation) N = 73	
Dimension of care	Positive impact	No impact or too soon to tell	Positive impact	No impact or too soon to tell
Quality	67.7% (46)	22.0% (15)	79.5% (58)	--[a]
Ability to respond in a timely way to patients' needs	66.2% (45)	--[a]	82.2% (60)	--[a]
Efficiency	39.7% (27)	22.0% (15)	60.3% (44)	15.0% (11)
Safety	41.2% (28)	26.5% (18)	58.9% (43)	20.6% (15)
Patient-centeredness	69.1% (47)	17.7% (12)	79.5% (58)	--[a]
Equity	39.7% (27)	33.8% (23)	49.3% (36)	16.4% (12)
Information available for clinical decision making[a]	--[b]	--[b]	35.6% (26)	15.0% (11)

Source: HCIA Primary Care Redesign Clinician Survey: Round 1 (field period September 2014 to November 2014), Round 2 (field period May 2015 to July 2015).

Note: The number (and percentages) are limited to PCPs who reported that they were at least somewhat familiar with the HCIA program. Numbers might not sum to 100 percent because of rounding or nonresponses.

[a] Not reported because fewer than 11 respondents reported yes.

[b] This question was not asked in Round 1.

HCIA = Health Care Innovation Award; PCP = primary care provider.

NA = not available.

B. Conclusions about intermediate program effects on clinicians' behavior

Based on information available in the clinician surveys, we conclude that existing clinicians participated in team-based care and perceived the program as effective. Virtually all PCPs surveyed were aware of the program, and most believed the HCIA-funded initiative improved the quality, patient-centeredness, and timeliness of care. However, although the enhanced primary care teams provided new services to patients under 21st Century Care, we do not have evidence to assess how much of this change, if any, was the result of existing clinicians changing the way they delivered clinical services.

V. PROGRAM IMPACTS ON PATIENTS' OUTCOMES

This section of the report draws conclusions, based on available evidence, about the impacts of Denver Health's HCIA program on patients' outcomes in four domains: quality-of-care processes, quality-of-care outcomes, service use, and spending. We first describe the methods for estimating impacts (Section V.A) and then the characteristics of the 8 Denver Health treatment

FQHCs at the start of the intervention (Section V.B). We next demonstrate the extent to which treatment FQHCs were similar at the start of the intervention to the 15 FQHCs we selected as a comparison group (Section V.C). Finally, in Section V.D, we describe the quantitative impact estimates, their plausibility given implementation findings, and our conclusions about program impacts in each domain. The findings in this report update the impact results from the second annual report for Denver Health (Higgins et al. 2015), substantially revising the impact evaluation design, as well as extending the outcome period by six months and adding new outcomes to assess quality-of-care processes.

A. Methods

1. Overview

We estimated program impacts on patients' outcomes as the difference in outcomes for Medicare FFS patients served by Denver Health's 8 FQHCs and those served by 15 comparison FQHCs, adjusting for observed differences in patient and FQHC characteristics (including patients' outcomes) between these groups during the 18 months before the intervention began. We prespecified primary tests, describing the evidence we would need to conclude that the program was effective, and the awardee and CMMI reviewed these. Each test specified a population, outcome, time period, expected direction of effect, and threshold that we count as substantively important. The purpose of these primary tests was to focus the impact evaluation on hypotheses that would provide the most robust evidence about program effectiveness. We used the results from the primary and secondary tests (robustness checks) to draw conclusions about program impacts in each of the four evaluation domains. The remaining subsections describe each component of the impact evaluation in more detail.

2. Treatment group definition

The treatment group consists of Medicare FFS patients served by the 8 treatment FQHCs in six baseline quarters before the intervention began (May 1, 2011, to October 31, 2012) and 11 intervention quarters (November 1, 2012, to July 31, 2015). We defined the intervention period for the impact evaluation to start on November 1, 2012, as this was the first day of the first month after Denver Health began providing HCIA-funded services to its patients (on October 29, 2012; see Table II.1). We have defined the treatment group at the FQHC level, rather than the health system level, because we did not find a suitable health system (or health systems) in a similar market to serve as a comparison group.

To be a member of the treatment group in a given baseline or intervention quarter, each treatment beneficiary had to meet two criteria. We refer to beneficiaries as *assigned* to the treatment group in a quarter whenever both conditions are met:

1. The beneficiary had to be attributed to one of the eight treatment FQHCs *on or before the first day of the period* (either baseline or intervention). We attributed beneficiaries to FQHCs using an algorithm similar to that used by CMMI for the Comprehensive Primary Care (CPC) initiative (and similar to the one used for the other HCIA-PCR awardees under this project). However, we adapted these attribution rules slightly to align better with Denver Health's own definition of its target population. Specifically, in each baseline and

intervention month, we attributed beneficiaries to the health care provider that delivered the plurality of primary care services in the past 18 months (as opposed to 24 months for CPC and the other HCIA-PCR awardees). We counted each FQHC as its own provider (identified by its CMS Certification Number [CCN]; CMS is the Centers for Medicare & Medicaid Services). For non-FQHC practices, in contrast, we counted each physician, nurse practitioner, and physician assistant as a separate provider (identified by his or her National Provider Identifier). This means we attributed all FQHC users to the FQHCs they visited most often, unless they had more primary care visits with a clinician that did not work at an FQHC than combined visits at the FQHC. When there were ties—that is, with more than one provider tied for the plurality of visits—we attributed the beneficiary to the FQHC or provider he or she visited most recently.

2. The beneficiary had to have *observable outcomes for at least one day in the quarter* (either baseline or intervention)—and also to have a Colorado address in the Medicare Enrollment Database. Outcomes are observable in a quarter for beneficiaries enrolled in Medicare FFS (Part A and B), alive, and with Medicare as their primary payer.

This definition of the treatment group has two practical implications:

1. Using this definition, a beneficiary is not a member of the treatment group unless he or she was already a member of the treatment group in the first quarter of the period (either baseline or intervention). We defined the treatment group this way because one goal of the Denver Health HCIA was to identify frequent users of the ED and hospital and to bring these people into primary care if they did not already have a PCP. Success in doing this could mean that the population of Denver Health primary care users, on average, became less healthy during the intervention period than primary care users at comparison practices (which we expect did not generally change their target populations by expanding services to similar high-risk beneficiaries). Because we cannot easily account for changes in practices' population composition over the intervention period, we define the treatment group to have essentially constant membership—that is, with no new members joining over time (although members will leave if they move out of state or become unobservable in Medicare FFS claims by death, gaining additional insurance coverage such that Medicare is no longer the primary payer, or enrolling in managed care).
2. In addition, *a beneficiary who was previously attributed to the treatment group will remain a member of the treatment group for the rest of the relevant period* (baseline or intervention), as long as he or she is still observable in Medicare claims and living in Colorado. This definition ensures that, during the intervention period, beneficiaries do not exit the treatment group solely because the intervention succeeded in reducing their use of acute services (which could in turn limit their need for visits at treatment FQHCs). The definition for the baseline period corresponds to that of the intervention period so that, across the two periods, interpretation of the population changes over time should be comparable.

3. Comparison group definition

The comparison group (unmatched) consists of Medicare FFS beneficiaries who, by the start of the relevant period (baseline or intervention), were attributed to an FQHC that we selected as a comparison FQHC. Attribution and assignment at the comparison practices followed the same rules as attribution and assignment at the treatment practices.

To select the comparison FQHCs, we first defined a pool of potential comparison practices that comprised all FQHCs, excluding Denver Health facilities, in the 17 counties of Colorado's Front Range urban corridor. The Front Range is the most populous part of the state, ranging from the Wyoming border in the north to Pueblo County in the south. From this pool of potential comparison FQHCs, we eliminated those that appeared to be obviously poor comparisons due to characteristics not shared by any of the treatment FQHCs. Specifically, we excluded from the comparison group those FQHCs that (1) served a narrow target population or an otherwise restricted population (for example, women's clinics or Indian Health Centers); (2) offered limited services (for example, mobile or school-based clinics); (3) did not have at least 30 attributed Medicare FFS beneficiaries in every year for which we needed data for this evaluation (2009 to 2015, inclusive); (4) were staffed largely by volunteers; or (5) were associated with multiple street addresses, suggesting that more than one clinic operated under an umbrella CCN. This left a pool of 21 potential comparison FQHCs.

We next constructed potential matching variables—including demographic characteristics of the assigned beneficiaries and mean utilization during the six quarters of the baseline period—on which we aimed to achieve a high degree of similarity between the treatment and comparison FQHCs. (Section V.C shows the balance we achieved between the two groups on the potential matching variables.) We were unable to achieve substantially better balance between the treatment and comparison FQHCs by matching than by not matching. (This was true even when we expanded the potential comparison pool to include FQHCs in states outside of Colorado, but with some similar policies and cultural characteristics to Colorado: Arizona, New Mexico, Oregon, and Washington.) Thus, we selected our 15 comparison FQHCs by taking the 21 FQHCs available in the potential comparison pool in Colorado and removing 6 FQHCs that had unusually low values on one variable in particular: the proportion of Medicare FFS beneficiaries who became unobservable in Medicare claims between the first and last quarters of the baseline period (most commonly due to switching from FFS into managed care).

We removed these six outlier FQHCs because differential attrition to managed care could cause differences in population composition over time between the treatment and comparison groups, and this could violate the assumptions of the difference-in-differences model used to estimate impacts (Section V.A.5). Differences in population composition are a particular concern for this evaluation because, as noted previously, Denver Health operates a managed care plan and—as we describe in Section V.D.1—Medicare FFS beneficiaries attributed to Denver Health's FQHCs do appear more likely than FFS beneficiaries elsewhere to move into managed care. When beneficiaries switch to managed care, they become unobservable in Medicare claims, thus exiting the evaluation population.

4. Construction of outcomes and covariates

We used Medicare claims from May 1, 2008, to August 31, 2015, for beneficiaries assigned to the treatment and comparison FQHCs to develop two types of variables: (1) outcomes, defined for each beneficiary in each baseline or intervention quarter; and (2) covariates, which describe a beneficiary's characteristics at the start of the baseline and intervention periods and are used in the regression models for estimating impacts to adjust for beneficiaries' characteristics before the period began. We used covariates defined at the start of each period, without updating them each quarter, to avoid controlling in each intervention quarter for previous quarters' program effects, as this would bias the effect estimates away from detecting true impacts. Appendix 1 provides details on the methods we used to construct these variables.

Outcomes. For each beneficiary, we calculated six outcomes that we grouped into four domains:

1. Domain: Quality-of-care processes
    a. Ambulatory-care follow-up visit within 14 days of a hospital discharge (binary variable for each beneficiary); calculated as whether all of an individual's discharges in a quarter were followed by an ambulatory visit with a primary care or specialist physician within 14 days of the discharge
2. Domain: Quality-of-care outcomes
    a. Inpatient admissions (number/quarter) for select ambulatory care-sensitive conditions (ACSCs); these conditions include hypertension, angina, diabetes long-term complications, uncontrolled diabetes, lower extremity amputation, chronic obstructive pulmonary disease or asthma in older adults (ages 40 and older), perforated appendix, and dehydration)
    b. Number of inpatient admissions followed by an unplanned readmission within 30 days (number/quarter)
3. Domain: Service use
    a. All-cause inpatient admissions (number/quarter)
    b. Outpatient ED visit rate (number/quarter); outpatient ED visits are defined as ED visits or observational stays that do not end in a hospital admission
4. Domain: Spending
    a. Total Medicare Part A and B spending (dollars/month)

Four of these outcomes—all but ACSC admissions and the measure of ambulatory follow-up within 14 days of hospital discharge—are outcomes that CMMI has specified as core for the evaluations of all HCIA programs. Our definition of the readmission measure, however, differs from CMMI's standard definition. CMMI typically defines readmissions as the proportion of inpatient admissions that end in an unplanned readmission. Instead, we analyzed impacts on the *number* of these unplanned readmissions across all beneficiaries per quarter, because this enables us to look at the total impact on readmissions across the treatment group, rather than

readmissions contingent on an inpatient admission. We made this decision, in consultation with CMMI, because the intervention might also affect the number of and type of admissions.

All outcomes are quarter-specific—meaning that we calculated them for each baseline and intervention quarter separately. We defined all outcomes for all treatment and comparison group members, except for the measure of 14-day follow-up after discharge. We calculated this measure among only those beneficiaries with at least one hospital discharge in the relevant quarter.

Covariates. The covariates include (1) 14 indicators for whether a beneficiary has each of the following chronic conditions: Alzheimer's and related dementia disorders, asthma, cancer, congestive heart failure, chronic obstructive pulmonary disease, chronic kidney disease, diabetes, hip fracture, hyperlipidemia, hypertension, ischemic heart disease, rheumatoid arthritis, stroke, and any serious mental health condition (which we define, following Denver Health's program criteria, to comprise several conditions including major depression, bipolar disorder, and schizophrenia); (2) Hierarchical Condition Category (HCC) score; (3) demographics (age, gender, and race or ethnicity); (4) original reason for Medicare entitlement (old age, disability, or end-stage renal disease); (5) dual eligibility for Medicare and Medicaid; and (6) number of primary care visits at an FQHC in the previous 18 months. We defined all covariates as of the start of the relevant period (baseline or intervention).

5. Regression model

We used a regression model to implement the difference-in-differences design for estimating impacts, and estimated the model using data from all baseline and intervention quarters. For each outcome, the model estimates the relationship between the outcome and a series of predictor variables, assuming that each of the predictor variables has a linear (additive) relationship with the outcome. The predictor variables include the beneficiary-level covariates (defined in Section V.A.4); whether the beneficiary is assigned to a treatment or a comparison FQHC; an indicator for each FQHC (which accounts for static—that is, time-invariant—differences between clinics in their beneficiaries' outcomes across the baseline and intervention periods); indicators for each post-intervention quarter; and an interaction of a beneficiary's treatment status with each post-intervention quarter.

The estimated relationship between the interaction term and the outcome in a given quarter is the impact estimate for that quarter. It measures the average difference between outcomes for beneficiaries assigned to the treatment and comparison clinics during that period, subtracting out the average differences between these groups during the six baseline quarters. By providing separate impact estimates for each intervention quarter, the model enables the program's impacts to change over time. We can also test impacts over discrete sets of quarters or years, which is needed to implement the primary tests discussed in the next section. Finally, the model quantifies the uncertainty in the impact estimates, allowing for statistical tests that determine whether observed differences in outcomes between the treatment and comparison groups are likely due to chance. The model uses robust standard errors to account for clustering of outcomes across quarters for the same beneficiary and a dummy variable for each FQHC (fixed effects) to

account for clustering of outcomes for beneficiaries assigned to the same panel. The model weights all beneficiaries equally. Appendix 2 provides details on the regression methods.

One key assumption of the difference-in-differences model is that baseline differences between the treatment and comparison groups are stable—meaning that any differences at baseline would persist during the intervention period, without growing or shrinking, were it not for the effects of the intervention. If this assumption were violated it could bias our results, as we would attribute changing differences to the intervention, when in fact something else caused them.

6. Primary tests

Table V.1 shows the primary tests for Denver Health, by domain. Each test specifies a population, outcome, time period, expected direction of effect, and threshold that we count as substantively important. The purpose of these primary tests is to focus the impact evaluation on hypotheses that will provide the most robust evidence about program effectiveness (see Appendix 3 for detail and a description of how we selected each test). We provided both the awardee and CMMI an opportunity to comment on the primary tests.

Our rationale for selecting these primary tests is as follows:

- **Outcomes.** Denver Health aimed to reduce spending through decreased use of acute care services, so we estimate impacts on Medicare Part A and B spending, as well as on all-cause inpatient admissions and ED visits. Denver Health further aimed to improve quality-of-care processes and outcomes, especially preventive care for people at high risk of future health care spending (people in Tiers 2, 3, and especially 4). Because of this we also assess impacts on the 30-day unplanned hospital readmission rate, admissions for select ACSCs, and the proportion of beneficiaries with inpatient admissions who had all their admissions in a quarter followed by a primary care or specialist visit within 14 days.
- **Time period.** Denver Health expected the impacts of 21st Century Care to grow over time, with small impacts during the first year of the program and more substantial impacts in the second and third years. Most of our primary tests thus cover one time period: the second and third years of 21st Century Care. This period starts at the beginning of the fifth intervention quarter (I5), which began on November 1, 2013, and ends with the 11th intervention quarter (I11), ending July 2015, one month after the program's end in June 2015. In the spending domain, however, we analyzed impacts over two time periods: (1) the period from I5 to I11 (that is, the same period we use for the other primary tests); and (2) the final year only of the program's operation (that is, I8 through I11). The decision to analyze impacts over two different time periods reflects trade-offs between precision and anticipated effect sizes. That is, analyzing impacts over the longer time period allows for more stable estimates, based on more data, but with a smaller anticipated effect size and a greater chance that the effects were not yet fully realized (because it might take longer than one year to implement the program fully). Effects in the final year of the program might be greater, but our estimates of these effects might be less stable.

Table V.1. Specification of the primary tests for Denver Health and Hospital Authority

Domain (number of tests in the domain)[a]	Outcome (units)	Time period for impacts (controlling for baseline differences)[b]	Population	Substantive threshold (expected direction of effect)[c]
Quality-of-care processes (1)	All inpatient admissions within a quarter were followed by an ambulatory care visit with a primary care or specialist provider within 14 days (binary [yes or no]/beneficiary/year)	Intervention quarters 5 through 11	All Medicare FFS beneficiaries in the treatment group	15.0% (+)
Quality-of-care outcomes (2)	Inpatient admissions for select[d] ambulatory care-sensitive conditions (#/1,000 beneficiaries/quarter)	Intervention quarters 5 through 11	All Medicare FFS beneficiaries in the treatment group	5.0% (-)
	30-day unplanned hospital readmissions (#/1,000 beneficiaries/quarter)	Intervention quarters 5 through 11	All Medicare FFS beneficiaries in the treatment group	5.0% (-)
Service use (2)	All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	Intervention quarters 5 through 11	All Medicare FFS beneficiaries in the treatment group	5.0% (-)
	All-cause outpatient ED visits (#/1,000 beneficiaries/quarter)	Intervention quarters 5 through 11	All Medicare FFS beneficiaries in the treatment group	5.0% (-)
Spending (2)	Total Medicare Part A and B spending ($/beneficiary/month)	Intervention quarters 5 through 11	All Medicare FFS beneficiaries in the treatment group	5.0% (-)
	Total Medicare Part A and B spending ($/beneficiary/month)	Intervention quarters 8 through 11	All Medicare FFS beneficiaries in the treatment group	5.0% (-)

Notes:

[a] We adjust the *p*-values from the primary test results for multiple comparisons made within each domain, but not across domains.

[b] The regressions we use for impact analysis control for differences between the treatment and comparison groups during the baseline period.

[c] For all outcomes, we extrapolated the substantive threshold from the literature (Peikes et al. 2011; Rosenthal et al. 2016). We chose these thresholds because, for each outcome, Denver Health either set a target for improvement that was larger than the threshold we determined as substantively important based on the literature or did not set an explicit target for the Medicare FFS population.

[d] The select ambulatory care-sensitive conditions include angina, asthma in older adults (ages 40 and older), chronic obstructive pulmonary disease, dehydration, diabetes long-term complications, heart failure, hypertension, lower extremity amputation or uncontrolled diabetes, and perforated appendix. We do not include the following additional conditions because Denver Health told us that, under 21st Century Care, staff were not monitoring admissions for these conditions in particular: asthma among younger adults (ages 18 to 39), bacterial pneumonia, and urinary tract infection.

ED = emergency department; FFS = fee-for-service.

- **Population.** For all primary tests, the evaluation population is Medicare FFS beneficiaries assigned to the treatment group, as described in Section V.A.2. Denver Health's impacts should be concentrated among beneficiaries in Tiers 3 and 4 who received intensive services. However, because Denver Health's risk-tiering algorithm used clinical data in addition to administrative data, it is difficult to replicate Denver Health's highest-tier populations using Medicare claims only; the algorithm also changed over the course of the award. For both reasons, we do not specify a separate high-risk population for our analysis. Our primary tests thus assess the impacts of risk stratification and resource targeting on quality-of-care processes, quality-of-care outcomes, service use, and spending among the Medicare FFS population *overall,* rather than the effects of extra services provided to people in the highest-risk tiers only.
- **Direction (sign) of the impact estimate.** For the measure of follow-up within 14 days of a hospital discharge, we expect the impact estimate to be positive, signaling an increase in the percentage of beneficiaries receiving follow-up care. For all other outcomes, we expect the impact estimates to be negative, indicating a reduction in service use or overall expenditures.
- **Substantive thresholds.** Some impact estimates could be large enough to be substantively interesting (to CMMI and other stakeholders) even if they are not statistically significant. For this reason, we have specified thresholds for what we call substantive importance. We express the threshold as a percentage change from the counterfactual—that is, the outcomes that beneficiaries in the treatment group would have had if they had not received the HCIA-funded intervention. We extrapolated thresholds from the literature for all outcomes in all time periods (Peikes et al. 2011; Rosenthal et al. 2016), either because (1) these values were smaller than those specified by Denver Health as its anticipated effects or (2) Denver Health did not specify by how much it expected to improve these outcomes among the Medicare FFS population, in particular.

7. Secondary tests (robustness checks)

We also conducted secondary quantitative tests to help corroborate the findings from the primary tests. This is important because some of the differences observed between the treatment and comparison groups in the primary test results could reflect limitations of the non-experimental impact evaluation design or random fluctuations in the data. We have greater confidence in the primary results if they are generally consistent with the expected broader pattern of results from the secondary tests.

We conducted two sets of secondary tests for Denver Health.

1. We estimated the program's impacts on all-cause admissions, outpatient ED visits, and total Medicare spending during an additional period, not specified in the primary tests: that is, during the first four intervention quarters (I1 to I4). Because we and the awardee expected program impacts to increase over time, with little or no impacts in the first few months of the program, the following pattern would be highly consistent with an effective intervention: little to no measured effects in the first year of the program, and then larger impacts in I5 through I11. In contrast, if we found very large differences in outcomes (favorable or

unfavorable) in the first four intervention quarters, this could suggest a limitation in the comparison group, not true intervention impacts.

For this set of secondary tests, because we anticipated a large degree of statistical uncertainty in the estimates, we prespecified a threshold for differences large enough to warrant rejecting the comparison group. Based on our assessment of statistical power and the likelihood of observing large point estimates due to chance alone, we decided to reject the comparison group only if an impact estimate from I1 to I4 was more than 15 percentage points smaller or greater than the awardee's anticipated effect on the outcome (for its full population) during the same period. We used this threshold with the goal of identifying truly implausible effects early in the intervention period, but still limiting the risk that we would reject the comparison group due to statistical noise.

2. Second, we reran all of the primary tests, but limiting the sample to Medicare FFS beneficiaries who did not subsequently switch into managed care. As we will describe in detail in Section V.D.1, there was greater attrition in the treatment group than the comparison group—meaning that sample sizes fell more quickly from one quarter to the next in the treatment than comparison group. his differential attrition was driven by differences between the two groups in the rate at which people became unobservable in claims data for a reason other than death (namely, switching into managed care [Medicare Advantage], or otherwise losing Medicare FFS as the primary payer of medical bills). Because differential attrition could violate the difference-in-differences assumption of stable differences between the treatment and comparison groups in the absence of the intervention, we repeated the primary tests limiting to the population that was continuously observable during the period (baseline or intervention) for as long as each beneficiary remained alive during that period. In effect, this dropped from analysis all those beneficiaries who switched into managed care in a later quarter in the same period. (For example, a beneficiary who switched in I6 was excluded from analysis in all intervention quarters.)

8. Synthesizing evidence to draw conclusions

Within each domain, we planned to draw one of five conclusions about program effectiveness based on the primary test results, the results of secondary tests, and the plausibility of those findings given the implementation evidence:

1. Statistically significant favorable effect (the highest level of evidence)
2. Substantively important (but not statistically significant) favorable effect
3. Substantively important (but not statistically significant) unfavorable effect
4. No substantively large effect
5. Indeterminate effect

We could not conclude that a program has a statistically significant unfavorable effect because, in consultation with CMMI, we decided to use one-sided statistical tests (which do not test for evidence of unfavorable effects). We used one-sided tests to increase the probability that, if a program truly did have impacts, we would be able to detect them.

Appendix 3 describes our decision rules for each of the five possible conclusions. In short, we concluded that a program had a statistically significant favorable effect in a domain if (1) at least one primary test result in the domain was favorable and statistically significant, after adjusting the statistical tests to account for multiple tests (if applicable) within a domain; or (2) the average impact estimate across all primary tests in the domain was favorable and statistically significant. In both cases, we also had to determine that the primary test results were plausible given the results of the secondary tests and implementation evidence. We concluded that a program had a substantively important favorable effect if the average impact estimate in the domain was in the hypothesized direction and substantively important—that is, greater than the substantive threshold—but not statistically significant, and if the result was plausible given the secondary tests and implementation evidence. In contrast, if the average impact estimate was unfavorable (opposite the hypothesized direction), larger than the substantive threshold, and unfavorable effects were plausible given the other evidence, we concluded the program had a substantively important unfavorable effect. If the tests in a domain did not meet any of these criteria, we instead used the following rules. First, if the tests for at least one outcome in the domain (or all outcomes in the domain together) had sufficient statistical power to detect an impact of the size of the substantive threshold with at least 75 percent probability, we concluded there was not a substantively large effect because we could be reasonably confident that we would have detected such an effect had there been one. Second, if the power was not sufficient (less than 75 percent) to detect this type of impact, we concluded the impact in the domain was indeterminate. Indeterminate means either that the program truly did not have effects that were substantively large, or that it did, but our statistical tests were not able to detect them. Finally, if the results for the primary tests in a domain were not plausible given the implementation evidence or the secondary, corroborating tests, we did not draw any conclusions about program impacts in that domain.

B. Characteristics of the treatment group at baseline

This section describes the characteristics of the treatment group at the start of the intervention (November 1, 2012). We also show this information in the second column of Table V.2. (Table V.2 serves a second purpose—to show the similarities and differences between the treatment and comparison FQHCs at the start of the intervention—which we describe in Section V.C.)

Beneficiaries' demographics and characteristics of Medicare enrollment. Denver Health's eight treatment FQHCs serve an unusually disadvantaged Medicare population. Beneficiaries during the baseline period were much younger than the national average (60 percent were younger than 65), reflecting the fact that almost 70 percent had disability as their original reason for Medicare entitlement. About 70 percent of treatment beneficiaries were dually eligible for Medicaid, a signal of low income, and many were racial minorities; 28 percent were non-Hispanic black and 20 percent Hispanic.

Table V.2. Characteristics of treatment and comparison FQHCs before the intervention start date (November 1, 2012)

Characteristic of FQHC	Treatment group (n = 8)	Comparison group (n = 15)	Absolute difference[a]	Standardized difference[b]	Medicare FFS average
Characteristics of the FQHCs overall					
Participating in the FQHC demonstration (%)	37.5	6.7	30.83	0.84	n.a.
Characteristics of an FQHC's location					
Located in an urban zip code (%)	100	100	100	0	NA
Medicare Advantage penetration rate in the county (2011) (%)	46.0	35.7	10.33	1.13	n.a.
Demographic characteristics of assigned patients at the start of the baseline period (May 2011)					
Number of assigned beneficiaries					
Fewer than 150 (%)	25.0	40.0	-15.00	-0.30	NA
150–399 (%)	37.5	40.0	-2.50	-0.05	NA
400–899 (%)	25.0	6.7	18.33	0.54	NA
900 or more (%)	12.5	13.3	-0.83	-0.02	NA
Mean number	413.0	454.3	-41.2	-0.07	NA
Age in years					
49 or younger (%)	23.5	23.5	-0.02	0.00	NA
50–64 (%)	36.1	35.5	0.58	0.11	NA
65–74 (%)	28.0	31.6	-3.54	-0.73	45.5[c]
75–84 (%)	9.9	7.3	2.55	0.62	25.4[c]
85 or older (%)	2.5	2.1	0.44	0.25	12.4[c]
Mean age	59.1	58.7	0.44	0.17	71[d]
Male (%)	45.8	42.8	3.03	0.58	45.3[c]
Race					
Black (%)	27.8	7.0	20.80	1.29	10.4[c]
Hispanic (%)	20.4	15.6	4.82	0.60	2.6[c]
Medicare-related characteristics of assigned patients at the start of the baseline period (May 2011)					
Dually eligible for Medicaid and Medicare (%)	70.2	61.4	8.73	1.12	21.7[e]
Original reason for Medicare entitlement					
Disability (%)	69.6	68.6	0.94	0.13	16.7[c]
ESRD (%)	2.8	1.3	1.42	1.89	0.9[c]
Health status and chronic conditions of assigned patients at the start of the baseline period (May 2011)					
Mean HCC risk score	1.1	1.1	-0.01	-0.06	1.0
Cancer (%)	2.8	3.5	-0.74	-0.50	NA
CHF (%)	8.4	9.4	-1.05	-0.45	15.3[f]
CKD (%)	15.6	14.4	1.15	0.34	16.2[f]
COPD (%)	9.5	13.7	-4.17	-1.22	11.8[f]
Diabetes (%)	33.0	32.8	0.23	0.05	28.0[f]
Serious mental health condition (%)[g]	13.9	11.9	2.04	0.48	NA
Service use and spending during the first 9 months of the baseline period (May 2011 to January 2012)					
All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	90.2	87.3	2.89	0.12	74[h]
Outpatient ED visits (#/1,000 beneficiaries/quarter)	211.5	243.9	-32.38	-0.53	105[i]
Medicare Part A and B spending ($/beneficiary/month)	851	823	28.29	0.16	860[j]

Table V.2 *(continued)*

Characteristic of FQHC	Treatment group (n = 8)	Comparison group (n = 15)	Absolute difference[a]	Standardized difference[b]	Medicare FFS average
Service use and spending during the second 9 months of the baseline period (February to October 2012)					
All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	80.0	85.0	-4.97	-0.17	74[h]
Outpatient ED visits (#/1,000 beneficiaries/quarter)	231.0	242.2	-11.22	-0.22	105[i]
Medicare Part A and B spending ($/beneficiary/month)	917	809	108.01	0.64	860[j]
Service use and spending across all 18 months of the baseline period (May 2011 to October 2012)					
Primary care visits (#/1,000 beneficiaries/quarter)	830.4	907.0	-76.55	-0.51	NA
Inpatient spending ($/beneficiary/month)	1350.8	1039.4	311.45	0.96	436[l]
Quality-of-care measures across all 18 months of the baseline period (May 2011 to October 2012)					
30-day unplanned hospital readmissions (#/1,000 beneficiaries/year)	16.0	16.4	-0.40	-0.04	NA
Select ACSC admissions (#/1,000 beneficiaries/quarter)[k]	9.6	11.3	-1.73	-0.32	NA
Receipt of an ambulatory care visit within 14 days of all hospital discharges in the quarter, among those with at least one discharge in the quarter (%)	44.6	49.7	-5.12	-0.67	NA
Beneficiary attrition across all 18 months of the baseline period (May 2011 to October 2012)					
Unobservable by the end of the baseline period[l] (%)	18.7	13.9	4.80	1.61	n.a.

Source: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at the Centers for Medicare & Medicaid Services. Zip code and county information (whether an urban zip code and 2011 Medicare Advantage penetration rate) was merged from the Area Resource File.

Notes: All FQHCs are weighted equally.

Absolute differences might not be exact due to rounding.

[a] The absolute difference is the difference in means between the treatment and comparison groups.

[b] The standardized difference is the difference in means between the treatment and comparison groups divided by the standard deviation of the variable, which is pooled across the treatment and comparison groups.

[c] Chronic Conditions Warehouse (2016a).

[d] Health Indicators Warehouse (2014a).

[e] Health Indicators Warehouse (2014b).

[f] Chronic Conditions Warehouse (2016b).

[g] We use diagnosis codes provided by Denver Health to identify serious mental health conditions. Conditions include major depression, personality disorders, schizophrenia and other psychotic disorders, and others.

[h] Health Indicators Warehouse (2014c).

[i] Gerhardt et al. (2014).

[j] Boards of Trustees (2013).

[k] The select ACSCs include heart failure, hypertension, angina, diabetes long-term complications, uncontrolled diabetes, lower extremity amputation, chronic obstructive pulmonary disease or asthma in older adults (ages 40 and older), perforated appendix, and dehydration. We do not include the following because Denver Health told us that, under 21st Century Care, staff were not monitoring admissions for these conditions in particular: bacterial pneumonia, urinary tract infection, and asthma among younger adults (ages 18 to 39).

[l] Unobservable beneficiaries are those who died, moved out of Colorado, are no longer enrolled in Medicare FFS Part A and B, or no longer have Medicare as the primary payer. We define the proportion unobservable by the end of the baseline period to be the number of beneficiaries in the sixth baseline quarter divided by the number of beneficiaries in the first.

Table V.2 *(continued)*

ACSC = ambulatory care-sensitive condition; CHF = congestive heart failure; CKD = chronic kidney disease; COPD = chronic obstructive pulmonary disease; ED = emergency department; ESRD = end-stage renal disease; FFS = fee-for-service; FQHC = federally qualified health center; HCC = Hierarchical Condition Category.

NA = not available.

n.a. = not applicable.

Health and health care utilization. Consistent with its disadvantaged population, Denver Health had greater service use per beneficiary than the national average, although spending and health status (as measured by HCC) were close to the national averages. Outpatient ED visits, in particular, were common among treatment beneficiaries: The treatment group averaged 212 visits per 1,000 beneficiaries per quarter during the first three quarters of the baseline period and 231 during the second three quarters of the baseline period—more than double the national average of 105. Mean Medicare Part A and B spending per beneficiary per month was $851 during the first three quarters of the baseline period and $917 during the second three quarters, compared with a national average of $860. The mean HCC score was 1.1, meaning that, based on chronic conditions, CMS would expect beneficiary spending in the subsequent year to be 1.1 times the national average.

Characteristics of the FQHCs. All of Denver Health's FQHCs are located in Denver, but the clinics ranged in size considerably; two (25 percent) had fewer than 150 assigned beneficiaries at the start of the baseline period, whereas one—located at the Denver Health main campus—had more than 900. At the start of the intervention, three of the FQHCs (38 percent) participated in CMMI's FQHC Advanced Primary Care Practice Demonstration.

C. Comparison of treatment and comparison characteristics at baseline

Table V.2 also shows the extent to which Denver Health's 8 treatment FQHCs and the 15 selected comparison FQHCs were similar at the start of the intervention. By construction, there were no differences between the two groups on urban location; all FQHCs were located in urban counties. There were, however, some differences between treatment FQHCs and comparison FQHCs on other variables observed in claims data or in the Area Resource File. In general, we consider differences of less than 0.25 standardized differences to be small and straightforward to account for with regression adjustment. We consider differences of 0.25 or larger to indicate more substantial differences between the two groups. (The 0.25 target is an industry standard; for example, see Institute of Education Sciences [2014]).

On average, the treatment FQHCs were similar to the comparison FQHCs with respect to some, but not all, of the six evaluation outcomes during the 18-month baseline period. For example, the treatment and comparison FQHCs had similar levels of inpatient admissions both in the first nine months and in the second nine months of the period, with standardized differences in both cases well within the target of 0.25. The groups also had similar outpatient ED visit rates in the second nine months (although the treatment FQHCs' rate was lower than the comparisons' by 32 visits per 1,000 beneficiaries per quarter, or 0.53 standardized differences, during the first nine months) and the two groups had similar spending during the first nine months (although spending was $108 per beneficiary per month higher, or 0.64 standardized differences higher, among the treatment than comparison FQHCs during the second nine months). Across all 18

months of the baseline period, the treatment FQHCs had similar rates of 30-day unplanned hospital readmissions to the comparison FQHCs, lower rates of ACSCs (by 0.32 standardized differences) and lower rates of ambulatory follow-up within 14 days of a hospital discharge (by 0.67 standardized differences).

Similarly, the treatment and comparison FQHCs were well balanced on some demographic and health measures, but substantial differences existed for others. For example, despite similar age profiles, mean HCC scores, and proportions with original Medicare entitlement due to disability, the treatment FQHCs served many more ethnic minorities (with a population that on average was 20 percent Hispanic and 28 percent black, compared with only 16 percent Hispanic and 7 percent black among the comparison FQHCs). Treatment FQHCs also had higher average rates of dual eligibility for Medicare and Medicaid (70 versus 61 percent) and served more than twice the proportion of beneficiaries who qualified for Medicare due to end-stage renal disease (2.8 versus 1.3 percent).

Finally, the treatment and comparison FQHCs were roughly the same size during the baseline period (with an average of 413 and 454 beneficiaries, respectively), although the treatment FQHCs had a smaller proportion of clinics with fewer than 150 assigned Medicare FFS beneficiaries, and a higher proportion with 400 to 899. The comparison FQHCs were located in counties with lower Medicare Advantage penetration rates, on average, than Denver (36 versus 46 percent), and the rates at which treatment and comparison beneficiaries become unobservable during the baseline period reflected this difference. Table V.2 shows that, on average, 19 percent of the beneficiaries at each treatment FQHC became unobservable between the first and last baseline quarters, compared with only 14 percent at the comparison FQHCs. This is one reason we conduct secondary tests (robustness checks) described in Section V.A.7, to estimate program impacts after restricting the evaluation population to those who never exit FFS to managed care.

In sum, the 15 comparison FQHCs are not perfectly matched to the treatment FQHCs during the baseline period on all variables shown in Table V.2, but, nevertheless, we achieve reasonable balance (within 0.25 standardized differences) on several of the most important practice characteristics. These include the mean FQHC size, mean age of the attributed beneficiaries, the proportion of beneficiaries qualifying for Medicare due to disability, mean HCC score, and the number of inpatient admissions and readmissions per 1,000 beneficiaries. Moreover, changing policies and market forces are likely to affect FQHCs in Colorado's Front Range urban corridor in similar ways over the evaluation period. Thus, the treatment and unmatched comparison FQHCs are likely to be similar on some important unobservable characteristics.

D. Beneficiaries' outcomes and intervention impacts

In this section, we first present sample sizes and mean outcomes, by quarter, for the treatment and comparison groups. These mean outcomes provide context for understanding the difference-in-differences estimates that follow; however, the differences in mean outcomes are not regression-adjusted and not impact estimates by themselves. Next, we present the results of the primary tests, by domain. Then, we present the results of the secondary tests and assess whether the primary test results are plausible given the secondary test results and the implementation evidence. We end with conclusions about program impacts in each domain.

1. Sample sizes

The sample sizes for impact estimation differ depending on the outcome. We present sample sizes by domain.

Quality-of-care processes. The sample size for the 14-day follow-up measure is smaller than that for all other outcomes because we define this measure only among Medicare FFS beneficiaries who had at least one hospital stay in the quarter. For the treatment group, the sample size ranged from 141 to 251 beneficiaries across the baseline and intervention quarters (Table V.3) and this accounted for about 6.5 percent of all treatment beneficiaries in each quarter. For the comparison group, the sample ranged from 301 to 478 across the baseline and intervention quarters, accounting for a similar proportion of the total comparison group.

Quality-of-care outcomes, service use, and spending. The sample sizes for all outcomes in these three domains were the same.

In the first baseline quarter (B1), the treatment group included 3,646 beneficiaries assigned to Denver Health's 8 FQHCs and the comparison group included 7,276 beneficiaries assigned to the 15 comparison FQHCs (Table V.4). The sample sizes decreased modestly from one quarter to the next during the six baseline quarters, resulting in a roughly 18 percent reduction in total sample size from B1 to B6 for the treatment group and a roughly 13 percent reduction for the comparison group. This decrease reflects sample attrition due to beneficiaries switching from FFS Medicare to managed care, dying, or moving out of state. There was no sample addition across the baseline quarters because, as described previously (Section V.A.2), the treatment and comparison groups are defined each quarter to include only those (observable) beneficiaries already assigned in the first quarter of the period.

In the first quarter of the intervention period, the treatment group included 3,746 beneficiaries and the comparison group included 6,679. This population represented less than 5 percent of Denver Health's target population, which at any given time comprised more than 100,000 people. The sample sizes in I1 were nevertheless slightly higher than those in B6 because we defined the sample to include any observable beneficiary attributed to a relevant FQHC and living in Colorado—with no conditions on observability in the baseline period. As in the baseline period, the sample sizes then decreased between successive quarters as beneficiaries exited the treatment and comparison groups, with no new sample addition. By the end of the intervention period, I11, the treatment group included 2,317 beneficiaries (62 percent of the I1 population) and the comparison group included 4,766 (71 percent of the I1 population). This differential attrition between the treatment and comparison groups was due almost entirely to differences in the rate at which beneficiaries switched to Medicare Advantage, and not to differences in death rates or moving out of state. Analysis of beneficiaries' reason for attrition shows that the difference between treatment and comparison beneficiaries in the proportion of people dying was less than 1 percentage point, whereas the difference in the proportion of people losing FFS coverage was more than 10 percentage points (results not shown).

Table V.3. Unadjusted mean outcomes for the measure of 14-day follow-up after a hospital discharge (quality-of-care processes) among Medicare FFS beneficiaries, by treatment status and quarter

		Number of Medicare FFS beneficiaries (FQHCs)		Mean outcomes		
Period	Quarter	T	C (not weighted)	T	C	Difference (%)
Among those with at least one inpatient admission in the quarter, all inpatient admissions in the quarter were followed by an ambulatory care visit with a primary care or specialist provider within 14 days of discharge (binary [yes or no]/beneficiary/year)						
Baseline	B1	251 (8)	478 (15)	42.6	47.7	-5.1 (-10.6%)
	B2	232 (8)	449 (15)	50.9	50.1	0.8 (1.5%)
	B3	192 (8)	399 (15)	41.1	53.6	-12.5 (-23.3%)
	B4	200 (8)	399 (14)	42.5	55.4	-12.9 (-23.3%)
	B5	180 (8)	390 (15)	40.0	51.0	-11.0 (-21.6%)
	B6	176 (8)	375 (15)	49.4	53.1	-3.6 (-6.8%)
Intervention	I1	216 (8)	416 (15)	36.6	50.5	-13.9 (-27.5%)
	I2	205 (8)	402 (15)	46.8	56.7	-9.9 (-17.4%)
	I3	203 (8)	354 (15)	53.2	53.7	-0.5 (-0.9%)
	I4	214 (8)	328 (15)	45.3	55.8	-10.5 (-18.8%)
	I5	200 (8)	349 (15)	47.0	55.6	-8.6 (-15.4%)
	I6	179 (8)	332 (15)	51.4	58.1	-6.7 (-11.6%)
	I7	201 (8)	307 (15)	55.2	58.0	-2.8 (-4.8%)
	I8	150 (8)	318 (15)	54.0	57.5	-3.5 (-6.2%)
	I9	169 (8)	342 (15)	46.2	52.3	-6.2 (-11.8%)
	I10	144 (8)	316 (15)	49.3	61.7	-12.4 (-20.1%)
	I11	141 (8)	301 (15)	54.6	57.5	-2.9 (-5.0%)

Source: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at the Centers for Medicare & Medicaid Services.

Notes: The baseline quarters are measured relative to the start of the baseline period on May 1, 2011. For example, the first baseline quarter (B1) runs from May 1, 2011, to July 31, 2011. The intervention quarters are measured relative to the start of the intervention period on November 1, 2012. For example, the first intervention quarter (I1) runs from November 1, 2012, to January 31, 2013. In each period (baseline or intervention), the treatment and comparison groups each quarter include beneficiaries (1) with a hospital discharge in the quarter; (2) who were attributed to a treatment or comparison FQHC, respectively, in the first quarter of the period; and (3) who met other sample criteria—specifically, they were enrolled in Medicare FFS with Medicare as the primary payer and lived in Colorado.

Table V.3 *(continued)*

The outcome means are unweighted. The difference between the treatment and comparison groups in a quarter is calculated by subtracting the mean outcome for the comparison group from the mean outcome for the treatment group. The percentage difference equals that difference divided by the mean outcome for the comparison group.

B = baseline; C = comparison; FFS = fee-for-service; FQHC = federally qualified health center; I = intervention; T = treatment.

2. Mean outcomes for the treatment and comparison groups, by domain and quarter

The mean outcomes for the treatment and comparison groups provide context for understanding the difference-in-differences estimates that follow, but they are not, themselves, estimates of program impacts. Measures in all domains showed relatively large fluctuations in values from one quarter to the next.

Quality-of-care processes. During the baseline period, 40.0 to 50.9 percent of treatment beneficiaries with any hospital stay in a baseline quarter had all of those stays followed by an ambulatory care visit within 14 days of discharge, and the same was true of 47.7 to 55.4 percent of comparison beneficiaries with a qualifying hospital stay (Table V.3). These percentages increased modestly during the intervention period, so that by I11 the value was 54.6 percent for the treatment group and 57.5 percent for the comparison group.

Quality-of-care outcomes. In the baseline period, the number of ACSC admissions per quarter ranged from 8.9 to 15.9 per 1,000 beneficiaries in the treatment group, and from 9.5 to 13.5 per 1,000 beneficiaries in the comparison group. For both groups, the rates remained similar throughout the intervention quarters, with no discernible trend, although both groups had slightly lower minimum and maximum values than during the baseline period (7.9 and 13.2 for the treatment group and 8.5 and 12.4 for the comparison group).

For the 30-day unplanned readmissions measure, the rates were highest in B1 for both the treatment and comparison groups (20.6 and 15.0 per 1,000 beneficiaries per quarter, respectively). The treatment group rate was higher than the comparison group's in every baseline and intervention quarter except one (I2), although no clear trend in outcomes emerged during the intervention period for either group.

Service use. All-cause inpatient admissions were higher for the treatment group than the comparison group in every baseline quarter, and in all but 3 of the 11 intervention quarters. Values ranged from 74.0 to 98.7 per 1,000 beneficiaries per quarter across the baseline and intervention periods in the treatment group and from 72.9 to 85.1 in in the comparison group. As with the measure of 30-day unplanned hospital readmissions, the rates were highest for both the treatment and comparison groups in B1, but there was no clear trend over time.

The outpatient ED visit rate fluctuated over time for the comparison group, perhaps with a slight upward trend, whereas the rate increased substantially over time (also with fluctuations) for the treatment group—especially during the intervention period. The treatment group rate was 244.5 per 1,000 beneficiaries per quarter in B1 and 245.7 in I1, but reached 320.8 in I11—an increase of roughly 30 percent between the start and end of the intervention period.

Table V.4. Unadjusted mean outcomes (quality-of-care outcomes, service use, and spending) measured for all Medicare FFS beneficiaries, by treatment status and quarter

	Number of Medicare FFS beneficiaries (FQHCs)		Inpatient admissions for ambulatory care-sensitive conditions (#/1,000 beneficiaries/quarter)			30-day unplanned hospital readmissions (#/1,000 beneficiaries/quarter)			All-cause inpatient admissions (#/1,000 beneficiaries/quarter)			Outpatient ED visit rate (#/1,000 beneficiaries/quarter)			Medicare Part A and B spending ($/beneficiary/month)		
Q	T	C (no wgt)	T	C	Diff (%)	T	C	Diff (%)	T	C	Diff (%)	T	C	Diff (%)	T	C	Diff (%)
Baseline period (May 1, 2011–October 31, 2012)																	
B1	3,646 (8)	7,276 (15)	15.9	12.0	4.0 (33.0%)	20.6	15.0	5.6 (37.3%)	98.7	85.1	13.7 (16.1%)	244.5	259.4	-15.0 (-5.8%)	$870	$789	$81 (10.3%)
B2	3,492 (8)	7,089 (15)	11.2	12.7	-1.5 (-12.0%)	12.9	12.6	0.3 (2.6%)	87.6	83.9	3.7 (4.4%)	227.0	267.2	-40.2 (-15.0%)	$827	$790	$38 (4.8%)
B3	3,353 (8)	6,931 (15)	11.6	12.1	-0.5 (-4.0%)	15.5	10.7	4.8 (45.3%)	82.6	75.2	7.4 (9.9%)	202.0	247.5	-45.5 (-18.4%)	$839	$714	$125 (17.4%)
B4	3,149 (8)	6,591 (15)	8.9	13.5	-4.6 (-34.2%)	18.4	11.8	6.6 (55.6%)	85.7	80.6	5.2 (6.4%)	221.9	262.8	-40.9 (-15.6%)	$930	$788	$142 (18.1%)
B5	3,046 (8)	6,453 (15)	9.8	9.5	0.4 (4.2%)	17.7	11.9	5.8 (48.6%)	86.7	75.9	10.7 (14.1%)	249.9	272.2	-22.3 (-8.2%)	$904	$801	$103 (12.9%)
B6	2,974 (8)	6,314 (15)	11.1	10.8	0.3 (3.0%)	13.4	12.0	1.4 (11.7%)	78.3	76.0	2.3 (3.1%)	225.5	265.0	-39.6 (-14.9%)	$908	$773	$135 (17.4%)
Intervention period (November 1, 2012–July 31, 2015)																	
I1	3,746 (8)	6,679 (15)	8.3	12.4	-4.2 (-33.4%)	16.0	11.4	4.6 (40.8%)	81.2	78.9	2.2 (2.9%)	245.7	266.7	-21.0 (-7.9%)	$807	$801	$6 (0.8%)
I2	3,513 (8)	6,298 (15)	9.4	9.8	-0.5 (-4.6%)	10.5	13.0	-2.5 (-19.1%)	74.0	80.3	-6.3 (-7.9%)	243.4	260.5	-17.1 (-6.6%)	$746	$828	$-82 (-9.9%)
I3	3,404 (8)	6,160 (15)	10.3	11.2	-0.9 (-8.2%)	15.3	14.4	0.8 (5.7%)	83.1	77.1	6.0 (7.8%)	256.3	296.2	-40.0 (-13.5%)	$904	$821	$83 (10.2%)
I4	3,266 (8)	6,025 (15)	13.2	8.5	4.7 (55.5%)	16.8	12.0	4.9 (40.9%)	89.1	73.7	15.4 (20.9%)	277.5	282.8	-5.3 (-1.9%)	$974	$796	$178 (22.4%)
I5	3,133 (8)	5,875 (15)	9.6	10.2	-0.6 (-6.2%)	17.6	12.4	5.1 (41.3%)	92.9	79.0	13.9 (17.6%)	257.4	278.6	-21.2 (-7.6%)	$969	$849	$120 (14.2%)
I6	2,915 (8)	5,566 (15)	7.9	11.5	-3.6 (-31.4%)	13.7	13.3	0.4 (3.2%)	78.9	81.2	-2.3 (-2.8%)	252.3	274.2	-21.9 (-8.0%)	$893	$879	$13 (1.5%)
I7	2,770 (8)	5,407 (15)	11.2	10.0	1.2 (12.1%)	15.5	12.2	3.3 (27.2%)	97.1	72.9	24.2 (33.3%)	274.0	271.7	2.3 (0.9%)	$959	$811	$148 (18.2%)
I8	2,641 (8)	5,301 (15)	8.7	8.5	0.2 (2.6%)	15.9	11.1	4.8 (42.9%)	80.7	77.5	3.1 (4.0%)	292.5	289.0	3.5 (1.2%)	$964	$868	$96 (11.1%)
I9	2,535 (8)	5,167 (15)	10.3	11.4	-1.2 (-10.2%)	16.6	11.4	5.1 (45.1%)	96.6	84.0	12.7 (15.1%)	273.7	286.0	-12.2 (-4.3%)	$1,029	$841	$188 (22.4%)
I10	2,410 (8)	4,890 (15)	10.8	9.8	1.0 (9.9%)	14.9	9.8	5.1 (52.2%)	82.2	82.0	0.2 (0.2%)	269.1	291.5	-22.3 (-7.7%)	$941	$938	$3 (0.3%)
I11	2,317 (8)	4,766 (15)	8.6	9.0	-0.4 (-4.3%)	18.1	13.2	4.9 (37.1%)	82.4	84.6	-2.1 (-2.5%)	320.8	281.5	39.4 (14.0%)	$936	$888	$48 (5.4%)

Table V.4 *(continued)*

Source: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at the Centers for Medicare & Medicaid Services.

Notes: The baseline quarters are measured relative to the start of the baseline period on May 1, 2011. For example, the first baseline quarter (B1) runs from May 1, 2011, to July 31, 2011. The intervention quarters are measured relative to the start of the intervention period on November 1, 2012. For example, the first intervention quarter (I1) runs from November 1, 2012, to January 31, 2013. In each period (baseline or intervention), the treatment and comparison groups each quarter include all beneficiaries who (1) were attributed to a treatment or comparison FQHC, respectively, in the first quarter of the period; and (2) met other sample criteria—specifically, they were enrolled in Medicare FFS in the quarter with Medicare as the primary payer and lived in Colorado.

The outcome means are unweighted. The difference between the treatment and comparison groups in a quarter is calculated by subtracting the mean outcome for the comparison group from the mean outcome for the treatment group. The percentage difference equals that difference divided by the mean outcome for the comparison group.

B = baseline; C = comparison; Diff = difference; ED = emergency department; FFS = fee-for-service; FQHC = federally qualified health center; I = intervention; Q = quarter; T = treatment; wgt = weight.

Spending. There was no clear trend in mean Medicare Part A and B spending over time for either the treatment group (ranging from $746 to $1,029 per beneficiary per month) or the comparison group (ranging from $714 to $938). However, treatment group spending exceeded comparison group spending in every baseline and intervention quarter except one (I2). The difference-in-differences model we used to estimate impacts, described in Section V.A.5, is designed to account for persistent differences like this between the treatment and comparison groups.

3. Results for primary tests, by domain

Overview. For three of the study domains—quality-of-care processes, quality-of-care outcomes, and spending—the regression-adjusted differences between the treatment and comparison groups were small (Table V.5), and none of these differences were statistically significant or larger than the substantive thresholds in either a favorable or an unfavorable direction. In contrast, in the service use domain, we found substantively large and *unfavorable* differences between the treatment group and the estimated counterfactual, driven by large estimated differences in the outpatient ED visit rate.

Quality-of-care processes. The likelihood of receiving an ambulatory care visit within 14 days of hospital discharge was 1.1 percent higher in the treatment group than its estimated counterfactual. (Our estimate of the counterfactual, or the outcome that the treatment group would have had in the absence of the intervention, is the treatment group mean minus the regression-adjusted difference-in-differences estimate.) This (favorable) difference was neither substantively large nor statistically significant, despite good statistical power (greater than 99 percent) to detect an effect of the size of the substantive threshold (which was 15 percent).

Quality-of-care outcomes. The rate of ACSC admissions for the treatment group during the primary test period was 3.3 percent higher than our estimate of the counterfactual, and the rate of unplanned readmissions was 6.1 percent higher. These higher rates for the treatment group were in the unfavorable direction. For ACSC admissions, these findings were driven primarily by greater reductions (that is, improvements) in the comparison group, relative to the treatment group. After combining results across the two outcomes in this domain, however, the combined effect (4.7 percent) was smaller than the substantive threshold of 5 percent. The statistical power to detect effects the size of the substantive threshold was poor for both ACSC admissions (15.6 percent) and 30-day unplanned readmissions (15.9 percent). Power was also poor (17.2 percent) for the combined effect in the domain.

Service use. The treatment group's admission rate was 4.4 percent higher, and the outpatient ED visit rate was 14.2 percent higher, than the estimated counterfactuals. As in the quality-of-care outcomes domain, these higher rates for the treatment group indicate unfavorable differences between the treatment and comparison groups. The combined effect was an estimated 9.3 percent—greater than the substantive threshold of 5 percent—and also unfavorable. Power to detect (favorable) effects that were the size of the substantive thresholds was poor for all outcomes in the domain: 28.2 percent for admissions, 29.2 percent for outpatient ED visits, and 33.2 percent for the two combined.

Table V.5. Results of primary tests for Denver Health and Hospital Authority

Primary test definition					Statistical power to detect an effect that is[a]		Results			
Domain (# of tests in domain)	Outcome (units)	Time period for impacts	Population	Substantive threshold[b] (expected direction of effect)	Size of the substantive threshold	Twice the substantive threshold[c]	Treatment group mean	Regression-adjusted difference between the treatment and estimated counterfactual[b] (standard error)	Percentage difference[d]	*p*-value[e]
Quality-of-care processes (1)	All inpatient admissions within a quarter were followed by an ambulatory care visit with a primary care or specialist provider within 14 days (binary [yes or no]/beneficiary/year)	Average over intervention quarters 5–11 (November 1, 2012, to July 31, 2015)	Medicare FFS beneficiaries with at least one hospital stay in the quarter and assigned to treatment FQHCs	15.0% (+)	95.0%	> 99.9%	51.1	0.6 (2.6)	1.1%	0.41
Quality-of-care outcomes (2)	Select inpatient admissions for select ambulatory care-sensitive conditions (#/1,000 beneficiaries/quarter)[f]	Average over intervention quarters 5–11 (November 1, 2012, to July 31, 2015)	All Medicare FFS beneficiaries assigned to treatment FQHCs	5.0% (-)	15.6%	23.0%	9.6	0.3 (1.7)	3.3%	0.51
	30-day unplanned readmissions (#/1,000 beneficiaries/quarter)	Average over intervention quarters 5–11 (November 1, 2012, to July 31, 2015)	All Medicare FFS beneficiaries assigned to treatment FQHCs	5.0% (-)	15.9%	23.8%	16.0	0.9 (2.7)	6.1%	0.54
	Combined (%)	Average over intervention quarters 5–11 (November 1, 2012, to July 31, 2015)	All Medicare FFS beneficiaries assigned to treatment FQHCs	5.0% (-)	17.2%	26.9%	n.a.	n.a.	4.7%	0.62
Service use (2)	All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	Average over intervention quarters 5–11 (November 1, 2012, to July 31, 2015)	All Medicare FFS beneficiaries assigned to treatment FQHCs	5.0% (-)	28.2%	55.2%	87.3	3.7 (5.9)	4.4%	0.61

Table V.5 *(continued)*

Primary test definition					Statistical power to detect an effect that is[a]		Results			
Domain (# of tests in domain)	Outcome (units)	Time period for impacts	Population	Substantive threshold[b] (expected direction of effect)	Size of the substantive threshold	Twice the substantive threshold[c]	Treatment group mean	Regression-adjusted difference between the treatment and estimated counterfactual[b] (standard error)	Percentage difference[d]	*p*-value[e]
	Outpatient ED visits (#/1,000 beneficiaries/quarter)	Average over intervention quarters 5–11 (November 1, 2012, to July 31, 2015)	All Medicare FFS beneficiaries assigned to treatment FQHCs	5.0% (-)	29.2%	57.3%	277.1	34.5 (16.6)	14.2%	0.96
	Combined (%)	Average over intervention quarters 5–11 (November 1, 2012, to July 31, 2015)	All Medicare FFS beneficiaries assigned to treatment FQHCs	5.0% (-)	33.2%	66.0%	n.a.	n.a.	9.3%	0.94
Spending (2)	Medicare Part A and B spending ($/beneficiary/month)	Average over intervention quarters 5–11 (November 1, 2012, to July 31, 2015)	All Medicare FFS beneficiaries assigned to treatment FQHCs	5.0% (-)	36.0%	71.3%	$956	$8 (51.4)	0.9%	0.52
	Medicare Part A and B spending ($/beneficiary/month)	Average over intervention quarters 8–11 (August 1, 2014, to July 31, 2015)	All Medicare FFS beneficiaries assigned to treatment FQHCs	5.0% (-)	31.9%	63.4%	$968	-$1 (59.7)	-0.1%	0.50
	Combined (%)	Varies by test	All Medicare FFS beneficiaries assigned to treatment FQHCs	5.0% (-)	34.4%	68.4%	n.a.	n.a.	0.4%	0.53

Table V.5 *(continued)*

Source: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at the Centers for Medicare & Medicaid Services.

Note: The results for each outcome are based on a difference-in-differences regression model, as described in the text. Estimates are calculated for Medicare beneficiaries who are observable in the relevant time period: that is, beneficiaries who are enrolled in Medicare FFS (Part A and B), are alive, and have Medicare as their primary payer. Additional sample restrictions apply to the quality-of-care process measures; see text for details.

[a] The power calculation is based on actual standard errors from the analysis. For example, for all-cause inpatient admissions, a 5.0 percent effect (from the estimated counterfactual of 87.3 – 3.7 = 83.6) would be a change of 4.2 admissions per 1,000 beneficiaries per quarter. Given the standard error of 5.9 from the regression model, we would be able to detect a statistically significant result 28.2 percent of the time if the impact was truly 4.2 admissions, assuming a one-sided statistical test at the $p = 0.10$ significance level.

[b] We show statistical power to detect a very large effect (twice the size of the substantive threshold) because this provides additional information about the likelihood that we will find effects if the program is indeed effective. If power to detect effects is less than 75 percent even for a very large effect, then the evaluation is extremely poorly powered for that outcome.

[c] The substantive threshold is the impact as a percentage of the counterfactual. The counterfactual is the outcome the treatment group would have had in the absence of the HCIA-funded intervention.Our estimate of the counterfactual is the treatment group mean minus the regression-adjusted difference-in-differences estimate.

[d] Percentage difference is calculated as the regression-adjusted difference between the treatment and comparison group, divided by the adjusted comparison group mean.

[e] *p*-values test the null hypothesis that the regression-adjusted difference-in-differences estimate is less than or equal to zero for outcomes in the quality-of-care processes domain, or greater than or equal to zero in all other domains (a one-sided test). Because each test is a one-sided test, as the difference-in-differences estimate approaches infinity in an unfavorable direction (negative for the quality-of-care process measure and positive for all other measures), the *p*-value approaches 1, whereas it would approach 0 in a two-sided test. We adjusted the *p*-values for the multiple (two) comparisons made within the quality-of-care outcomes domain, and (separately) for the two comparisons made within the service use domain, and for the two comparisons made within the spending domain.

[f] The select ambulatory care-sensitive conditions include heart failure, hypertension, angina, diabetes long-term complications, uncontrolled diabetes, lower extremity amputation, chronic obstructive pulmonary disease or asthma in older adults (ages 40 and older), perforated appendix, and dehydration. We do not include the following because Denver Health told us that, under 21st Century Care, staff were not monitoring admissions for these conditions in particular: bacterial pneumonia, urinary tract infection, and asthma among younger adults (ages 18 to 39).

ED = emergency department; FFS = fee-for-service; FQHC = federally qualified health center; HCIA = Health Care Innovation Award.

n.a. = not applicable.

Spending. The treatment group averaged $956 per beneficiary per month in Part A and B spending during the 5th through 11th intervention quarters, a value 0.9 percent (or $8 per beneficiary per month) higher than the estimated counterfactual. Monthly per-beneficiary spending was 0.1 percent (or $1) lower in the treatment group than the estimated counterfactual in the final four quarters of the intervention: I8 through I11. Both differences—and the combined difference of 0.4 percent—were much smaller than the substantive threshold of 5 percent. Statistical power to detect an effect the size of the substantive threshold was, once again, poor: 36.0 percent for the longer time period, 31.9 percent for the shorter time period, and 34.4 percent for the two tests combined.

Aggregate estimates for CMMI's core measures. The estimates presented for the CMMI core outcomes—that is, for 30-day unplanned readmissions, all-cause inpatient admissions, the outpatient ED visit rate, and Medicare Part A and B spending—have so far been expressed per 1,000 beneficiaries per quarter (or, for spending, per beneficiary per month). Table V.6 translates these rates or per-beneficiary-month estimates into estimates of aggregate impacts during the 21-month primary test period. (For the spending outcome, we estimated impacts over two time periods in the primary tests, but we present aggregate impacts only for the longer time period because it includes the shorter time period as well.) We calculated these aggregate impacts by multiplying the point estimates from Table V.5 by the average number of Medicare beneficiaries in the treatment group and by the number of quarters or months during the primary test period. Because the point estimates in Table V.5 are in the unfavorable direction for all four of the core outcomes (although generally small), the values in Table V.6 are positive—reflecting the estimated *additional* readmissions, inpatient admissions, and outpatient ED visits that occurred, and the Medicare Part A and B dollars spent, as a result of the HCIA-funded intervention. These aggregate estimates appear larger than the quarterly or monthly point estimates in Table V.5 because they are scaled to cover the entire 1.75-year period of the primary tests and the roughly 2,700 people, on average, in the treatment group during that same period. Even large aggregate estimates should be interpreted with caution, however, because the estimates were not statistically significant for any of the outcomes. The *p*-values for the aggregate estimates are the same as they are for the main results shown in Table V.5.

Table V.6. Results for primary tests for CMMI's core outcomes expressed as aggregate effects for all Medicare FFS beneficiaries in the treatment group

Outcome (units)	Aggregate impact estimate during the 5th through 11th intervention quarters (November 1, 2013, through July 31, 2015)	*p*-value
30-day unplanned readmissions (#)	+17	0.54
All-cause inpatient admissions (#)	+69	0.61
Outpatient ED visits (#)	+645	0.96
Medicare Part A and B spending ($)	+$470,955	0.52

Sources: Authors' calculation, based on analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at the Centers for Medicare & Medicaid Services.

Note: To estimate the aggregate impact during the primary test period (intervention quarters 5 through 11) we (1) multiplied the per beneficiary per quarter (or month) estimate from Table V.5 by the average number of Medicare FFS beneficiaries in the treatment group during the seven primary test quarters and then (2) scaled the estimate to 1.75 years by multiplying the resulting product by 7 (for outcomes measured per quarter) or 21 (for outcomes measured per month). The *p*-values are taken from Table V.5 and are therefore one-sided (testing that the program improved outcomes) and adjusted for multiple comparisons conducted within each outcome domain.

CMMI = Center for Medicare & Medicaid Innvation; ED = emergency department; FFS = fee-for-service.

4. Results for secondary tests

Estimates during the first intervention year (November 1, 2012, to October 31, 2015). As shown in Table V.7, the differences in admissions and spending for the treatment group and its estimated counterfactual were small (less than 2.5 percent) and not statistically significant during the first year of the intervention (I1 through I4). Differences in outpatient ED visits were somewhat larger, at 7.6 percent, and larger than the substantive threshold we stated for that outcome in the primary tests (5 percent). However, this difference was still much smaller than the threshold we set in Section V.A.7 as a criterion for rejecting the comparison group. Together, these secondary test results help support the credibility of the comparison group because we do not see large or statistically significant impact estimates during the first year of program participation, a period during which we and the awardee did not expect to see large program effects.

Estimates with the sample limited to prevent attrition into managed care. Results excluding any beneficiary who later switched into Medicare Advantage during the period were generally consistent with the primary test results. These secondary test results (Table V.7, bottom panel) were of roughly the same magnitude as the results of the primary test results (Table V.5), and suggest no changes to the interpretation of statistically significant or substantively important effect sizes. These results give us confidence that differential attrition between the treatment and comparison groups, although substantial, had no meaningful effect on our impact estimates.

5. Consistency of impact estimates with implementation findings

The primary test impact estimates are plausible in the domains of quality-of-care processes, quality-of-care outcomes, and spending. Despite evidence that Denver Health implemented its HCIA-funded program largely as planned, it is always plausible that a well-implemented program did not have its intended effects on patients' outcomes.

The primary test results in service use are less credible, however. In particular, the implementation evidence cannot explain how the program might have caused the large (14.2 percent) observed increase in the outpatient ED visit rate. It is possible that patient navigators might have diverted occasional 21st Century Care patients to the ED when those patients otherwise would not have gone. (For example, one could imagine a patient having a panic attack and calling a patient navigator; when the patient reported shortness of breath and tightness in the chest, the patient navigator might have transferred the call to a clinician who then felt compelled to send that patient to the ED, even if both the patient and the clinician suspected anxiety rather than a heart attack.) Still, it is difficult to believe this type of diversion to the ED—counter to the goals and expected processes of 21st Century Care—would have occurred so frequently as to cause such a large unintended impact on outpatient ED visits: an effect nearly three times the size of the substantive threshold. According to Denver Health's self-monitoring data reported to CMMI, of all the patient navigator contacts with patients during the award period, fewer than 0.01 percent ended in a referral to the ED. We have explored other possible explanations for the result, such as changes in billing practices for observational stays, which are included in our measure of outpatient ED visits, but we see no evidence of such changes (results not shown). We also asked Denver Health program administrators if they knew of anything that could explain the phenomenon. They noted that freestanding EDs have become increasingly common in Colorado. If these freestanding EDs were more likely to open in Denver than in other Colorado counties where comparison beneficiaries lived, it is possible that this could have created differences in ED use between the treatment and comparison beneficiaries—although it is not obvious this happened. In short, we see no plausible mechanism by which the program could have caused an increase in ED visits of the observed magnitude.

Table V.7. Results of secondary tests for Denver Health and Hospital Authority

Secondary test definition				Results			
Domain	Outcome (units)	Time period for impacts	Population	Treatment group mean	Regression-adjusted difference between the treatment and estimated counterfactual (standard error)[a]	Percentage difference[b]	*p*-value[c]
Secondary tests for a time period when large impacts are not expected							
Service use	All-cause inpatient admission rate (#/1,000 beneficiaries/quarter)	Average over intervention quarters 1–4	All observable Medicare FFS beneficiaries	81.9	1.9 (5.4)	2.3%	0.63
	Outpatient ED visits (#/1,000 beneficiaries/quarter)	Average over intervention quarters 1–4	All observable Medicare FFS beneficiaries	255.7	18.1 (14.9)	7.6%	0.89
Spending	Medicare Part A and B spending ($/beneficiary/month)	Average over intervention quarters 1–4	All observable Medicare FFS beneficiaries	$858	-$18 (50.3)	-2.1%	0.36
Secondary tests limiting to population that does not transfer into managed care							
Quality-of-care processes	All inpatient admissions within a quarter were followed by an ambulatory care visit with a primary care or specialist provider within 14 days (binary [yes or no]/beneficiary/year)	Average over intervention quarters 5–11	Medicare FFS beneficiaries in the treatment group who, throughout the period (baseline and intervention) are never enrolled in managed care or have a payer other than Medicare as the primary payer	50.1	0.2 (2.7)	0.5%	0.47
Quality-of-care outcomes	Inpatient admissions for select[d] ambulatory care-sensitive conditions (#/beneficiary/quarter)	Average over intervention quarters 5–11	Medicare FFS beneficiaries in the treatment group who, throughout the period (baseline and intervention) are never enrolled in managed care or have a payer other than Medicare as the primary payer	10.1	0.6 (1.8)	6.6%	0.64
	30-day unplanned hospital readmissions (#/1,000 beneficiaries/quarter)	Average over intervention quarters 5–11	Medicare FFS beneficiaries in the treatment group who, throughout the period (baseline and intervention) are never enrolled in managed care or have a payer other than Medicare as the primary payer	17.1	0.5 (2.8)	3.2%	0.57
Service use	All-cause inpatient admission rate (#/1,000 beneficiaries/quarter)	Average over intervention quarters 5–11	Medicare FFS beneficiaries in the treatment group who, throughout the period (baseline and intervention) are never enrolled in managed care or have a payer other than Medicare as the primary payer	90.5	3.7 (6.3)	4.3%	0.72

Table V.7 *(continued)*

Secondary test definition				Results			
Domain	Outcome (units)	Time period for impacts	Population	Treatment group mean	Regression-adjusted difference between the treatment and estimated counterfactual (standard error)[a]	Percentage difference[b]	*p*-value[c]
	Outpatient ED visits (#/1,000 beneficiaries/quarter)	Average over intervention quarters 5–11	Medicare FFS beneficiaries in the treatment group who, throughout the period (baseline and intervention) are never enrolled in managed care or have a payer other than Medicare as the primary payer	280.5	41.7 (17.7)	17.5%	0.99
Spending	Total Medicare Part A and B spending ($/beneficiary/month)	Average over intervention quarters 5–11	Medicare FFS beneficiaries in the treatment group who, throughout the period (baseline and intervention) are never enrolled in managed care or have a payer other than Medicare as the primary payer	$985	-$8 (54.8)	-0.8%	0.44
	Total Medicare Part A and B spending ($/beneficiary/month)	Average over intervention quarters 8–11	Medicare FFS beneficiaries in the treatment group who, throughout the period (baseline and intervention) are never enrolled in managed care or have a payer other than Medicare as the primary payer	$979	-$22 (62.6)	-2.2%	0.36

Notes: The results for each outcome are based on a difference-in-differences regression model, as described in the text. Estimates are calculated for Medicare beneficiaries who are living in Colorado and observable in the relevant time period: that is, beneficiaries who are enrolled in Medicare FFS (Part A and B), are alive, and have Medicare as their primary payer. We identify those who switch into managed care as those who had FFS as their primary payer and then cease to, but who remain alive and living in Colorado.

[a] The counterfactual is the outcome the treatment group beneficiaries would have had in the absence of the HCIA-funded intervention. Our estimate of the counterfactual is the treatment group mean minus the regression-adjusted difference-in-differences estimate.

[b] Percentage difference is calculated as the regression-adjusted difference-in-differences estimate, divided by the estimate of the counterfactual.

[c] *p*-values test the null hypothesis that the regression-adjusted difference-in-differences estimate is less than or equal to zero for outcomes in the quality-of-care processes domain, or greater than or equal to zero in all other domains (a one-sided test). Because each test is a one-sided test, as the difference-in-differences estimate approaches infinity in an unfavorable direction (negative for the quality-of-care process measure and positive for all other measures), the *p*-value approaches 1, whereas it would approach 0 in a two-sided test. Values are *not* adjusted for multiple comparisons.

[d] The select ambulatory care-sensitive conditions include heart failure, hypertension, angina, diabetes long-term complications, uncontrolled diabetes, lower extremity amputation, chronic obstructive pulmonary disease or asthma in older adults (ages 40 and older), perforated appendix, and dehydration. We do not include the following because Denver Health told us that, under 21st Century Care, staff were not monitoring admissions for these conditions in particular: bacterial pneumonia, urinary tract infection, and asthma among younger adults (ages 18 to 39).

ED = emergency department; FFS = fee-for-service; HCIA = Health Care Innovation Award.

6. Conclusions about program impacts on Medicare FFS beneficiaries, by domain

Based on all evidence currently available, we have drawn the following conclusions about program impacts of Denver Health's HCIA intervention. Table V.8 summarizes these conclusions and their support.

- **The program did not have a substantively large impact on quality-of-care processes.** The primary test result for the one outcome in this domain was neither substantively large nor statistically significant, despite good statistical power (greater than 99 percent) to detect effects if they existed. The secondary test results support this primary test result by supporting the comparison group, showing that (1) impact estimates were not overly large during the first year of 21st Century Care, before large impacts were expected; and (2) differential attrition between the treatment and comparison groups did not greatly influence the impact estimates. The conclusion of no substantively large effects in the domain is also consistent with implementation findings because, although the program was implemented reasonably well, it is plausible the program did not have its intended effects.
- **The program had an indeterminate effect on quality-of-care outcomes and spending.** The primary test results were neither substantively large nor statistically significant in either domain. However, for each outcome, the statistical power was poor (less than 40 percent) to detect effects the size of the substantive threshold. As a result, null findings from the primary tests in these domains could be due to (1) the program truly not having substantively large effects, at least among Medicare FFS beneficiaries; or (2) the program having substantively large effects but our tests failing to detect them. We have little evidence to judge which explanation is most plausible.
- **We cannot draw conclusions in the service use domain.** The primary test results showed a substantively large unfavorable estimate in this domain, driven by an estimated 14.2 percent increase in the treatment group's outpatient ED visit rate, relative to the counterfactual. This result is reasonable given the secondary test results, which generally support the comparison group (although, already in the first year of the intervention, the secondary tests suggest an unfavorable difference between the treatment group and its counterfactual that is greater [7.6 percent] than the substantive threshold [5.0 percent] from the primary tests, despite no expected large effects during this period). However, as we describe in Section V.D.5, the primary test results are not consistent with the implementation evidence, as we have no plausible explanation for how the program could have caused such a large unfavorable impact on outpatient ED visits. Although we believe the evaluation methods are sound overall, we cannot draw a conclusion in this one domain. Statistical power to detect effects was poor for all outcomes in the domain, and it is possible the large, 14.2 percent observed impact estimate for ED visits was due to chance.

Table V.8. Conclusions about the impacts of Denver Health's HCIA program on patients' outcomes, by domain

		Evidence supporting conclusion		
Domain	Conclusion	Primary test result(s)	Primary test result(s) plausible given secondary tests?	Primary test result(s) plausible given implementation evidence?
Quality-of-care processes	No substantively large effects	• No substantively large or statistically significant effects; well-powered to detect effects on the one outcome in the domain	Yes	Yes
Quality-of-care outcomes	Indeterminate effect	• No statistically significant effect on either outcome in the domain and no substantively important effect for the combined effect estimate; power was poor for all statistical tests in the domain	Yes	Yes
Service use	No conclusion	• A combined effect across the two measures in the domain that was unfavorable and larger than the substantive threshold	Yes	No
Spending	Indeterminate effect	• No statistically significant effect in either time period in the domain and no substantively important effect for the combined effect estimate; power was poor for all statistical tests in the domain	Yes	Yes

Sources: Tables V.5 and V.7.

HCIA = Health Care Innovation Award.

VI. DISCUSSION AND CONCLUSIONS

Denver Health used its $19.8 million in HCIA funds to implement 21st Century Care, a program with multiple components to better meet patients' medical, behavioral, and social needs. The program had four main components: (1) using administrative and clinical data to risk-stratify the patient population, enabling more efficient resource allocation within the health system; (2) upgrading and leveraging health IT to provide between-visit support, especially in the form of text-message reminders about upcoming appointments and recommended preventive services; (3) developing a new staffing model for Denver Health's eight FQHCs, integrating clinical pharmacists, behavioral health consultants, social workers, and (especially) nonclinician patient navigators to support routine clinical care; and (4) creating three new high-risk clinics to provide individualized care to patients with unusually complex health needs. Through these four intervention components, Denver Health aimed to improve preventive care for patients at risk of acute exacerbations and thus reduce acute care use, including ED visits and inpatient admissions. This, in turn, was expected to reduce total Medicare and Medicaid spending. Denver Health also expected that its text messaging program component would reduce no-shows and decrease the time staff needed to schedule appointments and remind patients about preventive care, freeing time for clinical visits that required staff–patient interaction.

Our impact results are largely indeterminate, although they cover only a small subgroup (less than 5 percent) of Denver Health's target population. Importantly, the treatment group excluded subpopulations where impacts might have been concentrated, including people without Medicare coverage or without a Denver Health primary care provider at the start of the HCIA intervention. Outcomes for the treatment beneficiaries in this evaluation—that is, Medicare FFS beneficiaries who received primary care services at Denver Health's 8 FQHCs before the intervention began—were not statistically or substantively better than those for comparison beneficiaries at 15 FQHCs in surrounding urban counties. In fact, in one evaluation domain (service use), outcomes appeared *worse* for the treatment than comparison beneficiaries, although this result—driven by an unusually high outpatient ED visit rate among the treatment group—might be spurious. The lack of observed effects in the other three evaluation domains (quality-of-care processes, quality-of-care outcomes, and spending) is unsurprising given generally poor statistical power. The evaluation was well powered to detect substantively large impacts on only one outcome, in the domain of quality-of-care processes: the proportion of beneficiaries with a hospital stay for whom all stays in the quarter were followed by an ambulatory visit within 14 days; however, we found no effects in this domain. For all other evaluation domains and outcomes—including all four of CMMI's core outcomes for the HCIA evaluations—we had poor statistical power (less than 50 percent) to detect effects the size of our prespecified thresholds of substantive importance.

The lack of favorable impact estimates does not appear to be due to major problems implementing the intervention. 21st Century Care operations began on schedule in fall 2012, and Denver Health implemented the four program components it described in its original HCIA application. Several measures capture the generally successful implementation:

- All three of Denver Health's high-risk clinics were open within six months of the program starting (that is, by April 2013), and Denver Health could identify eligible patients for them

based on early versions of its risk-tiering algorithm (continuously refined throughout the award period).

- HCIA-funded staff provided direct services to 18,626 unique patients at Denver Health. Patient navigators made more than 75,000 patient contacts and clinical pharmacists made nearly 20,000.
- As expected, 21st Century Health's highest-risk patients (Tier 4) received the greatest number of encounters (per person) with patient navigators and clinical pharmacists. This suggests program resources were allocated to high-cost, high-utilizing patients as intended.
- The text messaging service launched as planned. By the end of the intervention, Denver Health staff had invited 104,915 patients to participate and, of these, 26 percent (27,671) enrolled, receiving on average eight text messages each over the course of the intervention.
- Denver Health executive leaders and staff appeared committed to continuous program refinement. Over the course of the award period, Denver Health held 114 Lean events to discuss and improve various aspects of 21st Century Care.

Further, the lack of observed favorable effects appears not to be due to an inability to engage Denver Health clinicians as planned. Clinical staff who continued to work in the regular primary care clinics (as opposed to high-risk clinics created with HCIA funds) were not central to Denver Health's theory of action because—although clinicians had to work closely with new support staff—they were not themselves expected to provide new services. Nevertheless, in our surveys of Denver Health clinicians, most respondents reported that they felt the program improved the efficiency, safety, patient-centeredness, and equity of their care.

Instead, we believe the lack of observed favorable program impacts might be due to one of three factors (or some combination of them). First, it is possible we do not observe substantively large or statistically significant favorable effects because the program truly did not have these effects. This would mean the program, although implemented well, failed to reduce patients' needs for acute care and, in turn, reduce spending. Second, it is possible the Denver Health program had its intended effects for some of the target population, but not for Medicare FFS primary care users in particular. This would be possible, for example, if the most effective aspects of the program targeted other parts of the target population, such as children, uninsured patients, managed care plan members, or frequent users of the ED without a primary care provider, among others. Finally, it is possible the program *was* effective for Medicare beneficiaries at Denver Health's FQHCs, but that we failed to detect these program effects due to poor statistical power. The power calculations in Table V.5 show that, if the program did have effects the size of our prespecified substantive thresholds for most outcomes, it is unlikely (that is, the probability is less than 50 percent) that we would have detected statistically significant effects. Among these three possibilities, we have little information to judge which is most likely.

Our evaluation has two main limitations.

1. Denver Health's HCIA-funded intervention affected the entire health system, but we were unable to identify a similar health system (or health systems) in a similar market to serve as

a comparison group. For this reason, we designed our evaluation to analyze effects only on people using Denver Health's FQHCs, not the rest of the target population, and we defined the comparison group to comprise beneficiaries at other urban FQHCs in Colorado. The comparison group was unmatched, and some differences existed between the treatment and comparison groups during an 18-month baseline period before the intervention began. It is possible these baseline differences influenced our impact results.

2. As described in detail previously, because our evaluation covers only Medicare FFS beneficiaries, (a) sample sizes are relatively small, limiting statistical power; and (b) results are not generalizable to the full population affected by the program.

Both challenges have implications for future tests of CMMI-funded programs. The tests in this report highlight the importance of constructing a credible comparison group for evaluation and having high quality data to assess impacts on a substantial portion of the intervention-affected population. The evaluation presented in this report has many strengths—including robust evidence on program implementation, survey data on staff perceptions of program effectiveness, and a difference-in-differences model to estimate program impacts that makes use of high quality claims data from both pre- and post-intervention periods. However, future evaluations could be even stronger if they can overcome limitations such as those described here. One possible solution to the lack of a comparison health system, for example, might be to randomize patients within the program population, so that some receive the new, supplemental program services (such as meetings with program navigators or appointments at the high-risk clinics) but others do not. This would allow valid estimates of the impact of program services, even without an external comparison group. In addition, for an awardee such as Denver Health, the problem of small sample sizes despite a large target population might be solvable with more timely, high quality Medicaid data—both for FFS and managed care—and with data from Medicare Advantage. Adding these populations to the evaluation would improve statistical power and the relevance of the impact estimates to the intervention overall.

REFERENCES

Boards of Trustees, Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds. "2013 Annual Report of the Boards of Trustees of the Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds." Table V.D1. Washington, DC: Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds, 2013. Available at http://downloads.cms.gov/files/TR2013.pdf. Accessed August 13, 2014.

Chronic Conditions Data Warehouse. "Table A.1.a. Medicare Beneficiary Counts for 2005–2014." Baltimore, MD: Centers for Medicare & Medicaid Services, 2016a. Available at https://www.ccwdata.org/web/guest/medicare-tables-reports. Accessed June 29, 2016.

Chronic Conditions Data Warehouse. "Table B.2.a Medicare Beneficiary Prevalence for Chronic Conditions for 2005 Through 2014." Baltimore, MD: Centers for Medicare & Medicaid Services, 2016b. Available at https://www.ccwdata.org/web/guest/medicare-tables-reports. Accessed June 29, 2016.

Gerhardt, Geoffrey, Alshadye Yemane, Keri Apostle, Allison Oelschlaeger, Eric Rollins, and Niall Brennan. "Evaluating Whether Changes in Utilization of Hospital Outpatient Services Contributed to Lower Medicare Readmission Rate." *Medicare & Medicaid Research Review,* vol. 4, no. 1, 2014, pp. E1–E13.

Health Indicators Warehouse. "Average Age of Medicare Beneficiaries (mean)." Hyattsville, MD: National Center for Health Statistics, HIW, 2014a. Available at http://www.healthindicators.gov/Indicators/Average-age-of-Medicare-beneficiaries-mean_308/Profile/ClassicData. Accessed November 19, 2014.

Health Indicators Warehouse. "Medicare Beneficiaries Who Are Also Eligible for Medicaid (percent)." Hyattsville, MD: National Center for Health Statistics, HIW, 2014b. Available at http://www.healthindicators.gov/Indicators/Medicare-beneficiaries-eligible-for-Medicaid-percent_317/Profile/ClassicData. Accessed August 8, 2016.

Health Indicators Warehouse. "Hospital Inpatient Medicare Admissions (per 1,000 beneficiaries)." Hyattsville, MD: National Center for Health Statistics, HIW, 2014c. Available at http://www.healthindicators.gov/Indicators/Hospital-inpatient-Medicare-admissions-per-1000-beneficiaries_2001/Profile/ClassicData. Accessed August 13, 2014.

Higgins, Tricia Collins, Laura Blue, Lauren Hula, Boyd Gilman, Greg Peterson, Catherine DesRoches, Sandi Nelson, Keith Kranker, Kate Stewart, Frank Yoon, and Lorenzo Moreno. "Evaluation of the Health Care Innovation Awards (HCIA): Primary Care Redesign Programs. Second Annual Report: Findings for Denver Health and Hospital Authority." Princeton, NJ: Mathematica Policy Research, September 30, 2015.

Institute of Education Sciences. "What Works Clearinghouse: Procedures and Standards Handbook, Version 3.0." Washington, DC: U.S. Department of Education, IES, 2014. Available at http://ies.ed.gov/ncee/wwc/DocumentSum.aspx?sid=19. Accessed September 15, 2014.

Johnson, Tracy L., Daniel Brewer, Raymond Estacio, Tara Vlasimsky, Michael J. Durfee, Kathy R. Thompson, Rachel M. Everhart, and Holly Batal. "Augmenting Predictive Modeling Tools with Clinical Insights for Care Coordination Program Design and Implementation." *eGEMs (Generating Evidence & Methods to Improve Patient Outcomes),* vol. 3, no. 1, article 14, 2015a.

Johnson, Tracy L., Deborah J. Rinehart, Josh Durfee, Daniel Brewer, Holly Batal, Joshua Blum, Carlos I. Oronce, Paul Melinkovich, and Patricia Gabow. "For Many Patients Who Use Large Amounts of Health Care Services, the Need Is Intense Yet Temporary." *Health Affairs,* vol. 34, no. 8, 2015b, pp. 1312–1319.

Laibson, David. "Impatience and Savings." NBER Reporter: research summary. Cambridge, MA: National Bureau of Economic Research, 2005. Available at http://www.nber.org/reporter/fall05/laibson.html. Accessed August 17, 2016.

Peikes, Deborah, Stacy Dale, Eric Lundquist, Janice Genevro, and David Myers. "Building the Evidence Base for the Medical Home: What Sample and Sample Size Do Studies Need? White Paper." AHRQ Publication No.11-0100-EF. Rockville, MD: Agency for Healthcare Research and Quality, October 2011.

Rosenthal, M.B., S. Alidina, M. Friedberg, S. Singer, D. Eastman, Z. Li, and E. Schneider. "A Difference-in-Differences Analysis of Changes in Quality, Utilization, and Cost Following the Colorado Multi-Payer Patient-Centered Medical Home Pilot." *Journal of General Internal Medicine,* 2016, vol. 31, no. 3, March 2016, pp. 289–296.

CHAPTER 4

FINGER LAKES HEALTH SYSTEMS AGENCY

Randall Blair, Rachel Shapiro, Rebecca Coughlin, Greg Peterson, Boyd Gilman, Laura Blue, Keith Kranker, Kate Stewart, Sheila Hoag, and Lorenzo Moreno

This page has been left blank for double-sided copying.

FINGER LAKES HEALTH SYSTEMS AGENCY

CHAPTER SUMMARY

Introduction. Finger Lakes Health Systems Agency (FLHSA) used its $26.6 million Health Care Innovation Award (HCIA) to transform primary care delivery in 68 practices in the greater Rochester, New York, area. The intervention's investments in transforming participating practices into patient-centered medical homes (PCMHs) targeted all patients served by these practices, and its intensive care management services targeted high-risk Medicare and Medicaid beneficiaries. FLHSA aimed to reduce the total cost of care by 3 percent by improving intermediate health outcomes and quality of care for all patients—particularly high-risk Medicare and Medicaid beneficiaries—thus reducing potentially preventable hospital admissions, hospital readmissions, and avoidable emergency department (ED) visits. FLHSA received the HCIA in July 2012 and began implementing the intervention with its first cohort of 19 practices in January 2013. A second cohort of 29 practices joined the intervention in July 2013, and a third cohort of 20 practices joined in July 2014. All practices remained in the intervention throughout the original three-year award period ending June 2015.

Objectives. This report aims to (1) describe the design and implementation of FLHSA's HCIA-funded intervention, including the role of primary care providers (PCPs) in the intervention and the extent to which anticipated changes in providers' behavior occurred; (2) assess impacts of the intervention on quality-of-care processes and outcomes, service use, and Medicare Part A and B spending during the first three years of the award; and (3) use both implementation and impact findings to identify possible explanations for the observed impacts.

Methods. We reviewed FLHSA's program documents and self-monitoring metrics, and conducted interviews with FLHSA leadership and program staff. In addition, we conducted two rounds of site visits to a select number of participating practices (four practices in each round), during which we interviewed clinicians, care managers, and other practice staff. We also administered two rounds of a survey to participating clinicians and one round of a survey to practice staff who received training from FLHSA with the HCIA funds. To estimate impacts, we compared outcomes for Medicare fee-for-service (FFS) patients served by 37 of the 68 participating practices with outcomes for Medicare FFS patients assigned to 108 matched comparison practices that did not participate in the HCIA program, adjusting for any differences in outcomes for the two groups during a one-year baseline period. These 37 practices enrolled in the first and second cohorts of the intervention (16 of the 19 practices in Cohort 1 and 21 of the 29 practices in Cohort 2). Because FLHSA targeted high-risk Medicare and Medicaid patients with intensive care management, we estimated the intervention's impact on high-risk Medicare FFS patients served by the practices—in addition to the intervention's impact on all Medicare FFS patients served by the practices—for all outcomes except those in the quality-of-care process domain. (Outcomes in the quality-of-care process domain were not estimated separately for high-risk and all Medicare FFS patients, as these outcomes were prespecified for patients who had a diagnosis of diabetes or vascular disease, or at least one hospital stay in the past three months.)

Program design and implementation. The intervention had three components: (1) a practice transformation component to redesign primary care processes, culture, and workforce to transform 68 participating practices into PCMHs; (2) an intensive care management component for high-risk patients; and (3) a community-wide outcomes-based payment model component to ensure sustainability of program activities after the HCIA period ends.

FLHSA implemented the practice transformation and care management components of the intervention largely as planned. FLHSA's 68 enrolled practices exceeded its initial goal of 65, and all enrolled practices successfully hired care managers. FLHSA practice improvement advisors delivered services to each practice through weekly or biweekly meetings to identify and work on quality improvement projects. FLHSA clinical advisors coached and mentored practice-based care managers in regularly scheduled meetings to integrate the care manager into the care team at the practice. In addition, FLHSA held monthly learning collaboratives to facilitate learning across practice champions (primary care physicians from each practice who served as the main points of contact with FLHSA program staff) and care managers. After participating in the intervention, practices reported an increased use of electronic health records (EHRs) to generate population-based and patient-specific reports. The practices also implemented weekly care team huddles and monthly care team meetings. In addition, care managers reported that about half of patients showed improved levels of activation. Findings from the trainee survey of care managers and practice champions were largely positive. For example, care managers reported spending time as expected based on the program design. Still, despite a program goal to link high-risk patients with community services, only 57 percent of surveyed care managers reported that they routinely helped patients access nonmedical services.

The third component, working with two local insurers to implement a community-based payment model, was not implemented as planned. Two commercial insurers developed communitywide outcomes-based payment models by the end of the award period, although the development of the models took longer than expected. Although all practices have the option of participating in these payment models, recent changes in the market unrelated to the program will likely mean that few will choose to do so. Over the intervention period—and unanticipated by FLHSA at the time of its application for HCIA funding—two regional accountable care organizations (ACOs) formed, and most practices participating in the HCIA-funded intervention joined one of these ACOs. The ACOs will provide member practices with the ability to sustain practice transformation activities and, to some extent, care management activities. Therefore, these practices no longer have a need for the communitywide payment models. Practices that have not joined ACOs can still enroll in the commercial models to help them sustain the practice transformation and care management activities.

Clinicians' perceptions of intervention effects on the care they provided to patients. FLHSA's program design required PCPs to actively integrate care managers into practice care teams and implement care team huddles. The available evidence suggests that FLHSA engaged PCPs as planned, with most of the surveyed PCPs reporting that the intervention improved the quality and patient-centeredness of care at their practices. In the second round of the clinician survey, slightly more than half of the surveyed clinicians felt the intervention improved the efficiency and timeliness of care. However, a large portion of respondents (more than 25 percent)

reported that the program had no effect on efficiency or timeliness of care, and more than one-third felt the program had no effect on safety or equity of care or the information available for clinical decision making.

Impacts on patients' outcomes. The impact estimates indicate that, during the original three-year award period, the intervention improved Medicare FFS patients' outcomes in the quality-of-care processes domain, did not improve outcomes in the service use domain (either for high-risk or all Medicare FFS patients), and had an indeterminate effect on the quality-of-care outcomes and spending domains (for both high-risk and all Medicare FFS patients). Specifically, there was evidence of a statistically significant favorable effect in the quality-of-care process domain, driven by a 5 percent impact for inpatient admissions followed by an ambulatory care visit with a primary care or specialist provider within 14 days. The favorable impacts were modest in size, however (they were smaller than the prespecified threshold for substantively large effects). There were no statistically significant or substantively large effects in the other three domains. Because the statistical power to detect effects was good for the service use domain, these findings likely mean the program did not have substantively large effects on service use (outpatient ED visits and inpatient admissions). The evaluation was not well powered for outcomes in the quality-of-care outcome domain (inpatient admissions for ambulatory care-sensitive conditions and 30-day unplanned hospital readmissions) or for spending, so the lack of measured effects might be because the program truly did not have effects or it did but our test failed to detect them. It is unclear whether the estimates for Medicare FFS beneficiaries would generalize to Medicare Advantage and Medicaid beneficiaries who are in FLHSA's target population but, due to data availability, were not in the evaluations' treatment group.

Conclusion. Evaluation evidence indicates that, for Medicare FFS beneficiaries, FLHSA improved quality-of-care process measures by a modest amount but did not reduce service use during the original three-year award period. The lack of effects does not appear to be a result of a failure to engage PCPs or implement the program as planned. Rather, the lack of effects might be a result of (1) unforeseen implementation barriers, including limited staff time to devote to transformation activities and care management; (2) overly ambitious goals given the relatively small portion of patients receiving intensive services; (3) the relatively short intervention duration covered in this impact analysis; and (4) limited room for improvement among some practices with respect to care management and PCMHs. Impact estimates might change after including the final 12 months of program operations (July 2015 to June 2016), the period when FLHSA expected to observe the largest impacts. We will report final evaluation results in an addendum to this report.

Summary of intervention and impact results for FLHSA

Intervention description			
Awardee description		Community health planning and convening organization in Rochester, New York	
Award amount ($ millions)		$26.6 million	
Award extended beyond June 2015?		Yes (12 months)	
Location		6 counties in greater Rochester area[a] (urban, suburban, and rural)	
Target population		All patients served by 68 primary care practices, which enrolled in the intervention in three cohorts	
Interventions		Identified care gaps among the full patient population at participating practices and developed care plans for high-risk patients • 5 HCIA-funded practice improvement advisors helped practices improve team communication, use EHRs to identify care gaps, and streamline workflows • PCPs were each paid $20,000 to participate in the intervention • 70 care managers hired to (1) coach high-needs patients on self-management, (2) coordinate care with providers, and (3) connect patients with social services	
Metrics of intervention delivered		• Weekly huddles at all practices by June 2015 • Care managers hired at all practices • Care manager services provided to 17,484 patients	
Impact evaluation methods			
Core design		Difference-in-differences model with matched comparison group	
Treatment group	Definition	Medicare FFS beneficiaries attributed to 37 practices FLHSA enrolled by July 1, 2013[b]	
	# of beneficiaries during primary test period[c]	9,271 to 15,638	
Comparison group definition		Medicare FFS beneficiaries attributed to 108 matched comparison practices	
Impact results: Quality-of-care processes domain			
Ambulatory care visit within 14 days of discharge (% of beneficiaries/quarter)		Comparison mean[d]	67.6%
		Impact estimate (% difference)	+3.1 pp (+4.6%)*
Received recommended lipid test, for patients with IVD (% of beneficiaries/year)		Comparison mean[d]	76.4
		Impact estimate (% difference)	-0.6 pp (-0.7%)
Received an HbA1c test, for patients with diabetes (% of beneficiaries/year)[e]		Comparison mean[d]	88.6%
		Impact estimate (% difference)	+0.9 pp (+1.0%)
Received a complete lipid profile, for patients with diabetes (% of beneficiaries/year)[e]		Comparison mean[d]	80.2%
		Impact estimate (% difference)	+2.1 pp (+2.6%)
Combined impact estimate[f]		+1.9%**	
Impact conclusion[g]		**Statistically significant favorable effect**	
Impact results: Quality of care outcomes domain			
30-day unplanned hospital readmissions (#/1,000 beneficiaries/quarter)		Comparison mean[d]	14.3
		Impact estimate (% difference)	+0.1 (0.7%)
Inpatient admissions for ACSC conditions (#/1,000 beneficiaries/quarter)		Comparison mean[d]	16.0
		Impact estimate (% difference)	+0.3 (+1.6%)
Combined impact estimate[f]		+3.7%[h]	
Impact conclusion[g]		Indeterminate effect	
Impact results: Service use domain			
All-cause inpatient admissions (#/1,000 beneficiaries /quarter)		Comparison mean[d]	83
		Impact estimate (% difference)	+3.1 (+3.7%)
Outpatient ED visits (#/1,000 beneficiaries/quarter)		Comparison mean[d]	173.3
		Impact estimate (% difference)	-3.5 (-2.0%)
Combined impact estimate[f]		+0.6%[i]	
Impact conclusion[g]		No substantively large effect	
Impact results: Spending domain			
Medicare Part A and B spending ($/beneficiary/month)		Comparison mean[d]	$825
		Impact estimate (% difference)	+$11 (+1.3%)
Combined impact conclusion[f]		0.8%[j]	
Impact conclusion[g]		Indeterminate effect	

Note: See this chapter for details on the intervention, impact methods, and impact results.

Summary of intervention and impact results for FLHSA *(continued)*

[a] Livingston, Monroe, Ontario, Seneca, Wayne, and Yates.

[b] Our impact evaluation covers 37 practices that enrolled in the intervention in the first two cohorts of participating practices. We excluded Cohort 3 practices because they joined late in the award period and neither we nor the awardee expected the program to affect patients' outcomes during the original 3-year award period. We will include Cohort 3 practices in our future final impact analyses.

[c] For some outcome measures the sample is limited to a relevant subset of beneficiaries.

[d] The comparison mean is the estimate of the outcome the treatment group beneficiaries would have had if they had not received the intervention. It is equal to the mean for the treatment group over the intervention quarters (in the primary test period) minus the impact estimate.

[e] Unlike the estimates for other awardees, we did not estimate impacts on receipt of all four recommended diabetes processes of care because FLHSA did not target all of these measures. Instead, we focused on the two processes FLHSA did target: HbA1c tests and lipid profiles.

[f] The combined estimate is the average across all the individual estimates in each domain, in which the impact estimate for each individual outcome is expressed as percentage change relative to the comparison group.

[g] We drew conclusions at the domain level based on the results of prespecified primary tests, secondary tests (robustness checks), and consistency with implementation evidence. For each domain, we could draw one of five conclusions: (1) Statistically significant favorable effect (the highest level of evidence), (2) Substantively important (but not statistically significant) favorable effect, (3) Substantively important (but not statistically significant) unfavorable effect, (4), No substantively large effect, and (5) Indeterminate effect. Section V.A.8 of this report describes the decision rules we used to reach each of these possible conclusions.

[h] FLHSA's combined impact estimate for the quality-of-care outcomes domain comprises the estimates of two measures in this table (30-day unplanned readmissions and ACSC admissions among all beneficiaries) and two measures not reported in this table (30-day unplanned readmissions and ACSC admissions among only high-risk beneficiaries) but that are reported in the full chapter for FLHSA.

[i] FLHSA's combined impact estimate for the service use domain comprises the estimates of two measures in this table (all-cause inpatient admissions and outpatient ED visits among all beneficiaries) and two measures not reported in this table (all-cause inpatient admissions and outpatient ED visits among only high-risk beneficiaries) but that are reported in the full chapter for FLHSA.

[j] FLHSA's combined impact estimate for the spending domain comprises the estimates of one measure in this table (Medicare Part A and B spending among all beneficiaries) and one measure not reported in this table (Medicare Part A and B spending among only high-risk beneficiaries) but that is reported in the full chapter for FLHSA.

*Significantly different from zero at the .10 level, one-tailed test.

**Significantly different from zero at the .05 level, one-tailed test.

***Significantly different from zero at the .01 level, one-tailed test.

ACSC = ambulatory care-sensitive condition; ED = emergency department; EHR = Electronic Health Records; FFS = fee-for-service; FLHSA = Finger Lakes Health Systems Agency; HCIA = Health Care Innovation Award; IVD = ischemic vascular disease; NA = not available; n.a. = not applicable; PCP = primary care provider; pp = percentage point.

This page has been left blank for double-sided copying.

I. INTRODUCTION

This report presents findings from the evaluation of the Health Care Innovation Award (HCIA) received by the Finger Lakes Health Systems Agency (FLHSA), with a focus on program impacts on patients' outcomes. Section II provides an overview of FLHSA's HCIA-funded intervention and the design of the impact evaluation. Section III describes the design and implementation of the intervention, including how the program could be expected to affect evaluation outcomes through changes in patients' and providers' behavior. In Section IV, we assess the evidence on the extent to which planned changes in providers' behavior occurred. Section V describes our methods for, and results and conclusions from, estimating program impacts on patients' outcomes in four domains: quality-of-care processes, quality-of-care outcomes, service use, and spending. Section VI draws conclusions by synthesizing the impact and implementation findings and describes the next steps for the evaluation.

The impact estimates in this report are preliminary because they include only two of the three cohorts of practices included in the intervention and cover only the original three-year award period of the HCIA (July 2012 through June 2015). Because FLHSA's HCIA-funded program extended through June 2016, we do not yet include the final 12 months of FLHSA's intervention in the impact analysis. We plan to report final results, including these 12 months, in an addendum to this report. Final results will also include Cohort 3 practices in the impact analysis; these practices began the HCIA program in July 2014, and would not be expected to generate program impacts by June 2015.

II. OVERVIEW OF FLHSA'S HCIA-FUNDED INTERVENTION AND IMPACT EVALUATION

A. FLHSA's HCIA-funded intervention

FLHSA, a community health planning organization and convening agency in Rochester, New York, received $26.6 million in HCIA funding to implement an initiative to transform primary care processes and delivery in 68 practices in six counties in the greater Rochester area (Table II.1, top panel). FLHSA recruited practices in three cohorts and selected practices that (1) served a large number of Medicare and Medicaid patients relative to other practices in the region; (2) had used electronic health records (EHRs) for at least six months; (3) had four to seven full-time equivalent (FTE) physicians (Cohort 1) or two to seven FTE physicians, nurse practitioners, or physician assistants (Cohorts 2 and 3); and (4) demonstrated a sufficient level of readiness to participate in the program. The intervention's target population was all patients served by these practices, with intensive care management services targeting high-risk patients. HCIA-funded services began on January 1, 2013, as planned, and extended 12 months beyond the original award end date (June 30, 2015) to end on June 30, 2016.

FLHSA's goals were to reduce the total cost of care by 3 percent by reducing potentially preventable hospital admissions by 25 percent, reducing 30-day readmissions by 25 percent, and reducing avoidable emergency department (ED) visits by 15 percent by the end of the award (Table II.1). FLHSA expected to achieve these outcomes through three intervention components: (1) a practice transformation component to help 68 participating practices become patient-

centered medical homes (PCMHs), (2) an intensive care management component for high-risk Medicare and Medicaid patients, and (3) a communitywide outcomes-based payment model component. FLHSA expected that these intervention components would increase the quality of care and patients' access to care, and increase their activation and self-management, thus reducing potentially preventable inpatient admissions, hospital readmissions, and avoidable outpatient ED visits. These reductions would, in turn, reduce total Medicare and Medicaid spending. (Section III.A.3 describes the awardee's theory of action in detail.)

Table II.1. Summary of FLHSA's HCIA program and our evaluation for estimating its impacts on patients' outcomes

Program description	
Award amount	$26,584,892
Award start date	July 1, 2012
Implementation date	Cohort 1: January 1, 2013 Cohort 2: July 1, 2013 Cohort 3: July 1, 2014
Award end date	June 30, 2016
Awardee description	FLHSA is a community health planning organization and convening agency in Rochester, New York, that serves nine counties in the greater Rochester area: Chemung, Livingston, Monroe, Ontario, Schuyler, Seneca, Steuben, Wayne, and Yates.
Intervention overview	FLHSA created an initiative to transform primary care processes and delivery in 68 practices in greater Rochester.
Intervention components	1. **Practice transformation.** FLHSA practice improvement advisors worked with practice champions and other practice staff to redesign primary care processes, culture, and workforce to transform 68 practices (recruited in three separate cohorts) into PCMHs. Practice improvement advisors held weekly or biweekly meetings with practice staff to identify and work on quality improvement projects to help staff transform their practices. Each project incorporated team-based care and quality improvement concepts. Practice transformation projects occurred at the practice level and sought to affect the overall practice operations. 2. **Care management.** The care management component focused on providing intensive care management services to high-risk patients. FLHSA clinical advisors helped participating practices to train and deploy practice-based care managers to provide intensive care management and link patients with community resources. Clinical advisors coached and mentored practice-based care managers in regularly scheduled meetings. Care managers, who were fully funded by the HCIA for the first two years of practice participation, screened practice populations to identify high-risk patients who qualified for intensive care management services by using a screening tool, reviewing practice population data and medical records, receiving a provider's recommendation, and through patients' self-referral. After patients were identified, care managers reached out to patients to invite them to participate. Care managers started with 20 to 40 intensive patients, eventually building up to a caseload of 40 to 60 patients, and contacted these patients at least monthly. Care managers worked with these patients until they felt patients would no longer benefit from their services or patients decided they no longer needed care management. 3. **Communitywide outcomes-based payment model.** FLHSA leadership worked with two insurers to develop a communitywide outcomes-based payment model to ensure sustainability of program activities and personnel after the HCIA period.
Target population	All patients served by 68 practices in six counties in the greater Rochester area, with more intensive services provided to high-risk patients. High-risk patients were generally defined as those with multiple chronic conditions and frequent hospitalizations, but the definition varied by practice.

Table II.1 *(continued)*

Target impacts on patient outcomes	• Reduce cost of care by 3 percent • Reduce potentially preventable hospital admissions by 25 percent • Reduce 30-day hospital readmissions by 25 percent • Reduce avoidable ED visits by 15 percent
Workforce development	The award fully funded 87 positions. Under the practice transformation component, FLHSA hired 1 practice improvement coordinator and 5 practice improvement advisors. Under the care management component (1) FLHSA hired 4 clinical advisors and 1 social work clinical coordinator; and (2) using HCIA funds allocated by FLHSA, the participating practices hired 70 practice-based care managers and, before December 2014, Trillium Health (a health services organization partnering with FLHSA) provided 6 practices with 6 community health workers.
Location	Urban (Rochester), suburban (Webster), and rural. The intervention operates in six counties in the Finger Lakes region of New York State: Livingston, Monroe, Ontario, Seneca, Wayne, and Yates. These counties include Rochester and Webster.
Impact evaluation	
Core design	Difference-in-differences with matched comparison group
Treatment group	Medicare FFS beneficiaries assigned to 37 of the 48 practices that joined the HCIA program in the program's two cohorts
Comparison group	Medicare FFS beneficiaries assigned to 108 comparison practices that did not participate in the HCIA program
Intervention component(s) included in impact evaluation	The first two components described earlier: practice transformation and care management. The impact estimates will capture the joint effects of both components. The communitywide outcomes-based payment model component is not included in the impact evaluation because it was not implemented during the follow-up period.
Extent to which the treatment group reflects the awardee's target population (for the component(s) evaluated)	Low: FLHSA's target population for all the components assessed in the evaluation includes Medicaid, Medicare managed care, and/or privately insured beneficiaries. Medicare FFS beneficiaries account for less than half of the target population.[a]
Study outcomes, by domain	1. **Quality-of-care processes** - LDL testing for patients with diabetes - A1c testing for patients with diabetes - LDL testing for patients with IVD - 14-day follow-up after hospitalization 2. **Quality-of-care outcomes** - 30-day unplanned readmissions - Inpatient admissions for ambulatory care-sensitive conditions 3. **Service use** - All-cause inpatient admissions - Outpatient ED visits 4. **Spending** - Medicare Part A and B spending - Medicare inpatient spending

Source: Review of FLHSA reports, including its original application, operational plan, and 14 quarterly narrative reports to CMS.

[a] This estimate is based on managed care, commercial patient, and Medicare FFS beneficiary counts for treatment practices self-reported by FLHSA in quarterly measurement and monitoring results.

CMS = Centers for Medicare & Medicaid Services; ED = emergency department; FFS = fee-for-service; FLHSA = Finger Lakes Health Systems Agency; HCIA = Health Care Innovation Award; IVD = ischemic vascular disease; LDL = low-density lipoprotein; PCMH = patient-centered medical home.

B. Overview of impact evaluation

To estimate impacts, we compared outcomes for Medicare fee-for-service (FFS) patients served by 37 of the 48 practices participating in either of the first two cohorts of the HCIA-funded intervention (the treatment group) with outcomes for Medicare patients served by 108 matched comparison practices, adjusting for any differences in outcomes between these two groups before the intervention began. We excluded 6 of the 48 participating practices because they were federally qualified health centers (FQHCs) and had no suitable comparison. We excluded another 2 practices because they served psychiatric or pediatric populations, and 3 practices because they had had no attributed Medicare patients in at least one quarter of the evaluation baseline period (January to December 2012 for Cohort 1 practices and July 2012 to June 2013 for Cohort 2 practices). Because FLHSA targeted high-risk Medicare and Medicaid patients with intensive care management, we estimated the intervention's impact on high-risk Medicare FFS patients served by the practices, in addition to the intervention's impact on all Medicare FFS patients served by the practices. Table II.1, bottom panel, summarizes our impact evaluation design.

We selected the 108 comparison practices for the evaluation from the pool of all primary care practices in New York State that served Medicare FFS beneficiaries and were located outside of the 6 counties in which the HCIA-funded intervention took place, outside the New York City metropolitan area, and outside the 13 counties in New York State with relatively high participation rates in two federal primary care initiatives: the Multi-Payer Advanced Primary Care Practice (MAPCP) Demonstration and the Comprehensive Primary Care (CPC) initiative. We selected comparison practices that were similar to the 37 treatment practices in terms of their practice characteristics and the characteristics of their Medicare patients before the intervention began.

We estimated impacts on outcomes, as measured in Medicare FFS claims data grouped into four domains: (1) quality-of-care processes, (2) quality-of-care outcomes, (3) service use, and (4) spending. Across the HCIA awardees in primary care redesign (PCR), we designed our impact evaluations to identify promising interventions or intervention components—consistent with evaluation goals from the Center for Medicare & Medicaid Innovation (CMMI) to find programs that could be scaled or retested in the future. Before conducting analyses, we specified a series of primary tests, describing the evidence we would need to conclude that the program was effective, and FLHSA and CMMI reviewed these tests. Each test specified a population, outcome, time period, expected direction of effect, and threshold that we count as substantively important. The purpose of these primary tests was to focus the impact evaluation on hypotheses that would provide the most robust evidence about program effectiveness. We used the results from the primary tests and robustness checks to draw conclusions about program impacts in each of the four evaluation domains. Because we sought to identify promise, rather than only those programs with unequivocally demonstrated success, we conducted one-sided statistical tests (that is, testing only for program benefits) and used a threshold for statistical significance of 0.1, which is not as strict as the conventional standard of 0.05.

Our impact evaluation design reflects the effects of two of the three components that constituted FLHSA's HCIA-funded intervention for all attributed Medicare patients: the practice

transformation and the intensive care management components. It does not capture the effects of the communitywide outcomes-based payment model of the intervention because the component was not implemented during the impact analysis follow-up period (from January 2013 to June 2015). The evaluation's treatment group includes all Medicare FFS beneficiaries attributed to the 37 treatment practices—or all Medicare FFS beneficiaries who received the plurality of their primary care services from physicians, nurses, and physician assistants affiliated with treatment practices. FLHSA expected the two intervention components—practice transformation and care management—to combine to affect outcomes for all patients served by treatment practices, even though the practices provided care management services only to high-risk patients. We used the same decision rule that CMMI uses for the CPC initiative to attribute Medicare beneficiaries to treatment and comparison panels.

The treatment group for the impact evaluation—which is limited to Medicare FFS beneficiaries—accounts for about 20 percent of FLHSA's total target population. The other 80 percent of patients have Medicare Advantage, Medicaid, commercial insurance, another form or insurance, or no insurance. It is unclear whether the estimates for Medicare FFS beneficiaries would generalize to other patients within FLHSA's target population.

III. PROGRAM IMPLEMENTATION

This section first provides a detailed description of FLHSA's HCIA-funded intervention, highlighting how it evolved over time and its theory of action. Second, it assesses the evidence on the extent to which the intervention was implemented as planned based on measures of program enrollment, service delivery, staffing, training, and timeliness. Third, the section summarizes the facilitators and barriers associated with implementation effectiveness.

We based our evaluation of FLHSA's program implementation on a review of its quarterly reports to CMMI and self-monitoring program metrics, telephone discussions and follow-up communications with program administrators, and information collected during site visit interviews with frontline staff at selected practices conducted in April 2014 and April 2015. We selected eight practices to visit (four during each year) that represented the three cohorts; were a range of system-owned and independent practices; and were in urban, suburban, and rural areas. We did not verify the quality of the performance data reported by FLHSA in its self-measurement and monitoring reports.

A. Program design and adaptation

1. Target population and patient identification, recruitment, and enrollment

In this section, we describe how FLHSA selected practices to participate in the HCIA-funded intervention, identified the patients the practices serve, and identified high-risk Medicare and Medicaid patients for intensive care management services.

Identification of practices for participation. To recruit practices in each of the three cohorts, FLHSA conducted outreach to practices in the target area, solicited calls for applications, and assessed practices on the following four criteria:

1. High proportion (relative to other practices in the region) of Medicare and adult Medicaid patients who received care at the practice in the two years before the application and were at risk for potentially avoidable hospitalizations, hospital readmissions, and preventable ED use
2. Use of EHR for at least six months
3. Number of PCPs:
   - Four to seven FTE physicians (Cohort 1), because FLHSA determined that this practice size would best support one full-time care manager
   - Two to seven FTE physicians, nurse practitioners, or physician assistants (Cohorts 2 and 3), because FLHSA recognized that many practices in the six-county region relied on nurse practitioners or physician assistants to carry out primary care activities
4. Award readiness, as indicated by interviews with practice staff to assess:
   - Leadership, access, teamwork, and data/clinical information systems (Cohort 1)
   - Practice's stated willingness to participate in award activities (Cohorts 2 and 3) (although used to select Cohort 1 practices, FLHSA staff discontinued use of a readiness assessment tool because they felt that this method of scoring did not accurately predict practices' level of preparation for program implementation)

Of the 92 practices that applied to participate in the intervention, FLHSA selected 68 to participate (19 of 37 applicants in Cohort 1, all 29 applicants in Cohort 2, and 20 of 26 applicants in Cohort 3). This exceeded FLHSA's initial target of 65 practices. The participating practices were located across the six counties served by FLHSA, and varied in structure and affiliation—they were either private practices, FQHCs, or part of a larger health system. Practices also varied in terms of the characteristics of their patient populations, such as race and ethnicity, age, comorbid conditions, and coverage source.

Target patient population. The target population for FLHSA's HCIA-funded intervention was all patients served by the 68 participating practices, with intensive care management services provided to high-risk patients (generally defined as Medicare and Medicaid patients with multiple chronic conditions and frequent hospitalizations, although the specific definition varied by practice). FLHSA expected practice transformation to affect all patients at the practices and care management to affect only the subset of patients classified as high risk.

Identification, recruitment, and enrollment of patients for care management. Care managers screened practice populations to identify high-risk patients who qualified to receive intensive care management services. They accomplished this in several ways: (1) screening practice populations using a screening tool (such as the LACE Index Scoring Tool for Risk Assessment of Hospital Readmissions [Van Walraven et al. 2010] or other similar tools); (2) reviewing practice population data to identify patients with chronic conditions (for example, diabetes or chronic obstructive pulmonary disease) or patients with out-of-range lab results (for example, high hemoglobin A1c levels); (3) reviewing medical records to find patients with recent hospitalizations or ED visits; (4) receiving a provider's recommendation; and (5) through

patients' self-referral. The methods care managers used to identify patients varied based on practice characteristics, such as the capabilities of the practice's EHR or the practice's affiliation with a larger health system. After care managers identified patients as high risk, they contacted patients and invited them to participate in intensive care management and enrolled those who were interested.

2. Intervention components

As noted earlier, FLHSA's intervention had three components—practice transformation, care management, and a communitywide outcomes-based payment model. In this section, we describe the design and implementation of each component.

Practice transformation. FLHSA practice improvement advisors worked with participating practices to redesign their primary care processes, culture, and workforce. Practice improvement advisors, who are experts in quality improvement processes to implement system change, met with practice champions (primary care physicians from each practice who served as the main points of contact with FLHSA program staff and led the practice transformation component at their practices) and other staff in weekly or biweekly meetings to identify and work on quality improvement projects. For each project, practice improvement advisors incorporated team-based care and recognized process improvement concepts. They also helped practices collect and use data to identify areas for practice improvement and test new ideas using a Plan-Do-Study-Act (PDSA) model, which many practices had not previously used. The PDSA model was originally developed as a method to implement change in complex systems and is frequently used for quality improvement projects in health. PDSA focuses on using an iterative process to develop and test new interventions (Taylor et al. 2014).

In their regularly scheduled meetings, FLHSA practice improvement advisors worked with practice staff to do the following:

- Establish communication pathways among practice staff, specifically among practice care teams, through weekly huddles (meetings during which the care team reviews patients' information to prepare for upcoming appointments) and monthly care team meetings (scheduled meetings of clinicians, nurses, and support staff to discuss specific patients' care plans)
- Improve practice staff's use of EHRs to improve care processes, helping them to use EHRs to generate population- and patient-based reports and identify gaps in care in immunizations and chronic care screening
- Improve workflows through the use of process mapping and cycle time analysis (for example, determining the length of waiting times and patients' visits to determine the most efficient use of clinicians' and patients' time)
- Conduct additional quality improvement projects identified by practice staff, using the PDSA model to identify and develop projects, test change, and document new processes

In addition to these meetings with staff at individual practices, FLHSA practice improvement advisors organized monthly, in-person learning collaboratives to support practice champions and facilitate learning across practices. These learning collaboratives provided practice champions with opportunities to discuss successes of, barriers to, and potential solutions for quality improvement projects. FLHSA staff also used learning collaboratives as forums to provide additional trainings related to practice transformation for practice champions.

FLHSA provided direct financial support for the practice transformation component. It paid each participating practice an incentive payment of $20,000 per year for each primary care physician affiliated with the practice to compensate the practices for the additional work participating in the intervention required.

Care management. For the first two years of practice participation, FLHSA paid the salaries for practice-based care managers hired as part of the intervention. FLHSA clinical advisors helped participating practices hire, train, and deploy care managers to provide intensive care management and link patients with community resources. The goal was to improve service delivery to high-risk patients with complex care needs (for example, those with chronic obstructive pulmonary disease, congestive heart failure, or diabetes). FLHSA clinical advisors coached and mentored practice-based care managers in regularly scheduled meetings to integrate the care manager into the practice's care team. Initially, clinical advisors worked with care managers to establish care teams and regular huddles at each practice, among practices that did not already hold huddles, as well as regular care team meetings. At the initial and subsequent meetings, clinical advisors discussed with care managers how to identify patients for care management, build and maintain a panel of intensively managed patients, and conduct activities to support population management. FLHSA clinical advisors also provided targeted technical support to care managers (for example, to help care managers report on clinical quality measures through their EHRs) and organized learning collaboratives to support and facilitate learning across care managers.

In addition, from January 2013 to December 2014, FLHSA partnered with Trillium Health, a neighborhood health center and health services organization, to integrate community health workers (CHWs) into six of the practices, all of which were FQHCs and had large proportions of high-risk patients. CHWs educated practice staff on the needs of the local community and helped care managers link patients with community resources. However, poor communication among FLHSA, Trillium, and the practices resulted in confusion over how these CHWs were to function at the practices and how their responsibilities differed from those of the care managers. Ultimately, all six CHWs left the practices to which they were assigned, and FLHSA decided to discontinue its work with Trillium; the six practices did not hire new CHWs. We excluded FQHCs from the impact analysis, so the brief integration of CHWs into these practices does not affect our analysis of program impacts on patients' outcomes.

Communitywide outcomes-based payment model. FLHSA leadership worked with two insurers, Excellus Blue Cross Blue Shield and MVP, to develop a communitywide outcomes-based payment model to ensure sustainability of program activities and personnel after the HCIA period. FLHSA leadership expected that the combined shared savings payments to practices

would cover continuing practice transformation costs and the cost of employing a care manager. However, after program implementation began, many practices joined one of two regional accountable care organization (ACOs), and will sustain program activities through their ACOs instead of through the Excellus and MVP payment models. Practices that are not part of an ACO still expect to receive shared savings for performance on specific quality and outcome measures through the payment models offered by Excellus and MVP.

3. Theory of action

Based on extensive review of FLHSA's program activities and goals, we developed a theory of action to depict the mechanisms through which program administrators expected the program to improve the outcomes we selected for the impact evaluation (Table II.1 lists these outcomes). FLHSA expected that its HCIA-funded intervention would improve outcomes through two pathways.

First primary pathway to improved outcomes. Practices transform the way they provide care, thereby improving quality of care and reducing service use and spending among all patients. Planned mechanisms of this pathway include the following:

1. **FLHSA practice improvement advisors provide targeted assistance to practice champions and other staff from each practice following a project-based and learning collaborative model to help transform primary care processes.** These projects will improve practice staff communications and use of EHRs to improve care processes and workflows.
2. **Primary care practice staff redesign many of their primary care processes.** Practice staff have new and redefined roles and responsibilities and are team-based, coordinated, and provide patient-centered care. Clinicians, nurses, care managers, and front- and back-office staff work in care teams and participate in regular weekly huddles and monthly care team meetings. Practices integrate the use of EHRs so they can use them to generate population- and patient-based reports to flag patients who are due for immunizations or tests and identify other gaps in care. Care managers follow up on identified care gaps by informing care teams during huddles before patients' visits, or by contacting patients to schedule appointments. Either the National Committee for Quality Assurance (NCQA) recognizes practices as PCMHs or the practices are closer to gaining this recognition; however, formal recognition is not an explicit goal of the project. Practice staff continue to work with FLHSA practice improvement advisors on projects, using the PDSA model to facilitate changes to incrementally improve practice operations associated with becoming a PCMH, focusing on components that will improve care quality. Practice staff regularly use data from these PDSA projects and from the EHR reports to identify and implement practice improvements.
3. **Physicians, nurse practitioners, physician assistants, and nurses have improved communication about patients through the weekly huddles and monthly care team meetings, leading to more efficient and focused visits for patients.** Patients leave visits with more of their questions answered and a better understanding of their health and care needs. Patients have more timely screenings and appointments.

4. **The practice transformation process improves quality of care and increases patients' access to care.** Patient-centered care improves clinical care by increasing screenings and other preventive care measures, thereby proactively identifying and treating medical issues. This in turn reduces inpatient admissions for ambulatory care-sensitive conditions (ACSCs) and all-cause inpatient admissions, which reduces total Medicare Part A and B spending.

Text box III.1. Example from FLHSA illustrating the program's theory of action for practice transformation

"Here's an example of something that came out of CMMI. The patient checks out and they're referred for a mammogram. They say, 'Ok, I'll give them a call, I'll schedule it.' Then the patient leaves [without making an appointment for a mammogram]. So what's going on? We talked about it as a CMMI team on a phone conference and we said, 'Let's look at this quality measure in particular. What can we do differently?' And that's why having an integrated team on the CMMI team can help. I said, 'Why does the patient leave without the appointment, what's the reason we don't make the appointment?' Because I know it's not because they don't want to make the appointment; there's got to be something going on. 'Well, when you call to make appointment,' the access associate tells me, 'They [radiology] ask key questions, like where was your last mammogram, have you ever had a lump or mass, have you ever had a mastectomy.' There's four or five questions that the access associate doesn't know. That's not a fun conversation to have at checkout, and there could be people behind them, it's awkward. You don't want to be asked that at checkout. So what can we do to retain the appointment before the patient leaves, because that's the key. So, we made a laminate with key questions. There's four or five key questions that we know they're going to ask us when we book the appointment. So, we hand them the wipe marker, and we say 'Oh, I see that Dr. [X] scheduled you for a mammogram. I can schedule that for you, could you please answer these few questions?' Then they get radiology on the phone [and] they make the appointment. Now, what was our outcome? We went back and we did a tally of all the patients that we checked out and made their appointment, and the patients that checked out and chose to make their own appointment. The data was unbelievable. One hundred percent of the patients that had their appointment secured before they left followed through and completed the visit with mammograms. Every patient in this sample that left [without scheduling an appointment] never followed through. That's an example of something that our CMMI team did as a group to say 'Where do we go?" and track the outcomes to see if it worked or did it not work and do we need to do something differently. Then Dr. [X] took that to his quarterly provider group [at the practice network].… So we got that out to the other practices, we sent them all the supplies, so what happened here we also shared with other practices."

Source: Interview with practice manager, April 2015 site visit.

Second primary pathway to improved outcomes. Practice-based care managers provide intensive care management services to high-risk adult patients, thereby improving quality of care and reducing service use and spending among them. Planned mechanisms of this pathway include the following:

1. **FLHSA helps practices integrate care managers into their care teams to support care coordination.** FLHSA clinical advisors coach and mentor practice-based care managers in regularly scheduled meetings to integrate the care manager into the care team at the practice. FLHSA clinical advisors also provide training (such as in motivational interviewing) and targeted technical support to care managers (for example, helping them identify high-risk patients and report on clinical quality measures through their EHRs) and organize learning collaboratives to support and facilitate learning across care managers.

2. **Care managers identify high-risk patients who qualify for intensive care management services.** As described in Section III.A.1, care managers accomplish this through one or more of the following methods: using a tool to screen practice populations, reviewing practice population data and medical records, receiving a provider's recommendation, and patients' self-referral. Each practice defines *high risk* slightly differently, depending on its patient population and system affiliation.
3. **Care managers reach out to high-risk patients to explain care management and invite them to participate.** If they agree, care managers obtain patients' consent to receive care management.
4. **Care managers provide direct services to high-risk patients.** Patients have direct access to a care manager who is familiar with their health status; they can call or meet with their care manager as frequently as necessary instead of using the ED as a first point of care. Care managers contact these patients regularly to help them manage their care. The number of contacts varies by practice and by patient; contacts can occur by telephone, in person at the practice, or through home visits. In their routine contacts with patients, care managers use a Patient Activation Measure (PAM; developed by Insignia Health) to assess patients' activation to improve their health. Care managers use the PAM three times—at the first care management visit, 90 days after beginning care management, and at discharge—to help them assess a patient's overall needs and continued need for intensive care management. Care managers also assess patients' needs on a case-by-case basis, contacting patients who require more guidance (for example, those recently discharged from the hospital) as often as daily, depending on their needs. In addition to medical and behavioral health needs, care managers identify social and transportation needs to reduce patients' barriers to accessing care. Care managers coordinate patients' care among medical and community providers and connect patients with community-based service organizations and transportation services for their medical appointments. Care managers work with patients until they feel patients would no longer benefit from care management or patients decide they no longer need it.
5. **The direct care management services and better access to medical care and social services improves patients' clinical care and self-management.** Patients are more informed and in charge of their care; they are more activated to manage and improve their health. Because of their regular meetings with care managers; increased self-management of care; and improved connection to medical, community-based, and transportation services, patients are expected to better adhere to treatment recommendations.
6. **These improvements in self care and clinical care keep high-risk patients' chronic conditions under better control,** thereby reducing outpatient ED visits, inpatient admissions for ACSCs, and all-cause inpatient admissions. After an admission for inpatient care, care managers follow up with patients to ensure their care needs are being met, thereby reducing the number of inpatient admissions followed by an unplanned readmission within 30 days. As a result of reduced care admissions and readmissions, total Medicare Part A and B spending decreases.

Text box III.2. Example from FLHSA illustrating the program's theory of action for care management

"One patient, when he first started, his [hemoglobin] A1c was 10 and he was in and out of the hospital. When I started working with him I said, 'What is the most important thing for you?' He was told he couldn't drive because his blood sugar was so out of control. He wanted to be able to drive again. I said, 'I understand that, but do you understand that it's not safe for you to drive while your blood sugar is not controlled?' So I asked him how willing he was to work on controlling his blood sugar to see if he'd be able to drive again. He was very interested in that. So I started asking him, 'What things would help you with that?' We started working on his diet and his food choices. He got his [hemoglobin] A1c down and he was told he could drive again."

Source: Interview with care manager, April 2015 site visit.

4. Intervention staff and workforce development

Table III.1 provides key details about staff involved in the HCIA-funded intervention. At the administrative level, FLHSA hired an HCIA-funded program director, data analyst, and program assistant (all of which are full-time positions) to oversee and support overall program implementation. FLHSA also hired practice improvement advisors to work with practice staff on practice transformation activities, and it hired a practice improvement coordinator who oversaw the practice improvement advisors and served as a practice improvement advisor for a group of participating practices. In addition, FLHSA hired clinical advisors and a social worker/resource coordinator to provide care managers with guidance and help integrate them into practice care teams. Clinical advisors were not in the original staffing plan; FLHSA decided to hire them after the program began. There were no adaptations to any other positions during the award.

At the practice level, FLHSA allocated HCIA funds to practices so that they could hire care managers. FLHSA initially allocated these funds based on the number of PCP hours in the practice, and later allocated funds based on practices' risk-adjusted patient panel size. As a result of these allocation methods, a few practices had more than one care manager and many practices had part-time or shared care managers. Most care managers were registered or licensed practical nurses; a few were social workers. Before December 2014, FLHSA's partner Trillium Health used HCIA funds to hire CHWs, who provided services at six practices.

Although HCIA funds did not cover their salaries, practice champions—PCPs who served as liaisons between FLHSA and practice staff—were important members of the intervention staff at each participating practice. These PCPs served as the primary advocates for practice transformation and integrated care management.

FLHSA provided a variety of staff training and workforce development activities to care managers and practice champions. First, care managers and practice champions attended monthly learning collaboratives, which provided opportunities to share lessons and challenges and to learn from the experiences of their peers at other sites. During learning collaboratives for practice champions, FLHSA staff provided training on team-based care, clinical data, and leadership. In care manager learning collaboratives, FLHSA staff provided training on strategies for using the PAM, teach-back methods, and case reviews. Second, care managers attended a comprehensive training on fundamental skills, such as data collection and entry, and received

supplementary trainings on such topics as motivational interviewing and how to use the PAM to assess patient activation, build care team relationships, and use EHRs. Cohort 1 care managers attended five consecutive, day-long training sessions (40 hours total), and Cohort 2 and 3 care managers attended a pair of two-day sessions (32 hours total). Care managers hired later in the process attended two 8-hour make-up sessions. Finally, Insignia Health trained care managers on, and provided support for, use of its PAM to provide intensive care management services to patients.

Table III.1. Key details about intervention staff

Program component	Staff member	Staff/team responsibilities	Adaptations
All components	Program director	• Oversaw program strategy and execution, managed program staff and relationships with external partners, conducted research, and disseminated findings	None
	Data analyst	• Analyzed clinical and financial data and obtained, collected, and analyzed data for program use	None
	Program assistant	• Provided administrative and logistical support to program and program staff	None
Practice transformation	FLHSA practice improvement advisors	• Provided technical support to practice champions • Assessed needs of individual practices and worked with practice staff to develop and test solutions (for example, assisting practices with Plan-Do-Study-Act cycles) • Worked with practices to help them transform into PCMHs (for example, identifying processes and resources for managing admissions, discharges, and transitions of patients)	None
	FLHSA practice improvement coordinator	• Oversaw practice improvement advisors • Served as practice improvement advisor for designated practices	None
	Practice champions	• Oversaw on-site implementation of practice transformation activities • Served as main point of contact with FLHSA program staff • Met regularly with FLHSA practice improvement advisor • Attended FLHSA learning collaboratives	None

Table III.1 *(continued)*

Program component	Staff member	Staff/team responsibilities	Adaptations
Care management	FLHSA clinical advisors	• Provided technical support to care managers (for example, helping care managers report on clinical quality measures through practice EHRs) • Met with each care manager at least biweekly to discuss challenges and provide education and training on topics such as motivational interviewing, EHR use, and care team relationships (for example, building rapport with other staff at the practice)	• Clinical advisors were not in the initial staffing plan; FLHSA added these positions after the program began. • As an alternative to individual meetings with care managers, clinical advisors piloted small-group meetings (3 to 8 care managers grouped by practice affiliation with a health system or medical group).
	FLHSA social worker/resource coordinator	• Provided resources and technical assistance to care managers to help connect patients with necessary services at community-based service organizations • Organized trainings and networking sessions to introduce care managers to community resources	None
	Care managers	• Provided intensive care management to high-risk patients • Worked with practice staff to define the embedded care management role and implement care management processes, such as daily huddles and weekly care team meetings • Communicated with practice providers by documenting care they provided in EHRs, discussing the patients at care team meetings and huddles, and meeting informally with the providers during the workday • On a monthly basis, care managers submitted data to FLHSA about their patients and the care services delivered, such as number of patients on their caseload, PAM scores, and insurance information • Met regularly with FLHSA clinical advisor and social worker • Attended FLHSA learning collaboratives	Guidance for care manager's role evolved: • Initially, FLHSA clinical advisors expected that care managers would spend 35 percent of their time on intensive care management, 25 percent on population management, 30 percent on care transitions, and 10 percent on developing relationships in the practice. • Over the course of the program, the clinical advisors revised this guidance after recognizing that the care manager's role encompassed more than these areas and will continue to evolve as the practice transforms more completely into a PCMH.
	Community health workers (CHWs)	• Educated practice staff in six practices on the needs of the local community • Connected patients to external resources	• In December 2014, because of challenges related to the integration of CHWs into the six practices, FLHSA stopped working with Trillium Health to identify and employ CHWs in these practices.

Sources: Interviews and document review.

Note: Primary care payment reform is a supplemental component and is not listed in the table.

EHR = electronic health record; FLHSA = Finger Lakes Health Systems Agency; PAM = Patient Activation Measure; PCMH = patient-centered medical home.

B. Implementation effectiveness

In this section, we examine the evidence on implementation effectiveness—that is, we analyze measures of the intervention as delivered and, when possible, compare those measures to the services the awardee intended to deliver. We assess the evidence on implementation effectiveness in five areas: (1) program enrollment, (2) service delivery, (3) staffing, (4) training, and (5) implementation timeliness. To conduct this analysis, we used data from interviews with program administrators and frontline staff and self-reported metrics included in FLHSA's self-monitoring and measurement reports to CMMI. Whenever possible, we report metrics through July 2015—one month after the end of the original award period—because the current impact analysis covers the period through July 2015, for a total of 28 intervention months.

1. Program enrollment

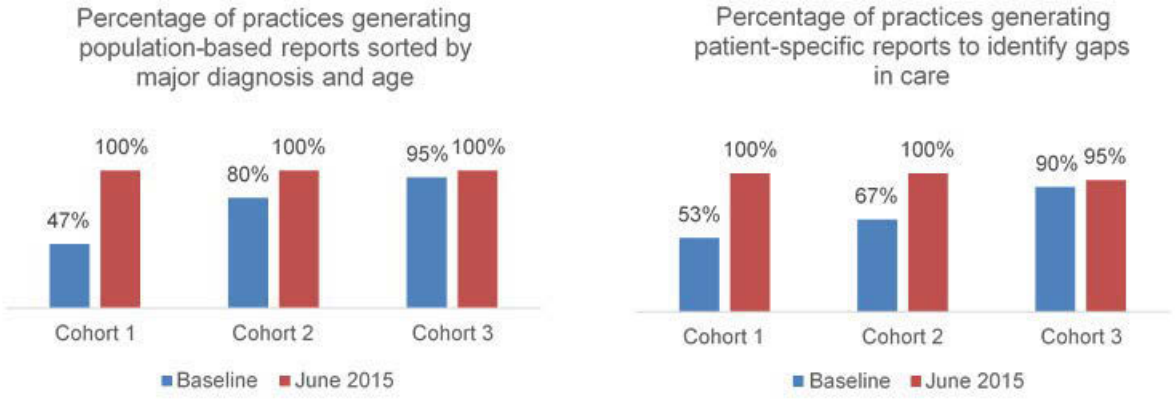
FLHSA successfully enrolled practices and reached more patients than expected. FLHSA recruited 68 practices (exceeding its target of 65; according to FLHSA, this is a little more than half of all primary care practices in FLHSA's six-county service area). All 68 practices participated for the duration of the original intervention period. As of July 2015, practices provided services to 17,484 unique patients, exceeding the target cumulative enrollment of 13,564 patients (about half of whom were expected to receive intensive care management; FLHSA did not have a stated target) for the entire award period.

2. Service-related measures

FLHSA's self-monitoring metrics indicate that practices effectively transformed the way they delivered care and improved on a variety of process measures in the time they participated in the program. Here, we describe FLHSA's self-reported metrics for each component. We report the metrics for each of the three cohorts separately, as each cohort has a different baseline (defined as the first month the practices were enrolled)—January 2013 for Cohort 1, July 2013 for Cohort 2, and July 2014 or Cohort 3. Although Cohort 1 practices had the most time to make progress, FLHSA staff noted that Cohorts 2 and 3 benefited from lessons learned early in the program. These metrics also indicate that practices were at different levels of transformation at baseline. For example, some practices already conducted huddles or used EHRs to generate reports on gaps in care.

Practice transformation. Practices successfully used EHRs to transform care. All three cohorts showed increases from baseline to July 2015 in the percentage of practices using EHRs to generate population-based reports, sorted by patients' age and diagnosis, and patient-specific reports to identify gaps in care (Figure III.1). Throughout the 28-month implementation period, almost all practices generated population-based reports from their EHRs and patient-specific reports (data not shown).

Figure III.1. FLHSA self-reported percentage of practices using EHRs to generate population and patient-specific reports, by cohort



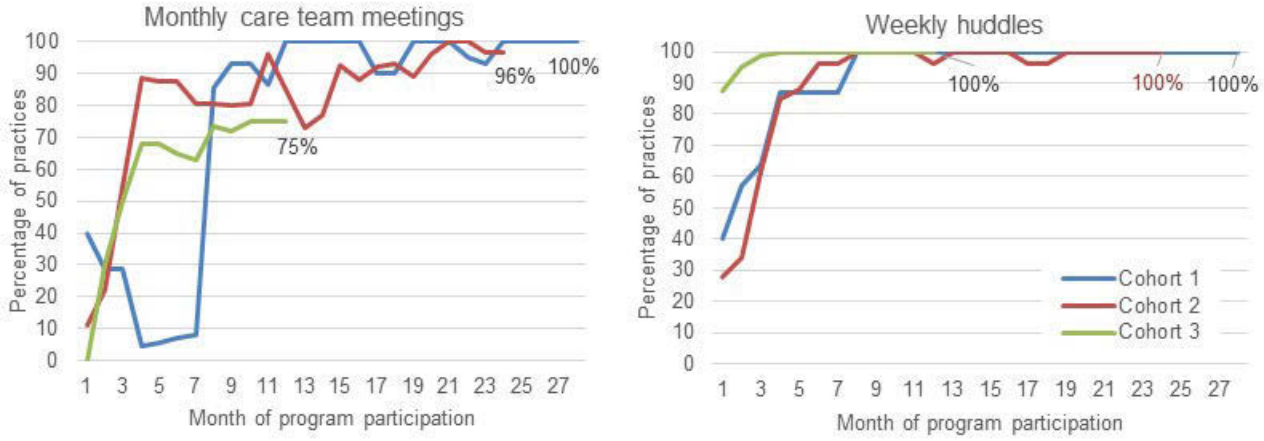
Source: Analysis of FLHSA's 12th quarter measuring and monitoring results. Prepared for CMMI, June 2015.

Note: This information is based on the awardee's self-reported data. We have not attempted to verify its completeness or quality. The baseline month of program participation varies by cohort. The baseline month is January 2013 for Cohort 1, July 2013 for Cohort 2, and July 2014 for Cohort 3.

CMMI = Center for Medicare & Medicaid Innovation; EHR = electronic health record; FLHSA = Finger Lakes Health Systems Agency.

In addition, practice staff successfully implemented monthly care team meetings and weekly huddles to coordinate care. From baseline to July 2015, across all cohorts, all practices had implemented weekly huddles and increasingly held monthly care team meetings (Figure III.2).

Figure III.2. FLHSA self-reported percentage of practices holding monthly care team meetings and weekly huddles, by cohort and month of program participation



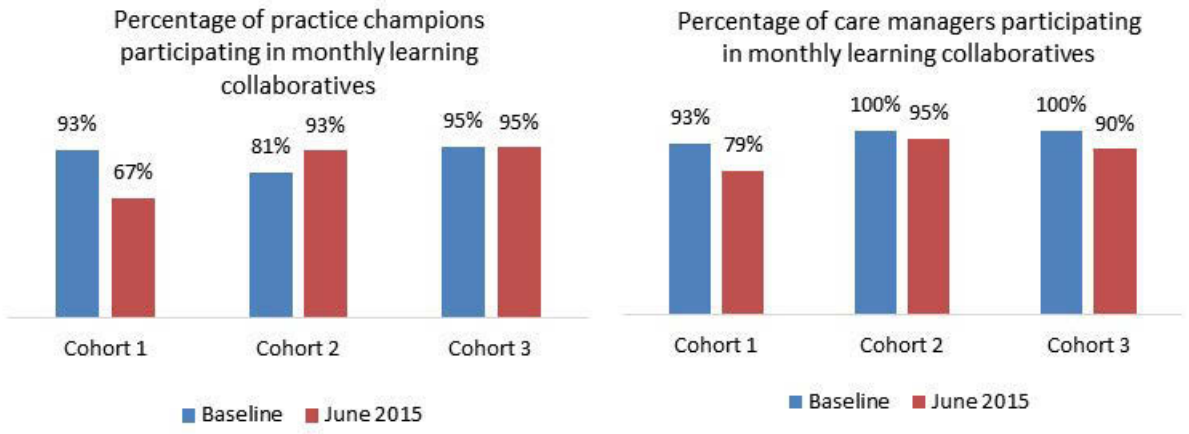
Source: Analysis of FLHSA's 12th quarter measuring and monitoring results. Prepared for CMMI, June 2015.

Note: This information is based on the awardee's self-reported data. We have not attempted to verify its completeness or quality. The first month of program participation varies by cohort. Month 1 is January 2013 for Cohort 1, July 2013 for Cohort 2, and July 2014 for Cohort 3. Labeled percentages indicate the percentage of practices in each cohort holding monthly care team meetings or weekly huddles as of June 2015.

CMMI = Center for Medicare & Medicaid Innovation; FLHSA = Finger Lakes Health Systems Agency.

By July 2015, two-thirds of Cohort 1 practice champions and almost all Cohort 2 and 3 practice champions participated in learning collaboratives (Figure III.3). As noted in FLHSA's 12th quarter (Q12) narrative, practice champions across the cohorts might have missed some of the learning collaboratives because of scheduling conflicts. Similarly, most care managers participated in learning collaboratives; however, as of July 2015, fewer Cohort 1 care managers participated in them than did the other two cohorts (Figure III.3).

Figure III.3. FLHSA self-reported percentage of practice champions and care managers participating in monthly learning collaboratives, by cohort



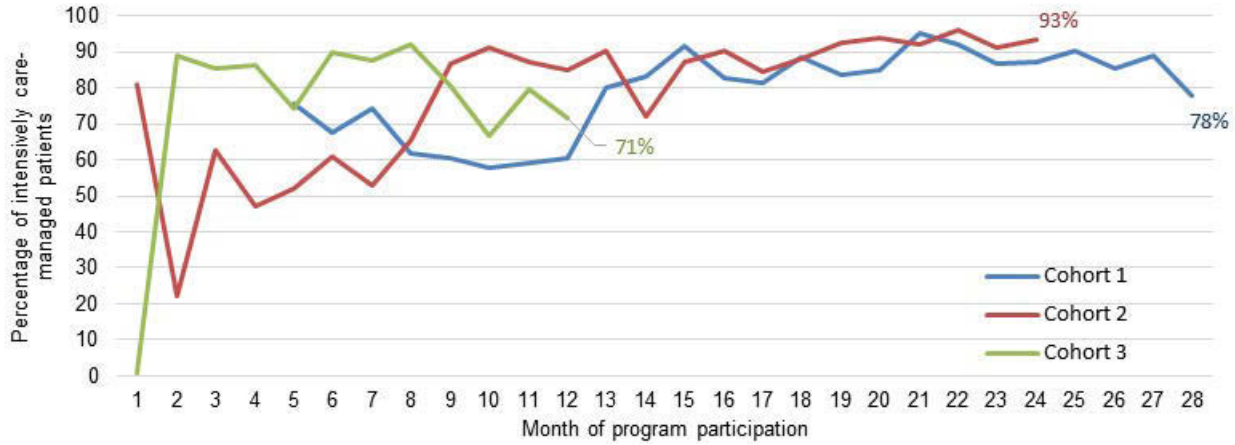
Source: Analysis of FLHSA's 12th quarter measuring and monitoring results. Prepared for CMMI, June 2015.

Note: This information is based on the awardee's self-reported data. We have not attempted to verify its completeness or quality. The baseline month of program participation varies by cohort. The baseline month is January 2013 for Cohort 1, July 2013 for Cohort 2, and July 2014 for Cohort 3.

CMMI = Center for Medicare & Medicaid Innovation; FLHSA = Finger Lakes Health Systems Agency.

Intensive care management. FLHSA did not collect information from participating practices on the number or type of patients who received intensive care management services, the duration of enrollment, the mode and frequency of contacts, or the specific issues discussed and addressed during the encounters. However, FLHSA tracked the extent to which care managers used the PAM to assess patient activation. By July 2015, care managers used the PAM to assess 71 percent of intensively care-managed patients in Cohort 1, 78 percent in Cohort 2, and 66 percent in Cohort 3 (Figure III.4).

Figure III.4. FLHSA self-reported percentage of intensively care-managed patients assessed using the PAM, by cohort and month of program participation



Source: Analysis of FLHSA's 12th quarter measuring and monitoring results. Prepared for CMMI, June 2015.

Note: This information is based on the awardee's self-reported data. We have not attempted to verify its completeness or quality. The baseline month of program participation varies by cohort. The baseline month is January 2013 for Cohort 1, July 2013 for Cohort 2, and July 2014 for Cohort 3. Labeled percentages indicate the percentage of practices in each cohort holding monthly care team meetings or weekly huddles as of July 2015.

CMMI = Center for Medicare & Medicaid Innovation; FLHSA = Finger Lakes Health Systems Agency; PAM = Patient Activation Measure.

3. Staffing measures

FLHSA successfully met or exceeded its staffing goals for the HCIA-funded intervention. Although FLHSA experienced some staff turnover, as of July 2015 it employed 11 program staff: 5 practice improvement advisors and 1 practice improvement coordinator (for 6 total, an increase from the original 3); 4 clinical advisors (a role added in response to practice needs); and 1 social work clinical coordinator. Participating practices also experienced some turnover in care managers, but as of July 2015 each practice met FLHSA's goal to employ at least 1 care manager, resulting in a total of 70 embedded care managers across the 68 practices. FLHSA hired 6 CHWs through its partner, Trillium Health, but as of December 2014 all 6 CHWs had left because of challenges integrating them into the practices.

4. HCIA-funded training

To assess perspectives of the care managers and practice champions on the effectiveness of the training they received, we administered the HCIA Primary Care Redesign Trainee Survey from January to March 2015 (25 to 27 months after implementation began in Cohort 1, 19 to 21 months in Cohort 2, and 7 to 9 months in Cohort 3). Of the 116 care managers and practice champions who worked at a participating practice and were familiar with the HCIA-funded program, 93 responded to the trainee survey (an 80 percent response rate).

Nearly all respondents (99 percent) reported receiving ongoing training through learning collaboratives (either for practice champions or care managers) (data not shown). Almost all care managers reported that they received care manager training (89 percent), training on patient activation mode (93 percent), and motivational interviewing training (98 percent).

In general, most care managers and practice champions (88 percent) rated the training as good or excellent (Table III.2). Respondents found the trainings useful, with more than 85 percent strongly or somewhat agreeing that topics covered in the trainings were relevant (89 percent) and useful in their work (92 percent), and that the training would help to improve their job performance (87 percent). Among care managers who received training and responded to the survey, the proportions with positive ratings were similar to the larger sample.

The trainee survey also collected information of the perceived effect of training on specific aspects of care. Most respondents indicated that the trainings had a positive impact on aspects of care related to the HCIA-funded intervention's first component, practice transformation. Care managers and practice champions reported that trainings had a positive impact on quality of care (77 percent), the ability to respond to patients' needs in a timely way (67 percent), patient-centeredness (84 percent), equity of care for all patients (69 percent), and relaying relevant information to the care team (74 percent) (Table III.2). Care managers were more likely than the full group of respondents to indicate that the trainings had a positive impact on these aspects of care.

Most respondents also indicated that trainings positively affected aspects of care related to the intervention's second component, care management. More than two-thirds of respondents reported that trainings had a positive impact on their ability to help patients access the care they need (81 percent), access nonmedical services (71 percent), and take control of their own care (72 percent) (Table III.2). Among the 45 care managers who responded to the survey, 87 percent reported that the trainings positively affected helping patients access the care they need, and more than two-thirds reported positive impacts on working with a diverse set of patients (70 percent) and explaining care information to patients and their families (76 percent).

Despite the overall positive findings, a substantial portion of care managers and practice champions felt that the training they received for the intervention had no effect on three aspects of care related to practice transformation or care management. The aspects of care with the highest percentage of respondents reporting no effect were the ability to clearly explain information about patients' care to patients and their families (31 percent), the ability to work with a diverse set of patients (37 percent), and using data to evaluate the respondent's performance to improve services provided to patients (24 percent) (data not shown). Care managers were generally more positive than the overall sample about the impact of the intervention on these three aspects of care related to practice transformation or care management, with the exception of using data to evaluate their performance to improve services provided to patients; 35 percent of care managers who responded to the survey felt the intervention did not have an impact on this aspect of care.

In addition, some respondents felt it was too soon to tell if the training had an impact on several other aspects of care. For example, 27 percent of respondents felt it was soon to tell if the training affected efficiency or cost-effectiveness of care (data not shown). Almost one in five (18 percent) felt it was too soon to know the impact of the training on their ability to respond to patients' needs in a timely way.

Table III.2. Care managers' and practice champions' perceptions of the effects of training on the care they provide to patients, from the trainee survey (separately by all respondents and restricted to care managers)

Survey question	Percentage (and number) of all respondents	Percentage (and number) of care managers
Rating of training received related to CMMI program[a]		
Excellent	34% (30)	38% (17)
Good	54% (47)	51% (23)
The topics covered were relevant to me[a]		
Strongly agree	54% (47)	51% (23)
Somewhat agree	35% (31)	33% (15)
The training experience will be useful in my work[a]		
Strongly agree	55% (48)	62% (28)
Somewhat agree	37% (32)	33% (15)
The training helped me to improve my performance or complete my new job responsibilities[a]		
Strongly agree	46% (40)	53% (24)
Somewhat agree	41% (36)	38% (17)
Training had a positive impact on[b]		
Quality of care	77% (69)	89% (41)
Ability to respond in a timely way to patients' needs	67% (60)	80% (37)
Efficiency or cost-effectiveness of care	54% (49)	61% (28)
Patient-centeredness of care	84% (76)	93% (43)
Equity of care for all patients	69% (62)	76% (35)
Training had a positive impact on respondents' ability to[b]		
Explain information about patients' care to patients and their families in lay terms	62% (56)	76% (35)
Relay relevant information to the care team	74% (67)	74% (34)
Work with diverse set of patients	53% (48)	70% (32)
Help patients access the care they need	81% (73)	87% (40)
Help patients access nonmedical services	71% (64)	74% (34)
Help patients take control of their own care	72% (65)	87% (40)
Use data to evaluate my performance to improve the services I provide to patients	51% (46)	46% (21)

Source: Mathematica's analysis of trainee survey.

[a] The denominator is 87 for all respondents (practice champions and care managers), and 45 for care manager respondents and includes all trainees who reported they received any formal training as part of the FLHSA program.

[b] The denominator for all respondents (practice champions and care managers) is 90 and includes all trainees who reported they received any formal (87) or informal (47) training as part of the FLHSA program. The denominator is 46 for care manager respondents and includes all care manager trainees who reported they received any formal (45) or informal (26) training as part of the FLHSA program.

CMMI = Center for Medicare & Medicaid Innovation; FLHSA = Finger Lakes Health Systems Agency.

Finally, the trainee survey collected information on how intervention staff spend their time. The data confirmed that care managers spent their time educating, communicating with, and counseling patients in ways that are consistent with FLHSA's program design (Table III.3). Almost all care managers (98 percent) reported that they routinely managed patients' care by educating patients about self-care. Most care managers also reported that they routinely helped to manage patients' care in the following ways: calling patients to check on medications, symptoms, or helping coordinate care between visits (91 percent); counseling patients on exercise, nutrition, and how to stay healthy (83 percent); attending team meetings or care conferences (83 percent); coaching patients (78 percent); and following up on care transitions (76 percent). Most surveyed care managers also reported that they took part in activities such as care team meetings and contacting patients to assist with medical, social, and behavioral needs (Table III.3). Most (83 percent) care managers reported that they attended team meetings and 78 percent reported coaching patients. More than half (57 percent) of care managers reported that they routinely assisted patients with accessing nonmedical services, such as housing, job training, or supplemental nutrition services.

Table III.3. Care managers' activities, as reported in the trainee survey (n = 46)

	Percentage (and number) of care managers who reported that they	
Activity	Personally help to manage patients' care through this activity *routinely*	Spend more than 2 hours on this activity on a typical work day
Call patients to check on medications, symptoms, or help coordinate care between visits	91% (42)	46% (21)
Execute standing orders for medication refills, ordering tests, or delivering routine preventive care	< 11	< 11
Educate patients about managing their own care	98% (45)	43% (20)
Counsel patients on exercise, nutrition, and how to stay healthy	83% (38)	33% (15)
Assist patients with accessing nonmedical services such as housing, job training, supplemental nutrition services (for example, SNAP benefits)	57% (26)	< 11
Attend medical appointment with patients	24% (11)	< 11
Conduct home visits with patients	< 11	< 11
Follow up on care transitions	76% (35)	< 11
Patient coaching	78% (36)	30% (14)
Attend team meetings/care conferences	83% (38)	< 11

Source: Mathematica's analysis of trainee survey.

Note: Questions with fewer than 11 responses are suppressed because the numerator is less than 11.

SNAP = Supplemental Nutrition Assistance program.

5. Program timeline

FLHSA successfully implemented both the practice transformation and the intensive care management components on schedule for all three cohorts of practices. Practices were recruited for each of the cohorts on or ahead of schedule, and care managers were hired at each participating practice within three months of its launch date. The Centers for Medicare & Medicaid Services (CMS) awarded FLHSA a one-year no-cost extension that enabled FLHSA staff to continue to support practices from all three cohorts financially and with technical assistance through June 2016. During this period, FLHSA changed its practice incentive strategy to reward practices for completing specific tasks rather than paying a set stipend per participating physician. FLHSA defined 16 deliverables (tasks) for which practices would be paid upon completion. FLHSA required that practices complete 4 of the 16 deliverables related to attending learning collaboratives and reporting requirements. FLHSA allowed practices to select from the remaining 12 deliverables, which included holding practice improvement meetings, using chronic care management billing codes, integrating behavioral health care into primary care, and implementing care management activities. FLHSA also stopped reimbursing practices for care managers' salaries during the no-cost extension and no longer required practices to employ care managers. Of the original 68 practices, 10 elected not to participate in the intervention during the no-cost extension period.

C. Summary of facilitators of and barriers to implementation

Several factors facilitated implementation of FLHSA's HCIA-funded intervention, although others hindered implementation. We described those factors in detail in the second annual report (Shapiro et al. 2016). Here, we summarize key facilitators and barriers (Table III.4).

Eight factors were particularly important in facilitating implementation of the intervention:

1. Practice staff perceived the intervention as a relative advantage compared with the standard delivery of care because of its increased emphasis on the care team, the presence of a practice-based care manager, and improved communication with patients.
2. Practice staff's ability to adapt the intervention to their own practices' needs helped them to implement the program effectively and achieve more patient-centered care.
3. The intervention helped practices use data to self-monitor and conduct quality improvement activities.
4. Practice staff increasingly engaged and worked with the care managers.
5. In addition to their commitment to integrating care managers, practice champions and practice managers were committed to transforming their practice workflows and improving care team communication.
6. Practices demonstrated strong team communication and collaboration through the use of care teams.
7. Most practice staff were committed to improving care through integrated care teams and worked toward transforming their practices into PCMHs.

8. Developing external payment models at two regional ACOs facilitated implementation of the FLHSA intervention by providing financial and technical support to assist with practice transformation and care management activities at those practices that were part of an ACO.

Several important barriers to implementation emerged. First, staff had limited time to devote to transformation activities, exacerbated by high caseloads and intervention-specific requirements. For practice champions, the HCIA-funded intervention was only part of their job, and they reported that transformation activities often took more time than they had available. Regular meetings—such as weekly care team huddles, monthly care team meetings, and learning collaboratives—added a substantial burden to their workloads, in addition to other practice transformation activities for the intervention. Care managers devoted all of their time to the intervention, but reported that they often had insufficient time to manage high patient caseloads and associated reporting requirements, in addition to other FLHSA requirements (such as attending learning collaboratives). Second, regional ACO system requirements for intensive care management further exacerbated time constraints for care managers; those system requirements sometimes conflicted with FLHSA requirements. For example, one ACO required that care managers should have a relatively high caseload for intensively managed patients (65 patients, compared with 40 to 60 patients for FLHSA). ACOs also imposed additional meetings and reporting requirements on care managers. A third barrier—ineffective integration of CHWs into participating practices—affected only six practices in the first cohort.

D. Conclusions about the extent to which the program, as implemented, reflects core design

FLHSA implemented its HCIA-funded intervention largely as planned. As previously noted, the 68 practices exceeded FLHSA's targets for the number of patients enrolled in the intervention. FLHSA practice improvement advisors delivered services for the intervention's practice transformation component, and FLHSA clinical advisors delivered services for the care management component. Practice champions spearheaded the practice transformation initiatives at each practice, and all 68 participating practices successfully hired care managers to provide targeted, intensive care management. In surveys, most practice champions and care managers reported that they (1) participated in learning collaboratives; and (2) felt that trainings had a positive impact on the quality, timeliness, and patient-centeredness of care, and relaying relevant information to the care team. Most respondents also felt the trainings positively affected their ability to help patients access medical and nonmedical services and improve their self-care.

By July 2015, most practices were well along a path of transforming primary care processes and delivery. All practices had adopted EHRs to generate population-based reports sorted by patients' age and diagnosis, and nearly all used patient-specific reports to identify gaps in care. More than three-quarters of practices held monthly care team meetings and all practices held weekly huddles. Care managers generally spent their time as expected, with most reporting that they routinely educated patients about self-care, calling patients between visits, counseling and coaching patients, attending team meetings or care conferences, and following up on care transitions.

Table III.4. Summary of key facilitators of and barriers to FLHSA program implementation

Item	Description based on findings in second annual report
	Facilitators
Perceived relative advantage of the program compared with the standard delivery of care (program characteristics)	Practice staff reported that several factors improved their care delivery since they began participating in the program, including an increased emphasis on the care team, the presence of an embedded care manager, and improved communication with patients. As a result of the program's focus on team-based care, practice staff reported either holding or increasing the frequency of huddles, improving the efficiency and effectiveness of pre-visit planning, and adapting to a team-based approach to care. In particular, practice staff appreciated the collaboration provided through the care team approach, viewing it as an advantage over the way they previously provided care. Interviewed providers appreciated the added degree of patient-centered care delivered by the care managers, which they felt had led some patients to better control their conditions.
Adaptability of the program to practices' and patients' needs (program characteristics)	Adaptability is built into the FLHSA program design; FLHSA practice improvement advisors and clinical advisors tailor their coaching and mentoring to the needs of the practices and care managers. In transforming practice workflows, FLHSA practice improvement advisors let practices chart their own course, identifying projects that would help them achieve more patient-centered care. FLHSA also allowed practices to use the approaches that worked best to identify and reach their targeted high-risk patients and provide them with care management. This often resulted in practices providing care management to different populations. In addition, FLHSA practice improvement advisors and clinical advisors did not limit practices to implementing a standardized model of care management; instead, they allowed practices to assess their patients' needs and tailor their use of the care manager in a way that best met their patients' needs.
Using data to self-monitor and conduct quality improvement activities (implementation process)	FLHSA helps practices to monitor their own progress, as well as how they compare with other participating practices' progress, by providing quarterly reports that summarize practice-level clinical, quality, and cost data; these quarterly reports supplement any reports that practices generate through their EHRs or receive from their hospital system or ACO.
Staff engagement related to the embedded care manager role (implementation process)	Providers reported that staff engagement with the embedded care manager increased over the course of practices' participation in the program. At first, respondents reported that some staff hesitated to embrace care managers, largely because they did not understand how the care managers should function in the practices. As providers grew to understand the care managers' role, saw them in action, and noticed changes in some patients' behaviors, providers started to appreciate the added care being provided and were more likely to refer high-risk patients to the care managers.
Leadership commitment (internal factors)	Practice champions and practice managers were committed to transforming their practice workflows; improving communication among members of the care team; and integrating care managers into the care team, particularly in light of national and statewide initiatives for new payment models based on the provision of patient-centered care and quality improvement. During site visits, staff pointed to the practice champions as a driving force behind practice change.

Table III.4 *(continued)*

Item	Description based on findings in second annual report
Team characteristics (internal factors)	Practices demonstrated strong team communication and collaboration. Much of FLHSA's coaching related to practice transformation focuses on building successful care teams. As a result, and perhaps not surprisingly, practice staff reported that these care teams helped move forward the practices' transformation efforts.
Implementation climate (internal factors)	Practice staff reported implementation climates that were favorable to practice transformation and integration of care managers. Most staff were committed to improving how they provided care and collaborated with care managers. Some practices were already moving toward becoming a PCMH before participating in the FLHSA program; at these practices, staff readily embraced team huddles and the opportunity to practice at the top of their licenses.
External payment models developed by two regional ACOs (external environment)	Since the beginning of the FLHSA program, many participating practices joined one of the two regional ACOs. Practice staff commented that the ACOs' support of practice transformation—by providing practices with population data, consultants to assist with PCMH certification, or financial support for care management—encouraged their practice transformation efforts.
Barriers	
Program resources in relation to the time required for transformation activities (implementation process)	Practice champions and other providers struggled to devote sufficient time to the transformation activities, reporting difficulty finding time to attend daily huddles, care team meetings, and learning collaboratives. Care managers also reported struggling to find sufficient time to perform all of the tasks required because of their high caseloads and the requirements that came with participating in the FLHSA program.
Regional ACO care manager requirements (external environment)	ACOs' system requirements for care managers and working with high-risk patients can be stricter than FLHSA's requirements. For example, one ACO requires care managers to have higher caseloads and has more restrictive protocols for identifying, enrolling, and providing services to care-managed patients. ACOs also required care managers to attend additional meetings and submit additional reports. FLHSA practice improvement advisors and clinical advisors reported that they worked closely with the ACOs to streamline guidance for care management such that FLHSA guidance did not conflict with system requirements.
Program execution in relation to integrating CHWs into practices (implementation process)	Although Trillium Health and FLHSA staff initially worked with selected practices to integrate CHWs, in hindsight it is clear that communications could have been improved, as several of these practices were unclear on how CHWs should function in their practices and did not assign work to them. FLHSA staff felt more effective management of the relationship between the practices and the CHWs could have prevented these issues. FLHSA leadership suggested that focused trainings and mentoring for CHWs and practices to clarify the CHW role and expectations of the position might have improved the integration of CHWs into practices.

Sources: Interviews with FLHSA and practice staff, Mathematica's analysis of clinician and trainee survey data, and self-reported awardee data.

Note: Other chapters present additional supporting data not previously available for the second annual report. However, FLHSA did not share any additional data relevant to implementation facilitators and barriers.

ACO = accountable care organization; CHW = community health worker; EHR = electronic health record; FLHSA = Finger Lakes Health Systems Agency; PCMH = patient-centered medical home.

Despite these overall positive implementation findings, a substantial proportion of trainee survey respondents felt the intervention would have no effect on specific aspects of care and there were still gaps in the care that they were supposed to provide routinely. About a quarter or more of respondents to the trainee survey reported that the program would have no effect on three aspects of care related to practice transformation or care management: (1) the ability to clearly explain information about patients' care to patients and their families; (2) the ability to work with a diverse set of patients; and (3) use of data to evaluate the respondent's performance to improve services provided to patients. In addition, although one goal for care managers was to link patients with community resources, only 57 percent of surveyed care managers indicated that they routinely helped patients access nonmedical services.

Taken together, the implementation findings suggest that FLHSA implemented its program as planned. This provides us with the opportunity to test the effect of the program on patient-level outcomes.

IV. CLINICIANS' PERCEPTIONS OF PROGRAM EFFECTS ON THE CARE THEY PROVIDED TO PATIENTS

This section describes the available evidence on the extent to which FLHSA's intervention had its intended effects on changing PCPs' behavior as a way to achieve desired impacts on patients' outcomes. As described in Section III.A.3, the intervention's theory of action requires that PCPs (1) were involved in quality improvement projects at the participating practices, (2) participated in care team meetings and implemented care team huddles, and (3) actively engaged care managers. We use data from two rounds of the HCIA Primary Care Redesign Clinician Survey to assess changes in providers' behavior and conclude whether the anticipated changes in their behavior occurred. Both surveys relied on self-reported responses and reflected clinicians' perceptions of the program, rather than measuring quantitatively direct program effects on the care they provided to patients.

A. Clinician survey

Survey methods. We administered the Clinician Survey in two rounds (fall 2014 and summer 2015). We sent the survey to PCPs working in the 68 practices at the time of each round (137 PCPs in Round 1 and 200 PCPs in Round 2), including physicians, nurse practitioners, and physician assistants. The survey respondents included practice champions and other clinicians working at the practices; they did not include care managers. A total of 86 and 117 clinicians participating in the intervention responded to the survey during the first and second rounds, respectively (resulting in a response rate of 70 percent in Round 1 and 61 percent in Round 2).

Survey results. Most clinicians who responded to the survey reported being somewhat or very familiar with the HCIA-funded intervention (84 percent in Round 1 and 85 percent in Round 2). As shown in Table IV.1, the intervention appears to have had its intended effects for most providers familiar with the intervention on dimensions related to care management, including patient-centeredness and quality of care, and the clinicians' ability to respond to patients' needs in a timely manner. However, only about half of surveyed clinicians thought the intervention had a positive effect on patients' safety, and fewer than half of surveyed clinicians

felt the intervention had a positive effect on equity of care or the information available for clinical decision making. The remaining clinicians responded that the intervention had no effect on these outcomes or that it was too soon to tell whether it did. Compared with the first round, clinicians in the second round of the survey were more likely to report that the intervention had a positive effect on care efficiency (55 percent compared with 38 percent in Round 1). The findings in both rounds of the survey were largely similar for other dimensions of care (Table IV.1).

Table IV.1. PCPs' perceptions of the effects of the program on the care they provided to patients, from the clinician surveys (both rounds)

	Percentage (and number) of PCPs reporting that the FLHSA program had the following effect on the care they provided to patients enrolled in their practice in the past year					
	First round of survey (2 to 23 months after program implementation) N = 72			Second round of survey (10 to 31 months after program implementation) N = 99		
Dimension of care	Positive impact	No impact	Too soon to tell	Positive impact	No impact	Too soon to tell
Quality	65% (47)	< 11	24% (17)	70% (69)	13% (13)	16% (16)
Ability to respond in a timely way to patients' needs	60% (43)	25% (18)	< 11	56% (55)	29% (29)	13% (13)
Efficiency	38% (27)	25% (18)	24% (17)	55% (54)	26% (26)	15% (15)
Safety	53% (38)	24% (17)	24% (17)	49% (49)	33% (33)	16% (16)
Patient-centeredness	69% (44)	< 11	18% (13)	73% (72)	13% (13)	13% (13)
Equity	44% (32)	36% (26)	18% (13)	36% (36)	42% (42)	19% (19)
Information available for clinical decision making	NA	NA	NA	41% (41)	39% (39)	18% (18)

Source: Clinician Survey Round 1 (field period September 2014 through November 2014) and Round 2 (field period May 2015 through July 2015).

Note: The number (and percentages) are limited to PCPs reporting they were at least somewhat familiar with the HCIA program.

We do not report numbers when the numerator is smaller than 11.

FLHSA = Finger Lakes Health Systems Agency; HCIA = Health Care Innovation Award; PCP = primary care provider.

NA = not available.

B. Conclusions about intermediate program effects on clinicians' behavior

Based on available information, the HCIA-funded intervention appears generally to have had its intended effects on how most PCPs at the participating practices provided care. More than 84 percent of PCPs surveyed were aware of the intervention, and most believed the HCIA-

funded intervention improved the quality and patient-centeredness of care at their practices. Slightly more than half felt that the intervention improved efficiency and timeliness of care. However, in the second round of the survey, more than one-quarter of PCPs thought that the program had no effect on efficiency or timeliness of care, and one-third or more felt the program had no effect on safety, equity of care, or the information available for clinical decision making. This suggests that for a small fraction of participating PCPs, the program is not transforming care in the way FLHSA had hoped. It is important to note that PCPs had participated in FLHSA's HCIA-funded intervention for varying amounts of time when they responded to the clinician survey. In particular, Cohort 3 PCPs had participated in the intervention for only 10 months when they responded, which could have influenced their responses. However, the findings between Rounds 1 and 2 of the survey are similar, suggesting length of PCP participation did not influence survey results.

These conclusions are based on clinicians' perceptions of program effects on the care they provided to patients; we do not have independent evidence on whether the specific changes in clinicians' behavior anticipated in the theory of action actually occurred. For example, we do not have data on how much time clinicians spent with care managers in developing and implementing care plans for high-risk patients.

V. PROGRAM IMPACTS ON PATIENTS' OUTCOMES

This section of the report draws conclusions, based on available evidence, about the impacts of FLHSA's HCIA program on patients' outcomes in four domains: quality-of-care processes, quality-of-care outcomes, service use, and spending. We first describe the methods for estimating impacts (Section V.A) and then the characteristics of the 37 HCIA treatment practices at the start of the intervention (Section V.B). We next demonstrate that the treatment practices were similar at the start of the intervention to the practices we selected as a comparison group, which is important for limiting potential bias in impact estimates (Section V.C). Finally, in Section V.D, we describe the quantitative impact estimates, their plausibility given implementation findings, and our conclusions about program impacts in each domain. The findings in this report update the impact results from the second annual report for FLHSA (Shapiro et al. 2015), extending the outcome period by 6 months and adding new outcomes to assess quality-of-care processes. Our conclusions in this report are preliminary because the analyses do not yet include the 12-month extension beyond the original award period, nor do they include an analysis of Cohort 3 practices.

A. Methods

1. Overview

We estimated program impacts on patients' outcomes as the difference in outcomes for Medicare FFS patients served by 37 treatment practices and those served by 108 matched comparison practices, adjusting for any differences in outcomes between these groups during the year before the intervention began. We prespecified primary tests, describing the evidence we would need to conclude that the program was effective, and the awardee and CMMI reviewed these. Each test specified a population, outcome, time period, expected direction of effect, and threshold that we count as substantively important. The purpose of these primary tests was to

focus the impact evaluation on hypotheses that would provide the most robust evidence about program effectiveness. We used the results from the primary and secondary tests (robustness checks) to draw conclusions about program impacts in each of the four evaluation domains. The remaining subsections describe each component of the impact evaluation in more detail.

2. Treatment group definition

The treatment group consisted of Medicare FFS patients served by 16 Cohort 1 practices and 21 Cohort 2 treatment practices (for a total of 37 treatment practices among the 48 participating practices in the first two cohorts) in 4 baseline quarters before the intervention began (January 1, 2012, to December 31, 2012 for Cohort 1 practices; and July 1, 2012, to June 30, 2013, for Cohort 2 practices), 10 intervention quarters for Cohort 1 practices (January 1, 2013, to June 30, 2015), and 8 intervention quarters for Cohort 2 practices (July 1, 2013, to June 30, 2015).

We excluded 6 of the 48 practices in the first two cohorts because they were FQHCs, and had no suitable comparison. (We attempted to match these 6 FQHCs to nonparticipating FQHCs in New York State using practice-level data from the Health Resources & Services Administration Data Warehouse. However, we found that participating FQHCs were much larger in size and their patient populations differed systematically from other FQHCs in the state.) We also excluded another 2 practices because they served psychiatric or pediatric populations, and 3 practices because they had no attributed Medicare patients in at least one quarter of the evaluation baseline period (January to December 2012 for Cohort 1 practices and July 2012 to June 2013 for Cohort 2 practices), and were thus incompatible with the statistical regression model used to measure impacts.

We constructed the treatment group in three steps.

1. First, we attributed beneficiaries to practices using the same decision rule that CMMI uses for the CPC initiative. Specifically, in each baseline and intervention month, we attributed beneficiaries to the primary care practice whose providers (physicians, nurse practitioners, or physician assistants) provided the plurality of primary care services in the past 24 months. When there was a tie, we attributed the beneficiary to the practice he or she visited most recently. This attribution method requires identifiers for the providers who worked in the treatment practices, as well as identifiers for providers in other practices in the region who could compete for patients; these identifiers determine the practice that provided the plurality of primary care services. SK&A, an outside health care data vendor, supplied identifiers for providers in the treatment practices.
2. Second, in each baseline and intervention period, we assigned each patient to the first treatment practice to which he or she was attributed in that period, and continued to assign him or her to that practice for all quarters in the period. This assignment rule—which is distinct from the attribution method—ensures that, during the intervention period, patients did not exit the treatment group solely because the intervention succeeded in reducing their service use (including visits at treatment panels). The definition for the baseline period corresponds to that of the intervention period so that, across the two periods, interpretation of the population changes over time should be comparable.

3. Third, we applied additional restrictions to refine the analysis sample in each quarter. A patient assigned to a treatment practice in a quarter was included in the analysis sample for that quarter if he or she (1) had observable outcomes for at least one day in the quarter and (2) lived in New York or Pennsylvania for at least one day of the quarter. For this sample, outcomes were observable for beneficiaries who were enrolled in Medicare FFS (Part A and B), were alive, and had Medicare as their primary payer.

In addition to this full treatment sample, we defined a subset of patients who were at high risk of hospitalizations and other expensive medical care. This high-risk subgroup enabled us to conduct primary tests (Section V.A.6) examining whether any observed effects were concentrated among high-risk members. This would be expected from the program's theory of action, given that FLHSA targets its care management services to high-risk Medicare and Medicaid beneficiaries. In each baseline quarter, we defined the evaluation's high-risk subgroup to consist of beneficiaries with a Hierarchical Condition Category (HCC) score in the top third of all treatment group members with observable outcomes at the start of the baseline period. The HCC score, developed by CMS, is a continuous variable that predicts a beneficiary's Medicare spending in the following year relative to the national average, with 1.0 indicating that the predicted spending is at the national average and 2.0 indicating that it is twice that average. The HCC score is likely correlated with utilization and cost data used by treatment practices to identify beneficiaries who would benefit from intensive care management services. In each intervention quarter, we defined the high-risk population to consist of beneficiaries whose HCC scores were in the top third of all observable Medicare beneficiaries assigned to the treatment panels at the start of the intervention period.

3. Comparison group definition

The comparison group consisted of Medicare FFS beneficiaries we assigned to 108 matched comparison practices in each of the baseline and intervention quarters. The comparison practices were similar to the treatment practices during the baseline period on factors that can influence patients' outcomes, especially those factors that FLHSA used when deciding which practices to recruit for the intervention. This section describes how we constructed the matched comparison group; Section V.C shows the balance we achieved between the two groups on the matching variables.

We identified the 108 comparison practices in four steps:

1. First, we limited the potential comparison practices to the approximately 2,000 primary care practices in New York State that were located outside of (1) the greater New York City area, (2) the 6 counties in which FLHSA was operating, and (3) 13 counties in New York that had relatively strong participation in federal primary care initiatives. (Relatively strong participation in federal primary care initiatives is defined as counties in which at least 10 percent of primary care practices participated in the MAPCP Demonstration or the CPC initiative). This formed the initial population of primary care practices that could feasibly be matched to treatment practices based on practice and patient characteristics. We excluded New York City because the demographics and market characteristics there are very different from the rest of the state. We excluded primary care practices in the 6 treatment counties

because FLHSA recruited many of the practices in those counties, and the remaining practices that were not participating could systematically differ from those that were (for example, in interest in participating in practice transformation activities). Similarly, we excluded the 13 counties with relatively strong participation in federal primary care initiatives because it is likely that practices in those counties that did not participate in federal initiatives were systematically different from those that did.

2. Second, we constructed matching variables, defined before the start of the intervention for all treatment and potential comparison practices. These variables include characteristics of the practices (for example, the number of PCPs in the practice and the practice's EHR use), as well as characteristics of all Medicare FFS beneficiaries assigned to the practices (for example, mean HCC score, Medicare Part A and B spending, and utilization in the baseline period) and characteristics of high-risk beneficiaries assigned to the practices. (Section II.C.4 provides additional detail on matching data and results.) We developed a Cohort 1 and Cohort 2 version of matching variables for each potential comparison practice—with different one-year baseline periods for each version—so that they could be matched to either a Cohort 1 or Cohort 2 treatment practice. As noted earlier, the baseline period was January 2012 to December 2013 for Cohort 1 and July 2012 to June 2013 for Cohort 2.

3. Third, we narrowed the pool of potential comparison practices by excluding those practices that (1) participated in either the MAPCP Demonstration or the CPC initiative and (2) had an average of fewer than 25 assigned Medicare FFS beneficiaries during the four baseline quarters. These exclusions made the comparison pool better resemble treatment practices because none of the treatment practices participated in CPC or MAPCP, and all participating practices in the treatment group had at least 25 assigned Medicare beneficiaries during the baseline period. These restrictions left a pool of 537 potential comparison practices.

4. Fourth, we used propensity-score methods to select 108 comparison practices from the pool of 537 that were similar to the 37 treatment practices on the matching variables. The propensity score for a given practice is the predicted probability, based on all matching variables, that the practice is part of the treatment group (Stuart 2010). The score collapses information from all of the matching variables into a single number for each practice that we used to assess how similar practices are to one another. We matched each treatment practice to one or more comparison practice with a similar propensity score, with the aim of generating a comparison group that was similar, on average, to the treatment group on the matching variables (see Section II.C.4 to assess balance between treatment and comparison groups after matching).

We required each treatment practice to match to at least one, but no more than six, comparison practices and that the overall ratio of comparison to treatment panels be 3:1. This matching ratio increases the statistical certainty in the impact estimates (relative to a 1:1 overall matching ratio) because it creates a more stable comparison group against which to compare the treatment group.

After completing the matching process, we assigned Medicare FFS beneficiaries to the comparison practices in each intervention quarter using the same rules we used for the treatment group (Section V.A.1). We also defined a high-risk subgroup of the comparison group using the

same rules as for the treatment group. That is, a beneficiary was in the high-risk group in the intervention quarter if his or her HCC score at the start of the intervention period was in the top third of all observable Medicare beneficiaries assigned to the treatment panels at the start of the intervention period.

4. Construction of outcomes and covariates

We used Medicare claims from August 1, 2009, to July 31, 2015, for beneficiaries assigned to the treatment and comparison practices to develop two types of variables: (1) outcomes, defined for each person in each baseline or intervention quarter; and (2) covariates, which describe a beneficiary's characteristics at the start of the baseline and intervention periods and are used in the regression models for estimating impacts to adjust for beneficiaries' characteristics before the period began. We used covariates defined at the start of each period, without updating them each quarter, to avoid controlling in each intervention quarter for previous quarters' program effects, as this would bias the effect estimates away from detecting true impacts. Appendix 1 provides details on the methods we used to construct these variables.

Outcomes. For each person, we calculated nine outcomes that we grouped into four domains:

1. Domain: Quality-of-care processes
    a. Diabetes hemoglobin A1c (HbA1c) (binary variable for each beneficiary); calculated as whether a beneficiary with diabetes had an HbA1c test during the previous 12 months
    b. Diabetes lipid profile (binary variable for each beneficiary); calculated as whether a beneficiary with diabetes had a lipid profile during the previous 12 months
    c. Ischemic vascular disease (IVD) lipid profile (binary variable for each beneficiary); calculated as whether a beneficiary with IVD had a complete lipid profile during the previous 12 months
    d. Ambulatory-care follow-up visit within 14 days of a hospital discharge (binary variable for each beneficiary); calculated as whether all of a beneficiary's discharges in a quarter were followed by an ambulatory visit with a primary care or specialist physician within 14 days of the discharge
2. Domain: Quality-of-care outcomes
    a. Inpatient admissions (number/beneficiary/quarter) for ambulatory care-sensitive conditions (ACSCs)
    b. Number of inpatient admissions followed by an unplanned readmission within 30 days (number/beneficiary/quarter)
3. Domain: Service use
    a. All-cause inpatient admissions (number/beneficiary/quarter)
    b. Outpatient ED visit rate (number/beneficiary/quarter); outpatient ED visits are defined as ED visits or observational stays that do not end in a hospital admission

4. Domain: Spending
    a. Total Medicare Part A and B spending (dollars/beneficiary/month)

Four of these outcomes—all but ACSC admissions and the four quality-of-care process measures—are outcomes that CMMI has specified as core for the evaluations of all HCIA programs. Our definition of the readmission measure, however, differs from CMMI's standard definition. CMMI typically defines readmissions as the proportion of inpatient admissions that end in an unplanned readmission. Instead, we analyzed impacts on the *number* of these unplanned readmissions across all beneficiaries per quarter, because this enables us to look at the total impact on readmissions across the treatment group, rather than readmissions contingent on an inpatient admission. We made this decision in consultation with CMMI because the intervention might also affect the number and type of admissions.

All outcomes are quarter-specific—meaning that we calculated them for each baseline and intervention quarter separately—except for the three quality-of-care process measures for IVD and diabetes. Because these three measures assess whether a beneficiary received recommended preventive care services over a year-long period, we calculated these measures over full years rather than quarters—for example, over the baseline year (the period corresponding to the four baseline quarters), over the first year of the intervention period (corresponding to the first four intervention quarters), and so on. We avoided calculating these measures for overlapping periods, meaning that no measurement year included services provided in another measurement year.

Finally, we defined all outcomes for all treatment and comparison group members, except for the four measures of quality-of-care processes. We calculated the measure of 14-day follow-up after discharge among only those patients with at least one hospital discharge in the relevant quarter. We calculated the diabetes measures among beneficiaries ages 18 to 75 with diabetes at the beginning of the period (baseline or intervention period), and calculated the measure of lipid screening among beneficiaries ages 18 or older with IVD at the beginning of the period.

Covariates. The covariates include (1) 18 indicators for whether a patient has each of the following chronic conditions: heart failure, chronic obstructive pulmonary disease, chronic kidney disease, diabetes, Alzheimer's and related dementia, depression, ischemic heart disease, cancer, asthma, hypertension, atrial fibrillation, stroke, hyperlipidemia, hip fracture, osteoporosis, rheumatoid arthritis, bipolar disorder, and schizophrenia; (2) HCC score; (3) demographics (age, gender, and race or ethnicity); and (4) original reason for Medicare entitlement (old age, disability, or end-stage renal disease). We defined all covariates as of the start of the relevant period (baseline or intervention).

5. Regression model

We used a regression model to implement the difference-in-differences design for estimating impacts. For each outcome, the model estimates the relationship between the outcome and a series of predictor variables, assuming that each of the predictor variables has a linear (additive) relationship with the outcome. The predictor variables include the patient-level covariates (defined in Section V.A.4); whether the patient is assigned to a treatment or a comparison

practice; an indicator for each panel (which accounts for differences between panels in their patients' outcomes at baseline); indicators for each post-intervention quarter (or, for the diabetes and IVD measures, for the final post-intervention quarter of the year-long measurement period); and an interaction of a beneficiary's treatment status with each post-intervention quarter (or, for the diabetes and IVD measures, the final post-intervention quarter of the year-long measurement period).

The estimated relationship between the interaction term and the outcome in a given quarter is the impact estimate for that quarter (or, for the diabetes and IVD measures, for the year ending with that quarter). It measures the average difference between outcomes for beneficiaries assigned to the treatment and comparison practices during that period, subtracting out any differences between these groups during the four baseline quarters. By providing separate impact estimates for each intervention quarter (or year, for the diabetes and IVD measures), the model enables the program's impacts to change the longer the practices are enrolled in the program. We can also test impacts over discrete sets of quarters or years; this is needed to implement the primary tests discussed in the next section. Finally, the model quantifies the uncertainty in the impact estimates, allowing for statistical tests that determine whether observed differences in outcomes between the treatment and comparison groups are likely due to chance. The model uses robust standard errors to account for clustering of outcomes across quarters for the same beneficiary and a dummy variable for each practice (fixed effects) to account for clustering of outcomes for beneficiaries assigned to the same practice. Appendix 2 provides details on the regression methods, including descriptions of the weights each beneficiary receives in the model.

6. Primary tests

Table V.1 shows the primary tests for FLHSA, by domain. Each test specifies a population, outcome, time period, expected direction of effect, and threshold that we count as substantively important. The purpose of these primary tests is to focus the impact evaluation on hypotheses that will provide the most robust evidence about program effectiveness (see Appendix 3 for detail and a description of how we selected each test). We provided both FLHSA and CMMI an opportunity to comment on the primary tests.

Our rationale for selecting these primary tests is as follows:

- **Outcomes.** FLHSA's central goal was to reduce ED visits, 30-day unplanned readmissions, ACSC admissions, and total medical spending. FLHSA did not explicitly state that it expected to reduced all-cause hospital admissions. However, through the expected reductions in ACSC admissions, FLHSA should also reduce all-cause admissions (although as a smaller percentage change). We plan to assess program effects on all five of these outcomes. We also include four quality-of-care process measures that, based on FLHSA's theory of action (Section III.A.3) and core monitoring indicators, we think the program could improve: (1) a measure for whether a beneficiary with diabetes received an HbA1c test, (2) a measure for whether a beneficiary with diabetes received a lipid profile, (3) receipt of a complete lipid profile for people with IVD, and (4) receipt of a follow-up ambulatory care visit with a primary care or specialist provider within 14 days of hospital discharge.

Table V.1. Specification of the primary tests for FLHSA

Domain (number of tests in the domain)[a]	Outcome (units)	Time period for impacts (controlling for baseline differences)[b]	Population	Expected direction of effect (+ or -) and substantive threshold (impact as percentage of the counterfactual)[c]
Quality-of-care processes (4)	Received an HbA1c test (binary [yes or no]/beneficiary/year)	Average over intervention quarters 5 through 12 for Cohorts 1 and 2, and 5 through 8 for Cohort 3[d]	Medicare FFS beneficiaries assigned to treatment groups with diabetes and ages 18 to 75	15.0% (+)
	Received a lipid profile(binary [yes or no]/beneficiary/year)	Average over intervention quarters 5 through 12 for Cohorts 1 and 2, and 5 through 8 for Cohort 3[d]	Medicare FFS beneficiaries assigned to treatment groups with diabetes and ages 18 to 75	15.0% (+)
	Received complete lipid profile in the year (binary [yes or no]/beneficiary/year)	Average over intervention quarters 5 through 12 for Cohorts 1 and 2, and 5 through 8 for Cohort 3[d]	Medicare FFS beneficiaries assigned to treatment groups with IVD and 18 or older	15.0% (+)
	All inpatient admissions within a quarter were followed by an ambulatory care visit with a primary care or specialist provider within 14 days (binary [yes or no]/beneficiary/year)	Average over intervention quarters 5 through 14 for Cohort 1, 5 through 12 for Cohort 2, and 5 through 8 for Cohort 3	Medicare FFS beneficiaries assigned to treatment groups with at least one hospital stay in the quarter	15.0% (+)
Quality-of-care outcomes (4)	Inpatient admissions for ambulatory care-sensitive conditions (#/beneficiary/quarter)	Average over intervention quarters 5 through 14 for Cohort 1, 5 through 12 for Cohort 2, and 5 through 8 for Cohort 3	Medicare FFS beneficiaries assigned to treatment groups	5.0% (-)
	30-day unplanned hospital readmissions (#/beneficiary/quarter)	Average over intervention quarters 5 through 14 for Cohort 1, 5 through 12 for Cohort 2, and 5 through 8 for Cohort 3	Medicare FFS beneficiaries assigned to treatment groups	5.0% (-)
	Inpatient admissions for ambulatory care-sensitive conditions (#/beneficiary/quarter)	Average over intervention quarters 5 through 14 for Cohort 1, 5 through 12 for Cohort 2, and 5 through 8 for Cohort 3	High-risk Medicare FFS beneficiaries assigned to treatment groups	15.0% (-)
	30-day unplanned hospital readmissions (#/beneficiary/quarter)	Average over intervention quarters 5 through 14 for Cohort 1, 5 through 12 for Cohort 2, and 5 through 8 for Cohort 3	High-risk Medicare FFS beneficiaries assigned to treatment groups	15.0% (-)

Table V.1 *(continued)*

Domain (number of tests in the domain)[a]	Outcome (units)	Time period for impacts (controlling for baseline differences)[b]	Population	Expected direction of effect (+ or -) and substantive threshold (impact as percentage of the counterfactual)[c]
Service use (4)	All-cause inpatient admissions (#/beneficiary/quarter)	Average over intervention quarters 5 through 14 for Cohort 1, 5 through 12 for Cohort 2, and 5 through 8 for Cohort 3	Medicare FFS beneficiaries assigned to treatment groups	3.0% (-)
	Outpatient ED visit rate (#/beneficiary/quarter)	Average over intervention quarters 5 through 14 for Cohort 1, 5 through 12 for Cohort 2, and 5 through 8 for Cohort 3	Medicare FFS beneficiaries assigned to treatment groups	5.0% (-)
	All-cause inpatient admissions (#/beneficiary/quarter)	Average over intervention quarters 5 through 14 for Cohort 1, 5 through 12 for Cohort 2, and 5 through 8 for Cohort 3	High-risk Medicare FFS beneficiaries assigned to treatment groups	5.0% (-)
	Outpatient ED visit rate (#/beneficiary/quarter)	Average over intervention quarters 5 through 14 for Cohort 1, 5 through 12 for Cohort 2, and 5 through 8 for Cohort 3	High-risk Medicare FFS beneficiaries assigned to treatment groups	15.0% (-)
Spending (2)	Medicare Part A and B spending ($/beneficiary/month)	Average over intervention quarters 5 through 14 for Cohort 1, 5 through 12 for Cohort 2, and 5 through 8 for Cohort 3	Medicare FFS beneficiaries assigned to treatment groups	2.0% (-)
	Medicare Part A and B spending ($/beneficiary/month)	Average over intervention quarters 5 through 14 for Cohort 1, 5 through 12 for Cohort 2, and 5 through 8 for Cohort 3	High-risk Medicare FFS beneficiaries assigned to treatment groups	3.0% (-)

[a] We will adjust the *p*-values from the primary test results for the multiple comparisons made within each domain, but not across domains.

[b] The regression models for estimating program impacts will control for differences in outcomes between the pre-intervention treatment and comparison groups.

[c] The substantive threshold is the impact as a percentage of the counterfactual. The counterfactual is the outcome the treatment group would have had in the absence of the HCIA-funded intervention.

[d] For most measures, we will take the average across 10 quarterly impact estimates (one for each intervention quarter from 5 through 14). For the diabetes and IVD process-of-care measures, we will use two annual impact estimates—one for the second program year (corresponding to intervention quarters 5 through 8) and one for the third program year (corresponding to intervention quarters 9 through 12).

ED = emergency department; FFS = fee-for-service; FLHSA = Finger Lakes Health Systems Agency; HCIA = Health Care Innovation Award; IVD = ischemic vascular disease.

- **Time period.** FLHSA obtained a one-year extension past its original HCIA funding end date of June 30, 2015. Under this extension, FLHSA continued to implement all components of its HCIA program until the new end date of June 30, 2016. To maximize our ability to detect program impacts, we plan to analyze program impacts from early 2014 to mid-2016 among patients in Cohort 1, from mid-2014 to mid-2016 for Cohort 2, and from mid-2015 to mid-2016 for Cohort 3. This corresponds to intervention quarters 5 through 14 (I5 through I14) for Cohort 1, I5 through I12 for Cohort 2, and I5 through I8 for Cohort 3. FLHSA officials expect the program to have no effects in a practice's first year of participation, half of the maximum effect in the second year, and the full effects in the third year and beyond. By mid-2016, the first cohort had experienced 3.5 years of the intervention (with potential for the maximum effect during the last six intervention quarters), the second cohort had experienced 3.0 years of the intervention (with potential for the maximum effect during the last four intervention quarters), and the third cohort had experienced 2.0 years of the intervention (with potential for half of the maximum effect during the last four intervention quarters). Most of the measures are defined quarterly, so to estimate impacts over the specified time period, we average the impact estimates for each quarter. In contrast, the process-of-care measures for IVD and diabetes are defined over a year. Therefore, our primary tests are the average of annual impact estimates—two annual estimates for Cohorts 1 and 2 (one at the end of I8 and another at the end of I12) and one annual estimate for Cohort 3 (at the end of I8).
- **Population.** FLHSA's practice transformation and care management components should generate impacts among all patients, but the impacts of care management services are expected to be concentrated among high-risk patients. To capture potential effects on all Medicare beneficiaries as well as high-risk Medicare beneficiaries, we include both groups in our primary tests on quality-of-care outcomes, service use and spending. For the diabetes and IVD process-of-care measures, we limit the population to beneficiaries ages 18 to 75 with diabetes or ages 18 and older with IVD, respectively, and who were observable in Medicare FFS claims for all 12 months of the measurement year. For the 14-day follow-up measure, we limit the sample in each quarter to those beneficiaries who had at least one index hospitalization during the quarter for which we could observe whether the person had a 14-day follow-up visit.
- **Direction (sign) of the impact estimate.** For the quality-of-care process measures, we expect the impact estimate to be positive, signaling an increase in the percentage of people receiving recommended care. For all other outcomes, we expect the impact estimates to be negative, indicating a reduction in service use or overall expenditures.
- **Substantive thresholds.** Some impact estimates could be large enough to be substantively interesting to CMMI and other stakeholders even if they are not statistically significant; for this reason, we have specified thresholds for what we call substantive importance. We express the threshold as a percentage change from the counterfactual—that is, the outcomes that beneficiaries in the treatment group would have had if they had not received the treatment. For the full patient population, the 3 and 2 percent thresholds we chose for all-cause hospitalizations and total spending, respectively, are 75 percent of FLHSA's expected effects among all three cohorts during the primary test period (I5 through I14). (We use 75

percent recognizing that FLHSA could still be considered successful if it approached, but did not achieve, its fully anticipated effects.) The 5 percent threshold for the remaining outcomes is extrapolated from the literature (Peikes et al. 2011), which suggests that impacts of this size should be considered substantial, even though they are smaller than the impacts FLHSA anticipates. (By the third year of the intervention, the awardee expects a decrease of 25 percent in potentially preventable hospitalizations and 30-day hospital readmissions, and a decrease of 15 percent in ED visits among its full patient population.)

For the high-risk patient population, the 5 and 3 percent thresholds we chose for all-cause hospitalizations and total spending, respectively, are 75 percent of our estimate of FLHSA's expected effects among high-risk beneficiaries for all three cohorts during the primary test period (I5 through I14). This estimate is based on the percentage of high-risk beneficiaries in the population and the portion of utilization and costs for which they account relative to patients who are not at high risk. The 15 percent threshold for the remaining outcomes is extrapolated from the literature (Peikes et al. 2011) for the same reason stated earlier (that is, the literature indicates effects of this size should be considered substantial, even though they are smaller than our calculation of FLHSA's expected effects for high-risk beneficiaries).

The 15 percent threshold for the quality-of-care process measures is extrapolated from the literature (Peikes et al. 2011; Rosenthal et al. 2016) because FLHSA did not specify by how much it expected to improve these outcomes.

Because the third annual report is designed to assess impacts during the original award period only, we plan to conduct the primary tests in the report only partially. Specifically, we will estimate impacts during the 5th through 10th intervention quarters for Cohort 1 (January 2014 through June 2015) and the 5th through 8th intervention quarters for Cohort 2 (July 2014 through June 2015). Cohort 3 is not included in this analysis, as no impacts are expected for these practices during this time frame. Future reports will cover the full 30 months from quarters 5 through 14 for Cohort 1, the full 24 months from quarters 5 through 12 for Cohort 2, and the full 12 months from quarters 5 through 8 for Cohort 3.

7. Secondary tests (robustness checks)

We also conducted secondary quantitative tests to help corroborate the findings from the primary tests. This is important because some of the differences observed between the treatment and comparison groups for the primary tests could result from the non-experimental design or random fluctuations in the data. We have greater confidence in the primary results if they are generally consistent with the expected broader pattern of results.

We conducted two sets of secondary tests for FLHSA.

1. First, we estimated the program's impacts on the full Medicare FFS population and the high-risk Medicare FFS population during the first 12 months after the practices joined the intervention (quarters I1 through I4). Because we and FLHSA expect program impacts to increase over time, with little or no impacts in the first few months of the program in quality-of-care outcomes, service use, and spending, no measured effects in the first 12 intervention months would be highly consistent with an effective program. In contrast, if we

found large differences in outcomes (favorable or unfavorable) in the first 12 intervention months, this could suggest a limitation in the comparison group, not true program impacts. (It should be noted, however, that impacts in quality-of-care processes could feasibly materialize in the first 12 months of the intervention, given that care managers were hired at all treatment practices within 3 months of the intervention's start date and they began providing care management services shortly thereafter.)

2. Second, we reestimated impacts on hospital admissions and spending among the full Medicare FFS population and the high-risk Medicare FFS population, limiting the sample to beneficiaries assigned to the treatment and comparison groups by the start of the period, either baseline or intervention. This restriction prevents addition to the intervention sample over time. It is possible that differences in sample addition between the treatment and comparison groups could bias the impact results to some degree if the sample members added over time differ from earlier sample members (for example, if they are younger and healthier). This could create differences in mean outcomes between the treatment and comparison groups that are unrelated to the HCIA-funded intervention. We have explored this possibility because, as we will describe in Section V.D.1, the rate of net sample growth during the intervention period was slightly higher for the treatment group (growth of 14.2 percent from I1 to I8) than for the comparison group (growth of 12.9 percent over the same period).

8. Synthesizing evidence to draw conclusions

Within each domain, we drew one of five conclusions about program effectiveness based on the results of primary and secondary tests, and the plausibility of those findings given the implementation evidence:

1. Statistically significant favorable effect (the highest level of evidence)
2. Substantively important (but not statistically significant) favorable effect
3. Substantively important (but not statistically significant) unfavorable effect
4. No substantively large effect
5. Indeterminate effect

We cannot conclude that a program has a statistically significant unfavorable effect because, in consultation with CMMI, we decided to use one-sided statistical tests (which do not test for evidence of unfavorable effects). We used one-sided tests to increase the probability that, if a program truly did have impacts, we would be able to detect them.

Appendix 3 describes our decision rules for each of the five possible conclusions. In short, we concluded that a program had a statistically significant favorable effect in a domain if (1) at least one primary test result in the domain was favorable and statistically significant, after adjusting the statistical tests to account for multiple tests (if applicable) within a domain; or (2) the average impact estimate across all primary tests in the domain was favorable and statistically significant. In both cases, we also had to determine that the primary test results were plausible given the results of the secondary tests and implementation evidence. We concluded that a

program had a substantively important favorable effect if the average impact estimate in the domain was substantively important but not statistically significant, and if the result was plausible given the secondary tests and implementation evidence. In contrast, if the average impact estimate was unfavorable (opposite the hypothesized direction), larger than the substantive threshold, and unfavorable effects were plausible given the other evidence, we concluded the program had a substantively important unfavorable effect. If the tests in a domain did not meet any of these criteria, we drew one of two conclusions. First, if the tests for at least one outcome in the domain (or all outcomes in the domain together) had sufficient statistical power to detect an impact of the size of the substantive threshold with at least 75 percent probability, we concluded there was not a substantively large effect because we are reasonably confident that we would have detected such an effect had there been one. Alternatively, if the power was not sufficient (less than 75 percent) to detect this type of impact, we concluded the impact in the domain was indeterminate. Indeterminate means either that the program truly did not have effects that were substantively large, or that it did, but our statistical tests were unable to detect them.

B. Characteristics of the treatment group at baseline

This section describes the characteristics of the treatment group at the start of the intervention (January 1, 2013, for Cohort 1 and July 1, 2013, for Cohort 2). We also show this information in the second column of Table V.2, which shows the characteristics of all treatment practices in Cohorts 1 and 2. (Table V.2 serves a second purpose—to show the equivalence of the treatment and comparison practices in the first two cohorts at the start of the intervention—which we describe in Section V.C.)

Characteristics of the practices overall. Our analysis includes 37 treatment practices at the start of the intervention, none of which are FQHCs. Almost all treatment practices had providers receiving payment from CMS for meaningful use of EHRs (95 percent). This latter proportion is consistent with FLHSA's targeting, as one of the program's eligibility criteria was an EHR system that practice staff used actively for at least a year. Treatment practices had 6.2 total clinicians, on average. The large majority of practices' clinicians in the treatment group had a primary care specialty.

Characteristics of the practices' Medicare FFS beneficiaries. The demographic characteristics of all Medicare FFS beneficiaries assigned to the treatment group during the baseline period were, overall, comparable to nationwide Medicare FFS averages. Beneficiaries in the treatment group also had hospital and ACSC admission rates, 30-day readmission rates, and HCC scores that were comparable to national averages. However, the mean outpatient ED visit rate (143/1,000 beneficiaries/quarter) was higher than the national average of 105. In part, this might reflect the proportion of dually eligible beneficiaries in treatment practices, which, at 31 percent, is higher than the national average of 20 percent among Medicare FFS beneficiaries. People who are dually enrolled in Medicare and Medicaid tend to have higher ED rates than Medicare beneficiaries who are not dually enrolled (MedPAC 2016).

Table V.2. Characteristics of treatment and comparison panels before the intervention start date (January 1, 2013, for Cohort 1 and July 1, 2013, for Cohort 2)

Characteristic	Treatment practices (N = 37)	Unmatched comparison pool (N = 537)	Matched comparison group (N = 108)	Absolute difference[a]	Standardized difference[b]	Medicare FFS national average
Exact-match variable[c]						
Non-FQHC	100.0	100.0	100.0	0	0	n.a.
Propensity-matched variables[d]						
Characteristics of a practice's location(s)						
Located in an urban zip code (%)	86.5	81.3	90.1	-3.6	-0.10	NA
Zip code poverty rate (%)[e,f]	14.9	13.9	18.2	-3.3	-0.27	NA
Located in a health professionals shortage area (primary care)[f]	94.6	56.8	79.7	14.9	0.35	NA
Medicare Advantage penetration rate[f]	57.5	31.6	38.9	18.7	1.46	NA
Characteristics of all Medicare FFS patients attributed to practices during the baseline year (January 1, 2012 to December 31, 2012, for Cohort 1 and July 1, 2012 to June 30, 2013, for Cohort 2 practices)						
Number of beneficiaries	393.9	390.2	402.1	-8.3	-0.03	n.a.
HCC risk score	1.12	1.18	1.12	-0.01	-0.04	1.0
All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	79.0	83.6	77.7	1.4	0.06	74[g]
Outpatient ED visit rate (#/1,000 beneficiaries/quarter)	142.9	125.0	142.6	0.3	< 0.01	105[h]
Medicare Part A and B spending ($/beneficiary/month)	2,142.4	2,324.3	2,222.4	-80.0	-0.14	860[i]
30-day unplanned hospital readmission (#/1,000 beneficiaries/quarter)[i]	12.3	12.4	11.1	1.2	0.16	NA
Inpatient admissions for ACSCs (#/1,000 beneficiaries/quarter)	14.5	16.2	13.6	0.9	0.12	11.8[j]
Dually eligible beneficiaries (%)	31.2	19.3	30.5	0.7	0.04	19.9[k]
Disability as original reason for Medicare entitlement (%)	43.5	28.9	40.4	3.1	0.17	16.7[l]
Age (years)	67.0	71.4	67.2	-0.2	-0.03	71[m]
Female (%)	59.6	58.2	58.3	1.3	0.16	54.7[l]
Race: white	81.2	88.5	82.4	-1.2	-0.06	81.8[l]
Receipt of recommended lipid profile, among those with diabetes ages 18 to 75 (%)	84.6	86.1	84.2	0.5	0.04	NA

Table V.2 *(continued)*

Characteristic	Treatment practices (N = 37)	Unmatched comparison pool (N = 537)	Matched compar-ison group (N = 108)	Absolute difference[a]	Standard-ized difference[b]	Medicare FFS national average
Receipt of recommended hemoglobin A1c test, among those with diabetes ages 18 to 75 (%)	89.8	88.7	90.1	-0.3	-0.04	NA
Receipt of recommended lipid profile, among those with IVD ages 18 or older (%)	79.3	81.6	80.1	-0.8	-0.08	NA
Receipt of an ambulatory care visit within 14 days of any hospital discharges in the quarter, among those with at least one discharge in the quarter (%)	67.9	62.1	67.4	0.5	0.06	NA
Characteristics of high-risk Medicare FFS patients attributed to practices during the baseline year (January 1, 2012 to December 31, 2012 for Cohort 1 and July 1, 2012 to June 30, 2013 for Cohort 2 practices)						
Number of high-risk beneficiaries	94.1	98.1	94.0	0.1	< 0.01	n.a.
HCC risk score	2.25	2.34	2.27	-0.01	-0.06	n.a.
All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	174.3	182.9	177.8	-3.4	-0.07	74[g]
Outpatient ED visit rate (#/1,000 patients/quarter)	235.7	203.3	227.5	8.1	0.07	105[h]
Medicare Part A and B spending ($/beneficiary/ month)	4487.3	4704.3	4576.5	-89.2	-0.08	860[i]
30-day unplanned hospital readmission (#/1,000 beneficiaries/quarter)	33.1	33.5	32.2	0.9	0.04	NA
Inpatient admissions for ACSCs (#/beneficiary/ quarter)	37.5	42.5	38.5	-1.1	-0.06	11.8[j]
Characteristics of the practices						
Meaningful use of EHR (%)[n]	94.6	68.2	93.3	1.3	0.04	n.a.
Patient-centered medical home[o]	10.8	7.3	9.6	1.2	0.04	
Owned by hospital or health system (%)	56.8	29.0	54.0	2.7	0.05	n.a.
Number of clinicians at practice	6.2	3.1	6.0	0.2	0.04	n.a.
Practices' clinicians with a primary care specialty (%)	93.9	93.2	92.1	1.8	0.10	n.a.

Sources: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at CMS. Zip code household income data merged from the American Community Survey ZIP Code Characteristics. Characteristics of the practices come from SK&A, a health care data vendor, and the National Committee for Quality Assurance.

Table V.2 *(continued)*

Notes: The characteristics for the treatment and their matched comparison practices are defined at the time the treatment practice joined the intervention (January 1, 2013, for Cohort 1 practices and June 1, 2013, for Cohort 2 practices).

The comparison group means are weighted based on the number of matched comparison practices per treatment practice. For example, if four comparison practices are matched to one treatment practice, each of the four comparison practices has a matching weight of 0.25.

Absolute differences might not be exact due to rounding.

[a] The absolute difference is the difference in means between the matched treatment and comparison groups.
[b] The standardized difference is the difference in means between the matched treatment and comparison groups divided by the standard deviation of the variable, which is pooled across the matched treatment and selected comparison groups.
[c] Exact match means that we required that non-FQHCs match only to non-FQHCs.
[d] Variables that we matched on through a propensity score, which captures the relationship between a practice's characteristics and its likelihood of being in the treatment group.
[e] Average poverty rate associated with each practice's zip code, merged from the American Community Survey.
[f] These variables were not included in the propensity-score model due to concerns that they would generate potential imbalances among the critical matching variables; crucial matching variables include all variables on patient and practice characteristics in this table.
[g] Health Indicators Warehouse (2014b).
[h] Gerhardt et al. (2014).
[i] Boards of Trustees (2013).
[j] This rate is for beneficiaries ages 65 and older (Truven Health Analytics 2015).
[k] MedPAC (2016).
[l] Chronic Conditions Data Warehouse (2014a, Table A.1).
[m] Health Indicators Warehouse (2014b).
[n] Meaningful use of EHRs is calculated as the percentage of practices with at least one provider (NPI) working in the practice who received financial incentives for meaningful use of certified EHRs through Medicare or Medicaid during the baseline period. Data on meaningful use of EHRs were merged from CMS data.
[o] NCQA Patient-Centered Medical Home (PCMH) Recognition. Data on practices with NCQA recognition were merged from the NCQA database.
ACSC = ambulatory care-sensitive condition; CMS = Centers for Medicare & Medicaid Services; ED = emergency department; EHR = electronic health record; FFS = fee-for-service; FQHC = federally qualified health center; HCC = Hierarchical Conditions Category; IVD = ischemic vascular disease; NCQA = National Committee for Quality Assurance; NPI = National Provider Identifier.
NA = not available.
n.a. = not applicable.

Characteristics of the practices' high-risk Medicare FFS beneficiaries. The high-risk beneficiaries in the treatment group had substantially greater health care needs during the baseline period than the full treatment group (Table V.2). Their mean HCC risk score was more than twice the mean for all treatment group members (2.3 versus 1.1), consistent with how the group was defined. Further, they had more than twice the all-cause inpatient admissions and Medicare spending than the full population of attributed beneficiaries.

C. Equivalence of treatment and comparison groups at baseline

Demonstrating that the treatment and comparison groups are similar at the start of the intervention is important for the evaluation design. This similarity increases the credibility of a key assumption underlying difference-in-differences models—that the change over time in outcomes for the comparison group is the same change that would have happened for the treatment group, had the treatment group not received the intervention.

Table V.2 shows that the 37 treatment practices and the 108 selected comparison practices were similar at the start of the intervention on most matching variables. By construction, there were no differences between the two groups on the exact matching variable—whether the practice was an FQHC. There were some differences between treatment and matched comparison group beneficiaries on the variables we matched through propensity scores (the second panel of Table V.2), but the standardized differences across the propensity-score matching variables are all within our target of 0.25 standardized differences, and most were within 0.15 standardized differences (the 0.25 target is an industry standard; for example, see Institute of Education Sciences [2014]). This includes patients' demographic characteristics, Medicare FFS beneficiaries' and high-risk Medicare FFS beneficiaries' utilization and costs, as well as four process-of-care measures upon which practices were matched in mid-2016. Similarly, all differences between treatment and comparison group practice characteristics (the third panel of Table V.2) are within our target of 0.25 standardized differences. This includes practices' EHR use, medical home designation, ownership, and number of clinicians.

However, there are three treatment–comparison differences in the characteristics of practices' locations (the first panel of Table V.2). Namely, the average poverty rate of zip codes in which treatment practices are located is lower than that of comparison practices (14.9 versus 18.2 percent in comparison) and the Medicare Advantage penetration rate in counties in which treatment practices are located is higher than that of comparison practices. In addition, a higher proportion of treatment practices are located in a health professionals shortage area than comparison practices (95 versus 80 percent of comparison practices). The difference-in-differences impact estimation model accounts for these baseline treatment–comparison differences. However, given differential Medicare penetration rates between treatment and comparison groups, it is important to conduct a sensitivity test that prohibits sample addition during the intervention period (Section V.A.7). This sensitivity test corrects for any bias that could result from differential sample addition to treatment and comparison groups over time (although it cannot correct for bias that might result from differential sample attrition to Medicare managed care).

We also separately assessed balance among the Cohort 1 practices (16 treatment and 51 comparison practices), because these practices changed from a subgroup of practices to the full set of practices in later quarters. Specifically, because only Cohort 1 practices can be followed up for I9 and I10 for this report, those are the only practices in the treatment and comparison groups in those two intervention quarters. It is important to show that the Cohort 1 treatment and comparison practices are balanced at baseline so that regression-adjusted differences in I9 and I10 can be interpreted as program impacts. Among Cohort 1 practices, there were some differences between the treatment and matched comparison group beneficiaries and practices on the variables we matched through propensity scores, but the standardized differences across the propensity-score matching variables are all within our target of 0.25 standardized differences, with the exception of whether practices are located in a health professionals shortage area (100 percent of treatment practices percent versus 79 percent of comparison practices; data not shown).

D. Beneficiaries' outcomes and intervention impacts

In this section, we first present sample sizes and mean outcomes, by quarter, for the treatment and comparison groups. These mean outcomes provide context for understanding the difference-in-differences estimates that follow; however, the differences in mean outcomes are not regression-adjusted and not impact estimates by themselves. Next, we present the results of the primary tests, by domain. Then, we present the results of the secondary tests (robustness checks) and assess whether the primary test results are plausible given the secondary test results and the implementation evidence. We end with conclusions about program impacts in each domain. These conclusions for FLHSA are preliminary because this report covers outcomes only through the end of the original award period (June 2015), whereas FLHSA's HCIA-funded intervention ran through June 2016, as described previously.

1. Sample sizes

The sample sizes for impact estimation differ depending on the outcome and intervention quarters. Notable across all outcomes that are calculated on a quarterly basis, sample sizes in the four baseline quarters and the first eight intervention quarters represent all 37 Cohort 1 and Cohort 2 treatment practices and their 108 matched comparison practices, whereas sample sizes in I9 and I10 represent only the 16 Cohort 1 practices and their 51 matched comparison practices. We present sample sizes by domain.

Quality-of-care processes

- The **HbA1c** and **lipid profile measures for people with diabetes** are defined among Medicare FFS beneficiaries with diabetes ages 18 to 75. The sample size for the treatment group and the weighted comparison group ranged from 1,963 to 2,434 across the baseline year and each of the two intervention years (Table V.3). This population accounted for about 16 percent of the total Medicare FFS sample in the treatment and comparison groups.
- The **lipid profile measure for people with IVD** is defined among Medicare FFS beneficiaries with IVD ages 18 or older. The sample size for the treatment group and the weighted comparison group ranged from 2,732 to 3,777 across the baseline year and each of the two intervention years (Table V.3). This population accounted for about 19 percent of the total Medicare FFS sample in the treatment and comparison groups. This percentage is higher than for the diabetes measure because (1) IVD (which is a broad disease category) is more common than diabetes among the treatment and comparison beneficiaries and (2) the diabetes measure excludes beneficiaries older than 75 but the IVD measure does not.
- The **14-day follow-up measure** is defined among Medicare FFS beneficiaries who had at least one hospital stay in the quarter. For the treatment group, the sample size ranged from 857 to 1,014 beneficiaries across the four baseline and first eight intervention quarters, accounting for about 6 percent of all treatment beneficiaries in each quarter (Table V.3). For the comparison group, the sample ranged from 2,091 to 2,432 across the four baseline and first eight intervention quarters (accounting for a similar proportion of the total comparison group). After weighting the comparison group to account for the larger number of comparison practices than treatment practices and for the difference in practice size between

treatment and comparison groups, treatment group sample sizes were similar, but slightly larger than, those in the treatment group: an average of 938 beneficiaries in the treatment group compared with 866 beneficiaries in the comparison group during the baseline and first eight intervention quarters. However, sample sizes dropped to an average of 666 and 556 in the treatment and comparison groups, respectively, in I9 and I10, as only Cohort 1 treatment and matched comparison practices are represented in these two quarters.

Quality-of-care outcomes, service use, and spending: all Medicare FFS beneficiaries. The sample sizes for all outcomes in these three domains were the same for all Medicare FFS beneficiaries. In the first baseline quarter (B1), the treatment group included 14,096 beneficiaries assigned to the 37 participating practices and the comparison group included 35,606 beneficiaries assigned to the 108 comparison practices (Table V.4). The sample sizes increased modestly during the four baseline quarters (by 6.3 percent from B1 to B4 for the treatment group and 7.7 percent for the comparison group). This net increase indicates that sample addition (due to beneficiaries being newly attributed to the treatment or comparison practices) exceeds sample attrition (due to beneficiaries dying, switching from Medicare FFS to managed care, moving out of state, or enrolling in Medicaid in addition to Medicare). The sample sizes dropped modestly from the last baseline quarter to the first intervention quarter, reflecting that the sample definition (Section V.A.1) retains sample members in successive baseline and intervention quarters, even if they are no longer attributed to the treatment or comparison panel, but not between the baseline and intervention periods. The sample increased modestly from I1 to I8, again reflecting greater sample addition than attrition over time. The net sample increase from I1 to I8 was slightly higher in the treatment group (14.2 percent) than the comparison group (12.9 percent). Both treatment and comparison group samples decreased by at least 30 percent from I8 to I9, reflecting the drop from 37 to 16 treatment practices and from 108 to 51 matched comparison practices.

Quality-of-care outcomes, service use, and spending: high-risk Medicare FFS beneficiaries. The sample sizes for all outcomes in these three domains were the same for high-risk Medicare FFS beneficiaries. In the first baseline quarter (B1), the treatment group included 3,585 high-risk beneficiaries assigned to the 37 participating practices and the comparison group included 9,231 beneficiaries assigned to the 108 comparison practices (Table V.5). The sample sizes decreased modestly during the four baseline quarters (by 6.1 percent for the treatment group and 3.5 percent for the comparison group from B1 to B4). The net sample decrease during the intervention period was slightly larger for the treatment group (15.3 percent from I1 to I8) than the comparison group (12.3 percent over the same time period). Both treatment and comparison group samples of high-risk Medicare FFS beneficiaries decreased by more than 40 percent from I8 to I9, reflecting the drop from 37 to 16 treatment practices and from 108 to 51 matched comparison practices.

Table V.3. Unadjusted mean outcomes (quality-of-care processes) observed among select Medicare FFS beneficiaries, by treatment status and quarter

		Number of Medicare FFS beneficiaries (practices)			Mean outcomes		
Period	Quarter(s)	T	C (not weighted)	C (weighted)	T	C	Difference (%)
Among those with diabetes and ages 18 to 75, received A1c screening (binary [yes or no]/beneficiary/year)							
Baseline	B1–B4[a]	2,228 (37)	5,904 (108)	2,434	90.4	90.0	0.4 (0.4%)
Intervention	I1–I4[a]	2,124 (37)	5,527 (108)	2,322	91.1	89.3	1.7 (2.0%)
	I5–I8[a]	1,963 (37)	5,141 (108)	2,149	89.5	88.7	0.8 (0.9%)
Among those with diabetes and ages 18 to 75, received lipid panel (binary [yes or no]/beneficiary/year)							
Baseline	B1–B4[a]	2,228 (37)	5,904 (108)	2,434	84.1	85.9	-1.8 (-2.1%)
Intervention	I1–I4[a]	2,124 (37)	5,527 (108)	2,322	84.3	85.2	-0.9 (-1.1%)
	I5–I8[a]	1,963 (37)	5,141 (108)	2,149	82.3	82.3	-0.0 (-0.0%)
Among those with IVD and ages 18 or older, received lipid panel (binary [yes or no]/beneficiary/year)							
Baseline	B1–B4[a]	3,217 (37)	9,779 (108)	3,777	78.8	78.5	0.2 (0.3%)
Intervention	I1–I4[a]	2,967 (37)	9,111 (108)	3,587	78.2	78.7	-0.5 (-0.6%)
	I5–I8[a]	2,732 (37)	8,581 (108)	3,329	75.8	76.5	-0.7 (-0.9%)
Among those with at least one inpatient admission in the quarter, all inpatient admissions in the quarter were followed by an ambulatory care visit with a primary care or specialist provider within 14 days of discharge (binary [yes or no]/beneficiary/year)							
Baseline	B1	925 (37)	2,139 (108)	813	66.3	66.8	-0.5 (-0.8%)
	B2	936 (37)	2,316 (108)	895	67.1	67.9	-0.8 (-1.2%)
	B3	890 (37)	2,293 (108)	856	69.1	66.8	2.3 (3.4%)
	B4	1,014 (37)	2,432 (108)	901	71.2	65.7	5.5 (8.4%)

Table V.3 *(continued)*

Period	Quarter(s)	Number of Medicare FFS beneficiaries (practices) T	C (not weighted)	C (weighted)	Mean outcomes T	C	Difference (%)
Intervention	I1	857 (37)	2,091 (108)	813	72.1	69.2	2.9 (4.2%)
	I2	923 (37)	2,136 (108)	877	74.1	66.3	7.8 (11.7%)
	I3	914 (37)	2,098 (108)	858	71.4	63.3	8.2 (12.9%)
	I4	933 (37)	2,224 (108)	866	72.5	63.9	8.6 (13.4%)
	I5	961 (37)	2,231 (108)	888	72.1	68.5	3.6 (5.3%)
	I6	906 (37)	2,268 (108)	865	70.4	64.1	6.4 (9.9%)
	I7	1,006 (37)	2,319 (108)	896	70.3	67.2	3.0 (4.5%)
	I8	985 (37)	2,311 (108)	869	71.6	65.3	6.3 (9.7%)
	I9	679 (16)	1,172 (51)	533	67.0	68.1	-1.1 (-1.6%)
	I10	653 (16)	1,240 (51)	579	72.9	64.5	8.4 (13.1%)

Sources: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at the Centers for Medicare & Medicaid Services.

Notes: The baseline quarters are measured relative to the start of the baseline period on January 1, 2012, for Cohort 1 and July 1, 2012, for Cohort 2. For example, the first baseline quarter (B1) for Cohort 1 runs from January 1, 2012, to March 31, 2012. The intervention quarters are measured relative to the start of the intervention period on January 1, 2013, for Cohort 1 and July 1, 2013, for Cohort 2. For example, the first intervention quarter for Cohort 1 (I1) runs from January 1, 2013, to March 31, 2013. In each period (baseline or intervention), the treatment group each quarter includes all beneficiaries who were assigned to a treatment panel by the start of the quarter and who met other sample criteria—that is, they were enrolled in FFS Medicare and were living in New York or surrounding areas. In each period, the comparison group includes all beneficiaries who were assigned to a comparison panel by the start of the quarter and who met the other sample criteria. See text for details.

The outcome means were weighted such that (1) each treatment beneficiary gets a weight of 1; and (2) each comparison beneficiary gets a weight that is the product of two weights: (a) a matching weight, equal to the reciprocal of the total number of comparison panels matched to the same treatment panel as the beneficiary's assigned panel, and (b) a practice size weight, which equals the average number of beneficiaries assigned to the matched treatment panel during the four baseline quarters divided by the average number of beneficiaries assigned to the beneficiary's comparison panel over those quarters. The difference between the treatment and comparison groups in a quarter is calculated by subtracting the mean outcome for the comparison group from the mean outcome for the treatment group. The percentage difference equals that difference divided by the mean outcome for the comparison group.

[a] The quality-of-care process measures were calculated over year-long periods, corresponding to the baseline and intervention quarters shown in the table.

B = baseline; C = comparison; FFS = fee-for-service; I = intervention; IVD = ischemic vascular disease; T = treatment.

Table V.4. Unadjusted mean outcomes (quality-of-care outcomes, service use, and spending) measured for all Medicare FFS beneficiaries, by treatment status and quarter

	Number of Medicare FFS beneficiaries (panels)			Inpatient admissions for ACSCs (#/1,000 beneficiaries/quarter)			30-day unplanned hospital readmissions (#/1,000 beneficiaries/quarter)			All-cause inpatient admissions (#/1,000 beneficiaries/quarter)			Outpatient ED visit rate (#/1,000 beneficiaries/quarter)			Medicare Part A and B spending ($/beneficiary/month)		
Q	T	C (no wgt)	C (wgt)	T	C	Diff (%)	T	C	Diff (%)	T	C	Diff (%)	T	C	Diff (%)	T	C	Diff (%)
Baseline period (1/1/12–12/31/12 for Cohort 1 and 7/1/12–6/30/13 for Cohort 2 practices)																		
B1	14,096 (37)	35,606 (108)	13,989	16.5	12.2	4.3 (35.3%)	12.0	11.2	0.8 (7.2%)	83.6	76.7	6.9 (9.0%)	147.3	139.5	7.8 (5.6%)	$703	$686	$17 (2.5%)
B2	14,556 (37)	36,968 (108)	14,462	15.1	14.5	0.6 (3.9%)	13.7	11.1	2.6 (23.1%)	84.8	79.3	5.5 (7.0%)	159.7	142.2	17.5 (12.3%)	$768	$754	$14 (1.8%)
B3	14,636 (37)	37,536 (108)	14,733	14.6	12.7	1.9 (15.0%)	14.3	9.7	4.6 (46.8%)	78.3	73.9	4.4 (5.9%)	154.4	139.7	14.6 (10.5%)	$723	$742	$-19 (-2.6%)
B4	14,986 (37)	38,350 (108)	15,090	16.5	13.5	3.0 (21.9%)	12.9	12.2	0.7 (5.8%)	87.2	78.0	9.2 (11.8%)	147.6	133.9	13.7 (10.2%)	$781	$763	$17 (2.3%)
Intervention period (1/1/2013–6/30/2015 for Cohort 1 and 7/1/2013–6/30/2015 for Cohort 2 practices																		
I1	13,692 (37)	35,279 (108)	14,221	16.7	13.1	3.5 (26.9%)	10.7	10.5	0.3 (2.7%)	78.5	75.2	3.3 (4.4%)	144.3	135.6	8.7 (6.4%)	$725	$736	$-11 (-1.5%)
I2	14,180 (37)	36,507 (108)	14,710	14.3	15.1	-0.8 (-5.4%)	11.9	11.6	0.3 (2.7%)	81.9	77.3	4.6 (5.9%)	158.3	145.8	12.5 (8.6%)	$784	$759	$25 (3.3%)
I3	14,364 (37)	36,872 (108)	14,851	14.0	13.4	0.6 (4.1%)	12.9	10.3	2.6 (25.4%)	81.0	75.1	5.8 (7.8%)	157.4	144.1	13.3 (9.2%)	$759	$747	$12 (1.6%)
I4	14,680 (37)	37,699 (108)	15,188	15.5	12.6	3.0 (23.6%)	13.1	11.6	1.6 (13.6%)	81.1	75.4	5.8 (7.7%)	163.3	142.2	21.0 (14.8%)	$795	$770	$25 (3.3%)
I5	14,769 (37)	38,057 (108)	15,220	17.1	12.6	4.5 (36.0%)	13.4	11.9	1.5 (12.6%)	84.0	76.1	8.0 (10.5%)	159.5	133.3	26.2 (19.7%)	$816	$757	$60 (7.9%)
I6	15,015 (37)	38,771 (108)	15,452	15.7	13.5	2.2 (16.1%)	13.0	9.6	3.4 (35.2%)	79.3	71.9	7.3 (10.2%)	165.6	149.5	16.2 (10.8%)	$804	$784	$21 (2.6%)
I7	15,263 (37)	39,067 (108)	15,570	15.3	13.4	1.9 (13.9%)	14.5	11.4	3.1 (27.2%)	87.0	77.1	9.9 (12.9%)	169.8	153.7	16.0 (10.4%)	$838	$808	$30 (3.7%)
I8	15,638 (37)	39,836 (108)	15,854	14.3	11.5	2.7 (23.5%)	12.1	10.2	1.8 (18.0%)	79.7	71.6	8.2 (11.4%)	164.7	151.0	13.7 (9.1%)	$811	$815	$-4 (-0.5%)
I9	9,271 (16)	19,652 (51)	9,261	18.4	13.9	4.5 (32.4%)	17.8	12.9	4.9 (37.8%)	97.8	78.9	19.0 (24.0%)	167.9	159.3	8.5 (5.4%)	$854	$821	$33 (4.1%)
I10	9,483 (16)	19,904 (51)	9,416	16.8	14.8	2.0 (13.4%)	15.6	14.9	0.7 (4.4%)	88.8	83.5	5.3 (6.4%)	191.3	171.5	19.8 (11.5%)	$894	$937	$-43 (-4.6%)

Source: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at the Centers for Medicare & Medicaid Services.

Notes: The baseline quarters are measured relative to the start of the baseline period on January 1, 2012, for Cohort 1 and July 1, 2012, for Cohort 2. For example, the first baseline quarter (B1) for Cohort 1 runs from January 1, 2012, to March 31, 2012. The intervention quarters are measured relative to the start of the intervention period on January 1, 2013, for Cohort 1 and July 1, 2013, for Cohort 2. For example, the first intervention quarter for Cohort 1 (I1) runs from January 1, 2013, to March 31, 2013. In each period (baseline or intervention), the treatment group each quarter includes all beneficiaries who were assigned to a treatment panel by the start of the quarter and who met other sample criteria—that is, they were enrolled in FFS Medicare and were living in New York or surrounding areas. In each period, the comparison group includes all beneficiaries who were assigned to a comparison panel by the start of the quarter and who met the other sample criteria. See text for details.

Table V.4 *(continued)*

The outcome means were weighted such that (1) each treatment beneficiary gets a weight of 1; and (2) each comparison beneficiary gets a weight that is the product of two weights: (a) a matching weight, equal to the reciprocal of the total number of comparison panels matched to the same treatment panel as the beneficiary's assigned panel, and (b) a practice size weight, which equals the average number of beneficiaries assigned to the matched treatment panel during the four baseline quarters divided by the average number of beneficiaries assigned to the beneficiary's comparison panel over those quarters. The difference between the treatment and comparison groups in a quarter is calculated by subtracting the mean outcome for the comparison group from the mean outcome for the treatment group. The percentage difference equals that difference divided by the mean outcome for the comparison group.

ACSC = ambulatory care-sensitive condition; B = baseline; C = comparison; Diff = difference; ED = emergency department; FFS = fee-for-service; I = intervention; Q = quarter; T = treatment; no wgt = unweighted; wgt = weighted.

Table V.5. Unadjusted mean outcomes (quality-of-care outcomes, service use, and spending) measured for high-risk Medicare FFS beneficiaries, by treatment status and quarter

	Number of Medicare FFS beneficiaries (panels)			Inpatient admissions for ACSCs (#/1,000 beneficiaries/quarter)			30-day unplanned hospital readmissions (#/1,000 beneficiaries/quarter)			All-cause inpatient admissions (#/1,000 beneficiaries/quarter)			Outpatient ED visit rate (#/1,000 beneficiaries/quarter)			Medicare Part A and B spending ($/beneficiary/month)		
Q	T	C (no wgt)	C (wgt)	T	C	Diff (%)	T	C	Diff (%)	T	C	Diff (%)	T	C	Diff (%)	T	C	Diff (%)
Baseline period (1/1/12–12/31/12 for Cohort 1 and 7/1/12–6/30/13 for Cohort 2 practices)																		
B1	3,585 (37)	9,231 (108)	3,533	41.0	37.8	3.2 (8.5%)	36.3	34.4	1.9 (5.4%)	192.5	184.6	7.9 (4.3%)	235.0	215.5	19.5 (9.1%)	$1,521	$1,512	$9 (0.6%)
B2	3,556 (37)	9,218 (108)	3,535	45.8	39.7	6.1 (15.5%)	36.0	32.8	3.2 (9.7%)	196.0	186.3	9.7 (5.2%)	262.4	225.5	36.9 (16.4%)	$1,652	$1,558	$94 (6.1%)
B3	3,428 (37)	9,039 (108)	3,456	33.8	32.5	1.4 (4.3%)	37.3	27.1	10.2 (37.6%)	170.7	165.9	4.8 (2.9%)	241.0	226.6	14.3 (6.3%)	$1,501	$1,512	$-11 (-0.7%)
B4	3,365 (37)	8,902 (108)	3,409	47.0	41.3	5.6 (13.6%)	30.9	37.2	-6.3 (-16.9%)	180.1	174.3	5.8 (3.3%)	230.0	206.1	23.9 (11.6%)	$1,496	$1,540	$-44 (-2.9%)
Intervention period (1/1/2013–6/30/2015 for Cohort 1 and 7/1/2013–6/30/2015 for Cohort 2 practices																		
I1	3,551 (37)	9,252 (108)	3,644	46.2	35.3	10.9 (30.9%)	30.1	27.1	3.1 (11.4%)	179.1	164.4	14.7 (8.9%)	218.2	230.0	-11.8 (-5.1%)	$1,527	$1,631	$-104 (-6.4%)
I2	3,513 (37)	9,186 (108)	3,606	36.2	43.3	-7.2 (-16.5%)	31.3	31.2	0.2 (0.5%)	184.2	175.6	8.6 (4.9%)	252.4	247.0	5.3 (2.2%)	$1,632	$1,554	$77 (5.0%)
I3	3,425 (37)	8,975 (108)	3,528	35.0	36.3	-1.3 (-3.6%)	34.7	28.5	6.2 (21.9%)	174.9	167.3	7.6 (4.5%)	247.4	236.8	10.6 (4.5%)	$1,532	$1,472	$60 (4.1%)
I4	3,357 (37)	8,802 (108)	3,470	40.8	32.4	8.4 (25.8%)	34.3	28.8	5.4 (18.8%)	184.7	164.3	20.4 (12.4%)	276.3	237.7	38.6 (16.2%)	$1,632	$1,533	$99 (6.4%)
I5	3,260 (37)	8,597 (108)	3,363	44.2	34.4	9.8 (28.4%)	39.0	30.4	8.6 (28.2%)	188.3	164.7	23.6 (14.3%)	239.7	211.0	28.7 (13.6%)	$1,728	$1,542	$185 (12.0%)
I6	3,178 (37)	8,430 (108)	3,279	39.3	38.1	1.2 (3.2%)	32.4	22.9	9.5 (41.6%)	166.8	147.2	19.6 (13.3%)	251.6	241.5	10.1 (4.2%)	$1,574	$1,488	$86 (5.8%)
I7	3,082 (37)	8,246 (108)	3,209	42.2	40.1	2.1 (5.1%)	38.6	34.4	4.2 (12.1%)	186.2	175.1	11.2 (6.4%)	262.2	258.1	4.2 (1.6%)	$1,632	$1,636	$-4 (-0.2%)
I8	3,008 (37)	8,112 (108)	3,167	36.6	32.3	4.3 (13.2%)	32.6	25.2	7.4 (29.2%)	169.9	155.3	14.6 (9.4%)	277.4	246.3	31.1 (12.6%)	$1,596	$1,669	$-73 (-4.4%)
I9	1,669 (16)	3,770 (51)	1,790	54.5	39.6	14.9 (37.8%)	56.9	35.3	21.6 (61.3%)	239.7	169.3	70.4 (41.6%)	267.2	275.8	-8.5 (-3.1%)	$1,794	$1,684	$111 (6.6%)
I10	1,624 (16)	3,666 (51)	1,749	45.6	48.6	-3.0 (-6.2%)	40.6	39.9	0.7 (1.8%)	190.9	189.8	1.1 (0.6%)	328.2	278.2	50.0 (18.0%)	$1,782	$1,933	$-151 (-7.8%)

Source: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at the Centers for Medicare & Medicaid Services.

Table V.5 *(continued)*

Notes: The baseline quarters are measured relative to the start of the baseline period on January 1, 2012, for Cohort 1 and July 1, 2012, for Cohort 2. For example, the first baseline quarter (B1) for Cohort 1 runs from January 1, 2012, to March 31, 2012. The intervention quarters are measured relative to the start of the intervention period on January 1, 2013, for Cohort 1 and July 1, 2013, for Cohort 2. For example, the first intervention quarter for Cohort 1 (I1) runs from January 1, 2013, to March 31, 2013. In each period (baseline or intervention), the treatment group each quarter includes all beneficiaries who were assigned to a treatment panel by the start of the quarter and who met other sample criteria—that is, they were enrolled in FFS Medicare and were living in New York or surrounding areas. In each period, the comparison group includes all beneficiaries who were assigned to a comparison panel by the start of the quarter and who met the other sample criteria. See text for details.

ACSC = ambulatory care-sensitive condition; B = baseline; C = comparison; Diff = difference; ED = emergency department; FFS = fee-for-service; I = intervention; Q = quarter; T = treatment; no wgt = unweighted; wgt = weighted.

2. Mean outcomes for the treatment and comparison groups, by domain and quarter

Quality-of-care processes. During the baseline year, 90.4 percent of treatment and 90.0 percent of comparison beneficiaries with diabetes and ages 18 to 75 received an HbA1c test (Table V.3). This percentage decreased slightly to 89.5 at the end of the second program year for the treatment group and declined to 88.7 for the comparison group. During the baseline year, 84.1 percent of treatment and 85.9 percent of comparison beneficiaries with diabetes and ages 18 to 75 received a lipid test (Table V.3). This percentage decreased slightly to 82.3 at the end of the second program year for the treatment and comparison groups.

During the baseline year, 78.8 and 78.5 percent of the treatment and comparison beneficiaries, respectively, ages 18 or older with IVD received the recommended lipid test (Table V.3). This percentage decreased to 75.8 in the second program year for the treatment group and declined to 76.5 for the comparison group.

In the first baseline quarter, about 66 and 67 percent of beneficiaries in treatment and comparison groups, respectively, who had any hospital stay in a baseline quarter had all of those stays followed by an ambulatory care visit within 14 days of discharge (Table V.3). This percentage increased for the treatment group during the subsequent five quarters (but not the comparison group), such that by I2 the rate was 74.1 percent for the treatment group but only 66.3 for the comparison group. However, the rate declined in both treatment and control groups from I3 to I8, resulting in rates of 71.6 and 65.3 percent in treatment and comparison, respectively, by I8. The rate increased again among Cohort 1 treatment practices from I9 to I10 (from 67.0 to 72.9 percent), while decreasing among Cohort 1 matched comparison practices during the same time period.

Quality-of-care outcomes. Among all Medicare FFS beneficiaries attributed to practices, the number of ACSC admissions fluctuated from I1 to I8 between the range of 14.0 and 17.1 per 1,000 beneficiaries in the treatment group and 11.5 and 15.1 per 1,000 beneficiaries in the comparison group (Table V.4). There was no distinguishable trend in either treatment or comparison groups during the intervention period. In addition, ACSC admissions were largely comparable among high-risk beneficiaries in the treatment and comparison groups during the baseline and the intervention periods (Table V.5).

For both the treatment and comparison groups, the number of 30-day unplanned readmissions among all Medicare FFS beneficiaries fluctuated from I1 to I8 between 10.7 and 14.5 per 1,000 beneficiaries in the treatment group and 9.6 and 11.9 per 1,000 beneficiaries in the comparison group (Table V.4). There was no distinguishable trend in either treatment or comparison groups during these quarters. However, in I9, readmissions increased to 17.8 per 1,000 beneficiaries in the treatment group, compared with only 12.9 per 1,000 beneficiaries in the comparison group. This increase was driven by a sharp increase in readmissions among high-risk beneficiaries in Cohort 1 practices in the group during I9 (56.9 per 1,000 beneficiaries in the treatment group versus only 35.3 per 1,000 beneficiaries in the comparison group; Table V.5).

Service use. All-cause inpatient admissions fluctuated over time and were generally similar between the treatment and comparison groups from I1 to I8, with all Medicare FFS beneficiaries

in the treatment group having moderately higher rates than all Medicare FFS beneficiaries in the comparison group (Table V.4). However, in I9, admissions jumped to 97.8 per 1,000 beneficiaries in the treatment group, compared with 78.9 per 1,000 beneficiaries in the comparison group. This increase was driven by an increase in admissions among high-risk beneficiaries in Cohort 1 practices in the group during I9 (239.7 for the treatment group versus 169.3 for the comparison group; Table V.5).

Outpatient ED visit rates among all Medicare FFS beneficiaries generally increased for the treatment and the comparison group from I1 to I10, with no distinguishable trends between the two groups (Table V.4). A similar phenomenon occurred among high-risk beneficiaries (Table V.5).

Spending. Mean Medicare Part A and B spending among all Medicare FFS beneficiaries generally increased for the treatment group and the comparison group from I1 to I8, but the increase was slightly larger for the treatment group (4.5 percent versus 2.3 percent in the comparison group; Table V.4). However, there was no distinguishable trend between treatment and comparison group spending for high-risk beneficiaries during this time (Table V.5). Similarly, there are no distinguishable trends between beneficiaries in Cohort 1 treatment and matched comparison practices in I9 and 110.

3. Results for primary tests, by domain

Overview. The impact estimates presented in this report are considered preliminary because they reflect only 6 (I5 through I10) of the 10 planned quarters (I5 through I14) for the final primary tests, and only one year (Year 2) of the two planned years (Years 2 and 3) for three of the four quality-of-care process measures. In addition, they include only two of the three intervention cohorts. An addendum to this report will present results from the full primary test period for all three cohorts.

For the quality-of-care processes study domain, we found statistically significant favorable effects of the HCIA-funded intervention (Table V.6). For the service use domain, we found no statistically significant effects in either a favorable or an unfavorable direction. However, the study had inconclusive results with respect to all-cause inpatient admissions due to limits in statistical power. Similarly, statistical power was limited for outcomes in the quality-of-care outcomes and spending domains, such that the analysis found an indeterminate effect of the HCIA-funded intervention on these domains.

Quality-of-care processes. The likelihood of receiving an HbA1c test or a lipid profile for diabetes was 1.0 and 2.6 percent higher, respectively, for the treatment group (a favorable estimate) than the estimated counterfactual. (Our estimated counterfactual—the outcome the treatment group members would have had in the absence of the HCIA-funded intervention—is the treatment group mean minus the difference-in-differences estimate.) We do not consider these favorable point estimates to be substantively large because both were smaller than the substantive threshold for these outcomes of 15 percent. In addition, these favorable results were not statistically significant.

Table V.6. Results of primary tests for FLHSA

Primary test definition					Statistical power[a] to detect an effect that is		Results			
Domain (# of tests in domain)	Outcome (units)	Time period for impacts	Population	Substantive threshold (expected direction of effect)[b]	Size of the substantive threshold	Twice the size of the substantive threshold[c]	Treatment group mean	Regression-adjusted difference between the treatment and estimated counterfactual (standard error)[b]	Percentage difference[d]	*p*-value[e]
Quality-of-care process (4)	Received an HbA1c test (binary [yes or no]/beneficiary/year)	Intervention quarters 5–8[f]	Medicare FFS beneficiaries ages 18 to 75 with diabetes assigned to treatment practices	15.0% (+)	>99.9%	>99.9%	89.5	0.9 (1.2)	1.0%	0.45
	Received a lipid profile(binary [yes or no]/beneficiary/year)	Intervention quarters 5–8[f]	Medicare FFS beneficiaries ages 18 to 75 with diabetes assigned to treatment practices	15.0% (+)	>99.9%	>99.9%	82.3	2.1 (1.5)	2.6%	0.24
	Received complete lipid profile in the year (binary [yes or no]/beneficiary/year)	Intervention quarters 5–8[f]	Medicare FFS beneficiaries ages 18 or older with ischemic vascular disease assigned to treatment practices	15.0% (+)	>99.9%	>99.9%	75.8	-0.6 (1.3)	-0.7%	0.51
	All inpatient admissions within a quarter were followed by an ambulatory care visit with a primary care or specialist provider within 14 days (binary [yes or no]/beneficiary/year)	Intervention quarters 5–10 for Cohort 1 and 5–8 for Cohort 2	Medicare FFS beneficiaries with at least one hospital stay in the quarter assigned to treatment practices	15.0% (+)	>99.9%	>99.9%	70.7	3.1* (1.6)	4.6%	0.08
	Combined (%)	Varies by test	Varies by test	15.0% (+)	>99.9%	>99.9%	n.a.	n.a.	1.9%**	0.04

Table V.6 *(continued)*

Primary test definition					Statistical power[a] to detect an effect that is		Results			
Domain (# of tests in domain)	Outcome (units)	Time period for impacts	Population	Substantive threshold (expected direction of effect)[b]	Size of the substantive threshold	Twice the size of the substantive threshold[c]	Treatment group mean	Regression-adjusted difference between the treatment and estimated counterfactual (standard error)[b]	Percentage difference[d]	*p*-value[e]
Quality-of-care outcomes (4)	Inpatient admissions for ambulatory care-sensitive conditions (#/1,000 beneficiaries/ quarter)	Intervention quarters 5–10 for Cohort 1 and 5–8 for Cohort 2	All observable Medicare FFS beneficiaries attributed to treatment practices	5.0% (-)	26.7	51.6	16.3	0.3 (1.2)	1.6%	0.50
	30-day unplanned hospital readmissions (#/1,000 beneficiaries/quarter)	Intervention quarters 5–10 for Cohort 1 and 5–8 for Cohort 2	All observable Medicare FFS beneficiaries attributed to treatment practices	5.0% (-)	23.0	42.1	14.4	0.1 (1.3)	0.7%	0.50
	Inpatient admissions for ambulatory care-sensitive conditions (#/1,000 beneficiaries/ quarter)	Intervention quarters 5–10 for Cohort 1 and 5–8 for Cohort 2	All observable high-risk Medicare FFS beneficiaries attributed to treatment practices	15.0% (-)	58.1	95.4	43.7	-0.7 (4.5)	-1.6%	0.50
	30-day unplanned hospital readmissions (#/1,000 beneficiaries/quarter)	Intervention quarters 5–10 for Cohort 1 and 5–8 for Cohort 2	All observable high-risk Medicare FFS beneficiaries attributed to treatment practices	15.0% (-)	42.8	82.1	40.0	5.0 (4.8)	14.2%	0.68
	Combined (%)	Intervention quarters 5–10 for Cohort 1 and 5–8 for Cohort 2	Varies by test	10.0% (-)	45.6	85.5	n.a.	n.a.	3.7%	0.67

Table V.6 *(continued)*

Primary test definition					Statistical power[a] to detect an effect that is		Results			
Domain (# of tests in domain)	Outcome (units)	Time period for impacts	Population	Substantive threshold (expected direction of effect)[b]	Size of the substantive threshold	Twice the size of the substantive threshold[c]	Treatment group mean	Regression-adjusted difference between the treatment and estimated counterfactual (standard error)[b]	Percentage difference[d]	*p*-value[e]
Service use (4)	All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	Intervention quarters 5–10 for Cohort 1 and 5–8 for Cohort 2	All observable Medicare FFS beneficiaries attributed to treatment practices	3.0% (-)	30.8	60.9	86.1	3.1 (3.2)	3.7%	0.63
	Outpatient ED visits (#/1,000 beneficiaries /quarter)	Intervention quarters 5–10 for Cohort 1 and 5–8 for Cohort 2	All observable Medicare FFS beneficiaries attributed to treatment practices	5.0% (-)	63.5	97.6	169.8	-3.5 (5.3)	-2.0%	0.45
	All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	Intervention quarters 5–10 for Cohort 1 and 5–8 for Cohort 2	All observable high-risk Medicare FFS beneficiaries attributed to treatment practices	5.0% (-)	32.2	64.0	190.3	13.7 (10.8)	7.8%	0.74
	Outpatient ED visits (#/1,000 beneficiaries/quarter)	Intervention quarters 5–10 for Cohort 1 and 5–8 for Cohort 2	All observable high-risk Medicare FFS beneficiaries attributed to treatment practices	15.0% (-)	89.9	>99.9	271.1	-20.3 (17.1)	-7.0%	0.29
	Combined (%)	Intervention quarters 5–10 for Cohort 1 and 5–8 for Cohort 2	Varies by test	7.0% (-)	77.1	99.7	n.a.	n.a.	0.6%	0.57

Table V.6 *(continued)*

Primary test definition					Statistical power[a] to detect an effect that is		Results			
Domain (# of tests in domain)	Outcome (units)	Time period for impacts	Population	Substantive threshold (expected direction of effect)[b]	Size of the substantive threshold	Twice the size of the substantive threshold[c]	Treatment group mean	Regression-adjusted difference between the treatment and estimated counterfactual (standard error)[b]	Percentage difference[d]	*p*-value[e]
Spending (2)	Medicare Part A and B spending ($/beneficiary/ month)	Intervention quarters 5–10 for Cohort 1 and 5–8 for Cohort 2	All observable Medicare FFS beneficiaries attributed to treatment practices	2.0% (-)	26.8	51.9	$836	$11 ($24.9)	1.3%	0.57
	Medicare Part A and B spending ($/beneficiary/ month)	Intervention quarters 5–10 for Cohort 1 and 5–8 for Cohort 2	All observable high-risk Medicare FFS beneficiaries attributed to treatment practices	3.0% (-)	24.8	46.8	$1,684	$6 ($83.8)	0.3%	0.50
	Combined (%)	Intervention quarters 5–10 for Cohort 1 and 5–8 for Cohort 2	Varies by test	2.5% (-)	27.0	52.3	n.a.	n.a.	0.8%	0.59

Source: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at the Centers for Medicare & Medicaid Services.

Note: The results for each outcome are based on a difference-in-differences regression model, as described in the text. Estimates are calculated for Medicare beneficiaries who are observable in the relevant time period: that is, beneficiaries who are enrolled in Medicare FFS (Part A and B), are alive, and have Medicare as their primary payer. Additional sample restrictions apply to the quality-of-care process measures; see text for details.

[a] The power calculation is based on actual standard errors from the analysis. For example, in the second-to-last row, a 3.0 percent effect on Medicare Part A and B spending (from the counterfactual of $1,684 + $6 = $1,690) would be a change of $51. Given the standard error of $84 from the regression model, we would be able to detect a statistically significant result 24.8 percent of the time if the impact was truly -$51, assuming a one-sided statistical test at the $p = 0.10$ significance level.

[b] The substantive threshold is the impact as a percentage of the counterfactual. The counterfactual is the outcome the treatment group would have had in the absence of the HCIA-funded intervention. Our estimate of the counterfactual is the treatment group mean minus the regression-adjusted difference-in-differences estimate.

[c] We show statistical power to detect a very large effect (twice the size of the substantive threshold) because this provides additional information about the likelihood that we will find effects if the program is indeed effective. If power to detect effects is less than 75 percent even for a very large effect, then the evaluation is extremely poorly powered for that outcome.

[d] Percentage difference is calculated as the regression-adjusted difference between the treatment and comparison groups, divided by the adjusted comparison group mean.

Table V.6 *(continued)*

[e] *p*-values test the null hypothesis that the regression-adjusted difference-in-differences estimate is less than or equal to zero for outcomes in the quality-of-care processes domain, or greater than or equal to zero in all other domains (a one-sided test). Because it is a one-sided test, as the difference-in-differences estimate approaches infinity in an unfavorable direction (negative for process-of-care measures and positive for all other measures), the *p*-value approaches 1, whereas it would approach 0 in a two-sided test. We adjusted the *p*-values for the multiple (three) comparisons made within the quality-of-care processes domain, and (separately) for the two comparisons made within the quality-of-care outcomes domain, and for the two comparisons made within the service use domain.

[f] For the three quality-of-care process measures for diabetes and ischemic vascular disease, we calculated outcomes over a year-long period (rather than quarters) covering intervention quarters 5 through 8.

ED = emergency department; FFS = fee-for-service; FLHSA = Finger Lakes Health Systems Agency; HCIA = Health Care Innovation Award.

n.a. = not applicable.

The likelihood of receiving a lipid profile for IVD was 0.7 percent lower for the treatment group (an unfavorable estimate) than the estimated counterfactual. We cannot conclude whether these unfavorable results are statistically significant because our one-sided statistical tests are designed to assess only improvements in outcomes.

The likelihood of receiving an ambulatory care visit within 14 days of hospital discharge was 4.6 percent higher in the treatment group than its estimated counterfactual, a favorable difference that was statistically significant. The combined estimate across the three measures in the quality-of-care processes domain was 1.9 percent, a favorable point estimate that was statistically significant. Although the estimates for ambulatory care visits—and for all quality-of-care process measures combined—were statistically significant, they were smaller than the substantive thresholds.

Quality-of-care outcomes. The rate of ACSC admissions for the treatment group during the primary test period was 1.6 percent higher than our estimate of the counterfactual for the full Medicare FFS population, but 1.6 percent lower than our estimate of the counterfactual for high-risk Medicare FFS beneficiaries. The rate of unplanned readmissions was 0.7 percent higher for the full Medicare FFS population and 14.2 percent higher than our estimate of the counterfactual for high-risk Medicare FFS beneficiaries. Higher rates in unplanned readmissions for the treatment group were in an unfavorable direction (indicating an increase in readmissions). However, no differences were substantively large for ACSC admissions or 30-day readmissions (the threshold is 5 percent for Medicare FFS beneficiaries and 15 percent for high-risk beneficiaries). After combining results across the two outcomes (and among both populations) in this domain, the combined effect was 3.7 percent smaller than the substantive threshold of 10.0 percent and in the unfavorable direction.

The statistical power to detect effects the size of the substantive threshold was poor to marginal for ACSC admissions (26.7 percent for the Medicare FFS population and 58.1 for high-risk Medicare FFS beneficiaries) and poor for 30-day unplanned readmissions (23.0 percent for the Medicare FFS population and 42.8 percent for high-risk Medicare FFS beneficiaries). Power was also poor (45.6 percent) for the combined effect in the domain.

Service use. The treatment group's admission rate was 3.7 percent higher for the full Medicare FFS population and 7.8 percent higher than our estimate of the counterfactual for high-risk Medicare FFS beneficiaries; these unfavorable differences were substantively large but not statistically significant. The treatment group's outpatient ED rate was 2.0 percent lower for the full Medicare FFS population and 7.0 percent lower than our estimate of the counterfactual for high-risk Medicare FFS beneficiaries; these favorable differences were not statistically significant or substantively large. After combining results across the two outcomes in this domain, the outcomes for the treatment group were similar to the estimated counterfactual. Power to detect effects that were the size of the substantive thresholds was poor for the admissions measure (30.8 and 32.2 for all patients and high-risk beneficiaries, respectively), marginal for the outpatient ED visit measure for all patients (63.5 percent), and good for the outpatient ED visit measure for high-risk beneficiaries (89.9 percent) and the combined outcome measure (77.1 percent).

Spending. For the full Medicare FFS population, the treatment group averaged $836 per beneficiary per month in Part A and B spending during the 5th through 10th intervention quarters, 1.3 percent (or $11) higher than the estimated counterfactual. Among high-risk Medicare FFS beneficiaries, spending was similar between the treatment and comparison groups. Among both groups, treatment-comparison differences were smaller than the substantive thresholds of 2 and 3 percent for all Medicare FFS beneficiaries and high-risk Medicare FFS beneficiaries, respectively. Statistical power to detect an effect the size of the substantive threshold was poor for individual outcomes as well as the combined outcome (ranging from 24.8 to 26.8 percent).

Aggregate estimates for CMMI's core measures. The estimates presented for the CMMI core outcomes—that is, for 30-day unplanned readmissions, all-cause inpatient admissions, the outpatient ED visit rate, and Medicare Part A and B spending—have so far been expressed per 1,000 beneficiaries per quarter (or, for spending, per beneficiary per month). Table V.7 translates these rates or per-beneficiary-per-month estimates into estimates of aggregate impacts during the 18-month primary test period presented in this report. We calculated these aggregate impacts by multiplying the point estimates by the average number of Medicare FFS beneficiaries in the treatment group and by the number of quarters or months during the primary test period. Although the point estimates are small for most of these measures, the aggregate estimates are fairly large because they are scaled to the full population of more than 13,000 Medicare FFS beneficiaries (and more than 2,600 high-risk Medicare FFS beneficiaries) assigned to practices and to the full 18 months of the primary test period. For example, the results in Table V.6 show the intervention was associated with an increase in Medicare Part A and B spending of $11 per beneficiary per month, or 1.3 percent relative to the estimated counterfactual. However, across more than 13,000 beneficiaries and 18 months, this small spending increase per beneficiary per month translates into an aggregate cost to the program of roughly $2.5 million. These large point estimates should be interpreted with caution because the estimates are not statistically significant for any of the outcomes (the *p*-values for these aggregate estimates are the same as for the main results shown in Table V.6).

4. Results for secondary tests

Estimates during the first intervention year (January 2013 to January 2014 for Cohort 1 and July 2013 to July 2014 for Cohort 2). As shown in Table V.8, most differences in quality-of-care outcomes, service use, and spending for the treatment group and its estimated counterfactual were small and not statistically significant during the first 12 months of the intervention (I1 through I4). Among the 10 patient outcomes, only one had a statistically significant difference in the treatment group and its estimated counterfactual: outpatient ED visits for high-risk Medicare FFS beneficiaries were 8.4 percent lower in the treatment group than in the comparison group. (This might reflect a treatment–comparison difference unrelated to the intervention, as impacts on service use were not yet expected during the intervention's first 12 months.) Overall, these results support the credibility of the comparison group because we do not see numerous or large differences (favorable or unfavorable) during the first year of practice participation, a period during which we and FLHSA did not expect to see program effects in quality-of-care outcomes, service use, or spending. This increased confidence in the comparison group, in turn, gives us

greater confidence in the primary test results and, eventually, the conclusions of the impact evaluation.

Table V.7. Results for primary tests for CMMI's core outcomes, expressed as aggregate effects for Medicare FFS beneficiaries in the treatment group

Outcome (units)	Aggregate impact estimate during the primary test period (January 1, 2014, through July 31, 2015)	*p*-value
All observable Medicare FFS beneficiaries attributed to treatment practices		
30-day unplanned readmissions (#)	+8	0.50
All-cause inpatient admissions (#)	+245	0.63
Outpatient ED visits (#)	-280	0.45
Medicare Part A and B spending ($)	+$2,537,437	0.57
All observable high-risk Medicare FFS beneficiaries attributed to treatment practices		
30-day unplanned readmissions (#)	+79	0.68
All-cause inpatient admissions (#)	+217	0.74
Outpatient ED visits (#)	-322	0.29
Medicare Part A and B spending ($)	+$275,925	0.50

Sources: Mathematica's calculation, based on analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at the Centers for Medicare & Medicaid Services.

Note: To estimate the aggregate impact during the primary test period (intervention quarters 5 through 10) we (1) multiplied the per beneficiary per quarter (or month) estimate from Table V.5 by the average number of Medicare FFS beneficiaries in the treatment group during the six primary test quarters, then (2) scaled the estimate to 18 months by multiplying the resulting product by 6 (or 18). The *p*-values are taken from Table V.6 and are therefore one-sided (testing that the program improved outcomes) and adjusted for multiple comparisons conducted within each outcome domain.

CMMI = Center for Medicare & Medicaid Innovation; ED = emergency department; FFS = fee-for-service.

However, there were favorable (and statistically significant) differences in quality-of-care process measures for the treatment group and its estimated counterfactual during the first 12 months of the intervention—particularly with respect to inpatient admissions followed by an ambulatory care visit (increase of 8.0 percent) and patients who received an HbA1c test during the year (increase of 2.0 percent). In these cases of admissions followed by an ambulatory visit, first-year impacts were larger in magnitude than impacts we found in later months, and appear immediately upon the start of the intervention period (and even potentially in the final two baseline quarters in the case of admissions followed by an ambulatory visit). For this reason, there is some doubt as to whether the HCIA intervention is solely responsible for these favorable impacts, given that HCIA-funded care managers (hired in the first intervention quarter—from January to March 2013 for the first cohort) would not likely have the ability to affect this measure until the second intervention quarter. Potentially, these favorable impacts in quality-of-care process measures also reflect quality improvement efforts in treatment practices that might have begun before, and outside the scope of, the HCIA intervention. For example, it is possible that some treatment practices initiated quality improvement efforts before joining the HCIA program, and these efforts—combined with HCIA activities that began in 2013—helped generate the positive impact in quality-of-care process measures we detected in the first 12 months of the intervention and beyond.

Table V.8. Results of secondary tests for FLHSA

Secondary test definition				Results			
Domain (# of tests in domain)	Outcome (units)	Time period for impacts	Population	Treatment group mean	Regression-adjusted difference between the treatment and estimated counterfactual (standard error) [a]	Percentage difference[a]	*p*-value[b]
Estimates during the first intervention year (January 1, 2013, to December 31, 2013, for Cohort 1, July 1, 2013, to June 30, 2014, for Cohort 2)							
Quality-of-care process (4)	Received an HbA1c test (binary [yes or no]/beneficiary/year)	Intervention quarters 1–4 for Cohorts 1 and 2	Medicare FFS beneficiaries aged 18-75 with diabetes assigned to treatment panels	91.1	1.8* (1.1)	2.0%	0.06
	Received a lipid profile(binary [yes or no]/beneficiary/year)	Intervention quarters 1–4 for Cohorts 1 and 2	Medicare FFS beneficiaries aged 18-75 with diabetes assigned to treatment panels	84.3	1.2 (1.4)	1.4%	0.20
	Received complete lipid profile in the year (binary [yes or no]/beneficiary/year)	Intervention quarters 1–4 for Cohorts 1 and 2	Medicare FFS beneficiaries aged 18 or older with ischemic vascular disease assigned to treatment panels	78.2	-0.1 (1.2)	-0.2%	0.54
	All inpatient admissions within a quarter were followed by an ambulatory care visit with a primary care or specialist provider within 14 days (binary [yes or no]/beneficiary/year)	Intervention quarters 1–4 for Cohorts 1 and 2	Medicare FFS beneficiaries with at least one hospital stay in the quarter assigned to treatment panels	72.5	5.4*** (1.6)	8.0%	0.00
Quality-of-care outcomes (4)	Inpatient admissions for ambulatory care-sensitive conditions (#/beneficiary/quarter)	Intervention quarters 1–4 for Cohorts 1 and 2	All observable Medicare FFS beneficiaries attributed to treatment practices	15.1	-0.9 (1.2)	-5.9%	0.22
	30-day unplanned hospital readmissions (#/beneficiary/quarter)	Intervention quarters 1–4 for Cohorts 1 and 2	All observable Medicare FFS beneficiaries attributed to treatment practices	12.2	-1.0 (1.2)	-7.4%	0.22
	Inpatient admissions for ambulatory care-sensitive conditions (#/beneficiary/quarter)	Intervention quarters 1–4 for Cohorts 1 and 2	All observable high-risk Medicare FFS beneficiaries attributed to treatment practices	39.5	-1.5 (4.3)	-3.7%	0.36
	30-day unplanned hospital readmissions (#/beneficiary/quarter)	Intervention quarters 1–4 for Cohorts 1 and 2	All observable high-risk Medicare FFS beneficiaries attributed to treatment practices	32.6	1.7 (4.3)	5.4%	0.65

Table V.8 *(continued)*

Secondary test definition				Results			
Domain (# of tests in domain)	Outcome (units)	Time period for impacts	Population	Treatment group mean	Regression-adjusted difference between the treatment and estimated counterfactual (standard error) [a]	Percentage difference[a]	*p*-value[b]
Service use (4)	All-cause inpatient admissions (#/1,000 beneficiaries/ quarter)	Intervention quarters 1–4 for Cohorts 1 and 2	All observable Medicare FFS beneficiaries attributed to treatment practices	80.6	-1.5 (3.1)	-1.8%	0.32
	Outpatient ED visits (#/1,000 beneficiaries/quarter)	Intervention quarters 1–4 for Cohorts 1 and 2	All observable Medicare FFS beneficiaries attributed to treatment practices	155.8	-3.2 (5.0)	-2.0%	0.26
	All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	Intervention quarters 1–4 for Cohorts 1 and 2	All observable high-risk Medicare FFS beneficiaries attributed to treatment practices	180.7	7.1 (10.1)	4.1%	0.76
	Outpatient ED visits (#/1,000 beneficiaries/quarter)	Intervention quarters 1–4 for Cohorts 1 and 2	All observable high-risk Medicare FFS beneficiaries attributed to treatment practices	248.6	-22.7* (14.9)	-8.4%	0.06
Spending (2)	Medicare Part A and B spending ($/beneficiary/month)	Intervention quarters 1–4 for Cohorts 1 and 2	All observable Medicare FFS beneficiaries attributed to treatment practices	$766	$9 ($24.1)	1.2%	0.64
	Medicare Part A and B spending ($/beneficiary/month)	Intervention quarters 1–4 for Cohorts 1 and 2	All observable high-risk Medicare FFS beneficiaries attributed to treatment practices	$1,581	$45 ($76.5)	2.9%	0.72
Estimates limiting the sample to prevent sample addition after the first baseline or intervention quarter							
Quality-of-care process (4)	Received an HbA1c test (binary [yes or no]/ beneficiary/year)	Intervention quarters 5–8 [c]	Medicare FFS beneficiaries ages 18 to 75 with diabetes assigned to treatment panels	90.1	0.9 (1.3)	1.0%	0.23
	Received a lipid profile(binary [yes or no]/beneficiary/year)	Intervention quarters 5–8 [c]	Medicare FFS beneficiaries ages 18 to 75 with diabetes assigned to treatment panels	83.1	2.1* (1.6)	2.6%	0.09
	Received complete lipid profile in the year (binary [yes or no]/beneficiary/year)	Intervention quarters 5–8 [c]	Medicare FFS beneficiaries ages 18 or older with ischemic vascular disease assigned to treatment panels	76.2	-0.5 (1.3)	-0.7%	0.65

Table V.8 *(continued)*

Secondary test definition				Results			
Domain (# of tests in domain)	Outcome (units)	Time period for impacts	Population	Treatment group mean	Regression-adjusted difference between the treatment and estimated counterfactual (standard error) [a]	Percentage difference[a]	*p*-value[b]
	All inpatient admissions within a quarter were followed by an ambulatory care visit with a primary care or specialist provider within 14 days (binary [yes or no]/beneficiary/year)	Intervention quarters 5–10 for Cohort 1 and 5–8 for Cohort 2	Medicare FFS beneficiaries with at least one hospital stay in the quarter assigned to treatment panels	72.3	4.0*** (1.7)	5.8%	0.01
Quality-of-care outcomes (4)	Inpatient admissions for ambulatory care-sensitive conditions (#/1,000 beneficiaries/ quarter)	Intervention quarters 5–10 for Cohort 1 and 5–8 for Cohort 2	All observable Medicare FFS beneficiaries attributed to treatment practices	17.0	-0.1 (1.3)	-0.7%	0.47
	30-day unplanned hospital readmissions (#/1,000 beneficiaries/quarter)	Intervention quarters 5–10 for Cohort 1 and 5–8 for Cohort 2	All observable Medicare FFS beneficiaries attributed to treatment practices	14.8	0.3 (1.5)	2.3%	0.59
	Inpatient admissions for ambulatory care-sensitive conditions (#/1,000 beneficiaries/ quarter)	Intervention quarters 5–10 for Cohort 1 and 5–8 for Cohort 2	All observable high-risk Medicare FFS beneficiaries attributed to treatment practices	44.8	-0.3 (4.7)	-0.7%	0.47
	30-day unplanned hospital readmissions (#/1,000 beneficiaries/quarter)	Intervention quarters 5–10 for Cohort 1 and 5–8 for Cohort 2	All observable high-risk Medicare FFS beneficiaries attributed to treatment practices	40.2	4.7 (5.1)	13.3%	0.82
Service use (4)	All-cause inpatient admissions (#/1,000 beneficiaries/ quarter)	Intervention quarters 5–10 for Cohort 1 and 5–8 for Cohort 2	All observable Medicare FFS beneficiaries attributed to treatment practices	87.4	3.7 (3.5)	4.5%	0.86
	Outpatient ED visits (#/1,000 beneficiaries/quarter)	Intervention quarters 5–10 for Cohort 1 and 5–8 for Cohort 2	All observable Medicare FFS beneficiaries attributed to treatment practices	165.1	-3.1 (5.8)	-1.8%	0.30
	All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	Intervention quarters 5–10 for Cohort 1 and 5–8 for Cohort 2	All observable high-risk Medicare FFS beneficiaries attributed to treatment practices	189.1	13.8 (11.4)	7.9%	0.89
	Outpatient ED visits (#/1,000 beneficiaries/ quarter)	Intervention quarters 5–10 for Cohort 1 and 5–8 for Cohort 2	All observable high-risk Medicare FFS beneficiaries attributed to treatment practices	259.2	-24.3* (18.0)	-8.6%	0.09

Table V.8 *(continued)*

Secondary test definition				Results			
Domain (# of tests in domain)	Outcome (units)	Time period for impacts	Population	Treatment group mean	Regression-adjusted difference between the treatment and estimated counterfactual (standard error) [a]	Percentage difference[a]	*p*-value[b]
Spending (2)	Medicare Part A and B spending ($/beneficiary/month)	Intervention quarters 5–10 for Cohort 1 and 5–8 for Cohort 2	All observable Medicare FFS beneficiaries attributed to treatment practices	$831	-$3 ($27.7)	-0.3%	0.46
	Medicare Part A and B spending ($/beneficiary/month)	Intervention quarters 5–10 for Cohort 1 and 5–8 for Cohort 2	All observable high-risk Medicare FFS beneficiaries attributed to treatment practices	$1,640	-$21 ($89.2)	-1.2%	0.41

Sources: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at the Centers for Medicare & Medicaid Services.

Notes: The results for each outcome are based on a difference-in-differences regression model, as described in the text. Estimates are calculated for Medicare beneficiaries who are observable in the relevant time period: that is, beneficiaries who are enrolled in Medicare FFS (Part A and B), are alive, and have Medicare as their primary payer. We defined high-risk beneficiaries as those with a Hierarchical Condition Category score in the top third among all treatment group members at the beginning of the baseline period (for outcomes in the baseline period) or intervention period (for outcomes in the intervention period).

[a] Percentage difference is calculated as the regression-adjusted difference between the treatment and comparison group, divided by the adjusted comparison group mean.

[b] The *p*-values from the secondary test results were not adjusted for multiple comparisons within or across domains.

[c] For the three quality-of-care process measures for diabetes and ischemic vascular disease, we calculated outcomes over a year-long period (rather than quarters) covering intervention quarters 5 through 8.

ED = emergency department; FFS = fee-for-service; FLHSA = Finger Lakes Health Systems Agency.

Estimates limiting the sample to prevent sample addition. We conducted additional secondary tests that limited the sample to those beneficiaries attributed at the start of the baseline or intervention period. These tests used the same time period as the primary tests: the 5th through 10th intervention quarters for Cohort 1 (January 2014 through June 2015) and the 5th through 8th intervention quarters for Cohort 2 (July 2014 through June 2015). The results of these secondary tests were generally consistent with the primary test results; they showed a small and statistically significant favorable impact of the intervention on hospitalizations followed by ambulatory care visits within 14 days, and inconclusive results with respect to all-cause inpatient admissions and outcomes in the quality-of-care outcomes and spending domains for the full Medicare FFS population and high-risk Medicare FFS beneficiaries.

However, these tests showed two statistically significant favorable effects of the intervention that were not present in primary tests: an increase in patients with diabetes who received a lipid profile in the quality-of-care process domain (of 2.6 percent) and a reduction in outpatient ED visits for high-risk Medicare FFS beneficiaries (of 8.4 percent). Both of these impacts could point to potential positive effects of the intervention on practices' long-standing patients. The favorable impact on outpatient ED visits for high-risk Medicare FFS beneficiaries could also reflect potential treatment–comparison differences on this measure that are unrelated to the HCIA intervention, because this impact first emerged within the first 12 months of the intervention's start date, when impacts on service use were not yet expected (see the first set of primary tests discussed previously).

5. Consistency of impact estimates with implementation findings

The impact estimates in the primary tests are plausible given the implementation findings. Notably, statistically significant (albeit not substantively important) impacts on ambulatory care visits with a primary care or specialist provider within 14 days of hospitalization likely reflect care managers' efforts to follow up with patients after a hospitalization, coordinate patients' care among medical and community providers, and connect patients with community-based service organizations and transportation services for their medical appointments.

However, the primary test results did not show any favorable effects during the first 6 quarters of the 10-quarter primary test period that were statistically significant or substantively important in the quality-of-care outcomes, service use, and spending domains. The implementation evidence shows the program was active during these 6 quarters. For example, as described in Section III.B.1, as of July 2015, practices provided services to 17,484 unique patients, exceeding the target cumulative enrollment of 13,564 patients (about half of whom were expected to receive intensive care management) for the entire award period. Even with a well-implemented intervention, it is possible that the program was unable to change beneficiaries' or providers' behaviors in ways that would affect impact outcomes in these domains during the primary test period covered in this report. In the case of FLHSA, it is possible that the program's large investments in care management and practice transformation helped generate modest positive impacts in quality-of-care processes, but did not translate into desired reductions in hospitalizations and costs during the first 6 quarters of the primary test period.

6. Conclusions about program impacts, by domain

Based on all evidence currently available, we draw the following conclusions about program impacts during the first 18 of the planned 30 months of the primary test period. Table V.9 summarizes these conclusions and their support.

- **The program had a statistically significant favorable effect on quality-of-care processes.** For the ambulatory care visit with a primary care or specialist provider within 14 days and the combined outcome in this domain, we found statistically significant favorable impacts. The secondary test results support these primary test results by (1) showing impacts in the first program year (when the intervention would presumably begin to register an effect on these quality-of-care processes) and (2) demonstrating that differential sample addition over time between the treatment and comparison groups did not drive results. However, given that impacts emerge between the end of the baseline period and the start of the implementation period, it is possible that they cannot be fully attributed to the HCIA-funded intervention. The point estimates suggest that the favorable impacts were modest in size (given that the estimates were smaller than the prespecified substantive thresholds).
- **The program had an indeterminate effect on quality-of-care outcomes.** The primary test results were not statistically significant for any outcome or population in this domain, and the combined test in the domain was not substantively large or statistically significant. However, the statistical power was poor to detect effects the size of the substantive threshold. As a result, null findings from the primary test in this domain could be due to (1) the program truly not having a substantively large effect or (2) the program having a substantively large effect but our tests failing to detect it.
- **The program had no substantively large effect on service use.** The primary test results were not statistically significant for any outcome or population in this domain, and the combined test in the domain was not substantively large or statistically significant. The statistical power was good to detect effects the size of the substantive threshold for outpatient ED visits among high-risk beneficiaries and the combined outcome (more than 75 percent). These conclusions are also consistent with implementation findings because, although the program was implemented reasonably well, it is plausible the program did not have intended effects in the service use domain.
- **The program had an indeterminate effect on Medicare spending.** The primary test results were not statistically significant for any outcome or population in this domain, and the combined test in the domain was not substantively large or statistically significant. However, the statistical power was poor to detect effects the size of the substantive threshold. As a result, null findings from the primary test in this domain could be due to (1) the program truly not having a substantively large effect or (2) the program having a substantively large effect but our tests failing to detect it. The fact that we observed no declines in service use (which FLHSA anticipated would drive reductions in spending)—and that some primary tests for service use were well powered—suggests that lack of effects on spending is the more likely explanation.

Table V.9. Preliminary conclusions about the impacts of FLHSA's HCIA program on patients' outcomes, by domain

		Evidence supporting conclusion		
Domain	Preliminary conclusion	Primary test result(s) that supported conclusion	Primary test result(s) plausible given secondary tests?	Primary test result(s) plausible given implementation evidence?
Quality-of-care process	Statistically significant favorable effect	• Estimate for an ambulatory care visit with a primary care or specialist provider within 14 days was favorable and statistically significant (after adjusting for four tests in domain) • Estimate for the combined outcome in the quality-of-care process domain was favorable and statistically significant	Yes	Yes
Quality-of-care outcomes	Indeterminate effect	• No substantively large or statistically significant effects; poorly powered to detect a substantively large effect in combined outcome in the domain; poorly to marginally powered to detect a substantively large effect in individual measures	Yes	Yes
Service use	No substantively large effect	• No statistically significant effects and the combined test for all outcomes in the domain was neither statistically significant nor substantively large; well-powered to detect a substantively large effect on ED visits for high-risk beneficiates and the combined outcome in the domain	Yes	Yes
Spending	Indeterminate effect	• No substantively large or statistically significant effects; poorly powered to detect a substantively large effect in individual measures and the combined outcome in the domain	Yes	Yes

Sources: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at the Centers for Medicare & Medicaid Services.

ED = emergency department; FLHSA = Finger lakes Health Systems Agency; HCIA = Health Care Innovation Award.

VI. DISCUSSION AND CONCLUSIONS

FLHSA used its $27 million in HCIA funding to implement the following three intervention components: (1) practice transformation to transform 68 participating practices into PCMHs; (2) intensive care management for high-risk patients; and (3) implementing a community-wide outcomes-based payment model. FLHSA expected that these intervention components would increase the quality of care and patients' access to care, and increase patients' activation and self-management, thus reducing inpatient admissions and outpatient ED visits. These reductions would, in turn, reduce total Medicare and Medicaid spending. FLHSA's goals were to reduce the total cost of care by 3 percent through a reduction of potentially preventable hospital admissions and 30-day readmissions by 25 percent and avoidable ED visits by 15 percent by the end of the award.

The results from our impact evaluation suggest FLHSA modestly improved process of care measures for Medicare FFS beneficiaries. Notably, the intervention improved the rate of ambulatory care visits with a primary care or specialist provider within 14 days of inpatient admissions by about 5 percent. The modest improvement in quality-of-care processes is encouraging, though stakeholders should consider the extent to which these favorable findings can be replicated in other settings. FLHSA convened primary care practices that were highly motivated to undergo improvement efforts, particularly to become PCMHs. The region also has a robust network of community service providers that are integrated with primary care practices. Therefore, these favorable impacts may generalize to other interventions involving practices that are highly committed to transformation in settings with good access to robust social services that patients can access shortly after a hospitalization.

However, there is no evidence that FLHSA improved quality-of-care outcomes, service use, or spending during the original three-year award period. Outcomes for Medicare FFS patients served by the 37 treatment practices were not statistically or substantively better than those for Medicare patients served by 108 matched comparison practices in the quality-of-care outcomes, service use, and spending domains. The evaluation was well powered to detect substantively large impacts on service use, but not quality-of-care outcomes or spending. (It is unclear whether the estimates for Medicare FFS beneficiaries could be generalized to other patients in FLHSA's target population, including Medicaid beneficiaries, Medicare Advantage beneficiaries, and those with private or no health insurance.)

The lack of favorable effects on service use for Medicare FFS beneficiaries does not appear to be a result of major problems implementing the intervention as planned. Indeed, FLHSA delivered a complex intervention consistent with its core design. Several measures capture the generally successful implementation:

- FLHSA successfully met or exceeded its staffing goals for the HCIA-funded intervention. As of July 2015, each practice met FLHSA's goal to employ at least 1 care manager, resulting in a total of 70 embedded care managers across the 68 practices.
- FLHSA clinical advisors met regularly with care managers to integrate the care manager into the care team at the practice. Clinical advisors also provided targeted technical support

to care managers (for example, to help care managers report on clinical quality measures through their EHRs) and organized learning collaboratives to facilitate learning across care managers.

- As of July 2015, FLHSA care managers provided services (both intensive and otherwise) to 17,484 unique patients, exceeding the target cumulative enrollment of 13,564 patients (about half of whom were expected to receive intensive care management) for the entire award period.
- Practice improvement advisors met regularly with practice champions and other staff to identify and work on quality improvement projects, improve communication pathways among practice staff, use EHRs to improve care processes, and improve practice workflows. FLHSA staff also organized monthly learning collaboratives to support practice champions and facilitate learning across practices. By July 2015, two-thirds of Cohort 1 practice champions and almost all Cohort 2 and 3 practice champions participated in learning collaboratives.
- Practice staff successfully implemented monthly care team meetings and weekly care team huddles to coordinate care. Whereas at the start of program implementation, only 40 percent of Cohort 1 practices, 11 percent of Cohort 2 practices, and no Cohort 3 practices reported holding monthly care team meetings, by July 2015, all Cohort 1 practices, 96 percent of Cohort 2 practices, and 75 percent of Cohort 3 practices held monthly care team meetings. Similarly, at the start of program implementation, 40 percent of Cohort 1 practices, 28 percent of Cohort 2 practices, and 88 percent of Cohort 3 practices reported they held weekly huddles, whereas, by July 2015, all practices held weekly huddles.

Further, the lack of effects on service use does not appear linked to any difficulties engaging PCPs as planned. PCPs are central to the awardee's theory of action because they, jointly with care managers, had to provide care coordination services to high-risk beneficiaries. The primary care clinician survey results indicate that most PCPs believed the HCIA-funded intervention improved the quality and patient-centeredness of care at their practices. (However, we have no evidence to assess whether and how the program changed PCPs' interactions with care teams, or whether it altered PCPs' fundamental treatment practices.) The lack of effects is also unlikely due to challenges integrating CHWs into the practices, as CHWs were largely hired at FQHCs in Cohort 1, and the impact analysis excluded all FQHCs. Similarly, the lack of effects is also unrelated to complications introducing the outcomes-based payment model, because the new model was largely expected to affect patients' care and outcomes after the award period.

These findings suggest that one of four factors might cause the lack of measured effects. First, although the program was generally implemented as planned, a few key implementation barriers might have limited the effectiveness of care management services in reducing utilization and costs. For example, both practice champions and care managers reported that they had limited time to devote to practice transformation and intensive care management activities, respectively. In particular, care managers' large caseloads could have negatively affected the quality or quantity of their interactions with patients, thus reducing the potential impact of care management services on patient activation, self-management, access to care, and health outcomes.

Second, the intervention might have set overly ambitious goals at the outset. A premise of the intervention's theory of action was that providing high-risk patients with intensive care management services would generate substantial reductions in readmissions, potentially preventable hospitalizations, and avoidable ED visits, on the order of 15 to 25 percent. FLHSA gave direct care management services to some 17,500 people out of 750,000 total patients, or 2 percent of all patients at participating practices. If substantial reductions in readmissions and potentially preventable hospitalizations were expected to be driven by this relatively small share of the patient population, they would have to be substantially, potentially unrealistically large.

Third, it is possible that impacts take longer to accrue than the primary test period in this impact analysis. For example, patients' enhanced management of their low-density lipoprotein and blood sugar levels would be unlikely to prevent hospitalizations within 12 months of activation, but could feasibly play a role in preventing hospitalizations during a longer time frame. Presumably, impacts on utilization and costs could grow larger during the final 12 months of program operations, which are not included in the current analysis.

Fourth, on the available baseline service metrics, some of the practices already conducted key practices supported by the intervention. For example, 40 percent of Cohort 1 practices held weekly huddles and 40 percent of practices held monthly care team meetings at baseline. Similarly, 80 percent of Cohort 2 practices used EHRs to generate population-based reports sorted by patients' ages and major diagnoses, and 67 percent of these practices used EHRs to generate patient-specific reports to identify gaps in care at baseline. As such, some practices might have had little room to improve, and thus less potential to generate impacts than if they had no prior experience with care management or PCMHs.

The next step for the evaluation is to add the final 12 months of program operations to the study period and include Cohort 3 practices in the impact analysis, completing the evaluation. FLHSA received a no-cost extension to continue its HCIA-funded intervention beyond the initial award period (which ended June 2015) through June 2016. As a result, we will update the implementation metrics in this report with the number of patients receiving managed care services as of mid-2016. We will also (1) incorporate Cohort 3 practices and their matched comparison practices into the impact analysis; (2) generate claims-based outcomes to cover the final 12 months of the primary test period (July 1, 2015, to June 30, 2016); (3) conduct the final primary tests incorporating these outcomes; and (4) update our conclusions, if necessary. We will report final evaluation results in an addendum to this report.

REFERENCES

Boards of Trustees, Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds. "2013 Annual Report of the Boards of Trustees of the Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds." Table V.D1. Washington, DC: Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds, 2013. Available at http://downloads.cms.gov/files/TR2013.pdf. Accessed August 13, 2014.

Centers for Medicare & Medicaid Services. "CSV Flat Files—Revised: Readmissions Complications and Deaths—National.csv." Baltimore, MD: CMS, 2014. Available at https://data.medicare.gov/data/hospital-compare. Accessed August 14, 2014.

Chronic Conditions Data Warehouse. "Table A.1.a. Medicare Beneficiary Counts for 2005–2014." Baltimore, MD: CMS, 2014a. Available at https://www.ccwdata.org/web/guest/medicare-tables-reports. Accessed June 29, 2016.

Chronic Conditions Data Warehouse. "Table B.2.a Medicare Beneficiary Prevalence for Chronic Conditions for 2005 Through 2014." Baltimore, MD: CMS, 2014b. Available at https://www.ccwdata.org/web/guest/medicare-tables-reports. Accessed June 29, 2016.

Gerhardt, Geoffrey, Alshadye Yemane, Keri Apostle, Allison Oelschlaeger, Eric Rollins, and Niall Brennan. "Evaluating Whether Changes in Utilization of Hospital Outpatient Services Contributed to Lower Medicare Readmission Rate." *Medicare & Medicaid Research Review,* vol. 4, no. 1, 2014, pp. E1–E13.

Health Indicators Warehouse. "Average Age of Medicare Beneficiaries (mean)." Hyattsville, MD: National Center for Health Statistics, HIW, 2014a. Available at http://www.healthindicators.gov/Indicators/Average-age-of-Medicare-beneficiaries-mean_308/Profile/ClassicData. Accessed November 19, 2014.

Health Indicators Warehouse. "Hospital Inpatient Medicare Admissions (per 1,000 beneficiaries)." Hyattsville, MD: National Center for Health Statistics, HIW, 2014b. Available at http://www.healthindicators.gov/Indicators/Hospital-inpatient-Medicare-admissions-per-1000-beneficiaries_2001/Profile/ClassicData. Accessed August 13, 2014.

Institute of Education Sciences. "What Works Clearinghouse: Procedures and Standards Handbook, Version 3.0." Washington, DC: U.S. Department of Education, IES, 2014. Available at http://ies.ed.gov/ncee/wwc/DocumentSum.aspx?sid=19. Accessed September 15, 2014.

MedPAC. "Data book: Beneficiaries dually eligible for Medicare and Medicaid. MACPAC, 2016. Available at https://www.macpac.gov/wp-content/uploads/2015/01/Dually-Eligible-Beneficiares-DataBook.pdf. Accessed September 15, 2016.

Peikes, Deborah, Stacy Dale, Eric Lundquist, Janice Genevro, and David Myers. "Building the Evidence Base for the Medical Home: What Sample and Sample Size Do Studies Need? White Paper." AHRQ Publication No.11-0100-EF. Rockville, MD: Agency for Healthcare Research and Quality, October 2011.

Rosenthal, M.B., S. Alidina, M. Friedberg, S. Singer, D. Eastman, Z. Li, and E. Schneider. "A Difference-in-Differences Analysis of Changes in Quality, Utilization, and Cost Following the Colorado Multi-Payer Patient-Centered Medical Home Pilot." *Journal of General Internal Medicine,* 2016, vol. 31, no. 3, March 2016, pp. 289–296.

Shapiro, Rachel, Randall Blair, Rebecca Coughlin, Boyd Gilman, Greg Peterson, Catherine DesRoches, Sandi Nelson, Laura Blue, Keith Kranker, Kate Stewart, Frank Yoon, and Lorenzo Moreno. "Findings for Finger Lakes Health Systems Agency." In Lorenzo Moreno, Boyd Gilman, Greg Peterson, Catherine DesRoches, Sheila Hoag, Linda Barterian, Laura Blue, Katharine Bradley, Emily Ehrlich, Kristin Geonotti, Lauren Hula, Keith Kranker, Rumin Sarwar, Rachel Shapiro, KeriAnn Wells, Joseph Zickafoose, Sandi Nelson, Kate Stewart, Frank Yoon. "Evaluation of Health Care Innovation Awards (HCIA): Primary Care Redesign Programs. Second Annual Report to CMS. Volume II: Individual Program Summaries." Princeton, NJ: Mathematica Policy Research, December 11, 2015, pp. 333–390.

Stuart, Elizabeth A. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science,* vol. 25, no. 1, 2010, pp. 1–21.

Taylor, M.J., C. McNicholas, C. Nicolay, A. Darzi, D. Bell, and J.E. Reed. "Systematic Review of the Application of the Plan–Do–Study–Act Method to Improve Quality in Healthcare." *BMJ Quality & Safety*, vol. 23, no. 4, 2014, pp. 290–298.

Truven Health Analytics. "AHRQ Quality Indicators, Prevention Quality Indicators v5.0 Benchmark Data Tables." Prepared for the Agency for Healthcare Research and Quality, U.S. Department of Health and Human Services. Santa Barbara, CA: Truven Health Analytics, March 2015. Available at http://www.qualityindicators.ahrq.gov/Downloads/Modules/PQI/V50/Version_50_Benchmark_Tables_PQI.pdf. Accessed August 18, 2015.

Van Walraven, C., I.A. Dhalla, C. Bell, E. Etchells, I.G. Stiell, K. Zarnke, P.C. Austin, and A.J. Forster. "Derivation and Validation of an Index to Predict Early Death or Unplanned Readmission After Discharge from Hospital to the Community." *Canadian Medical Association Journal,* vol. 182, no. 6, April 6, 2010, pp. 551–557.

CHAPTER 5

PACIFIC BUSINESS GROUP ON HEALTH

Sean Orzol, Rosalind Keith, Rumin Sarwar, Michael Barna, Greg Peterson, Boyd Gilman, Laura Blue, Keith Kranker, Kate Stewart, Sheila Hoag, and Lorenzo Moreno

This page has been left blank for double-sided copying.

PACIFIC BUSINESS GROUP ON HEALTH

CHAPTER SUMMARY

Introduction. Pacific Business Group on Health (PBGH), a nonprofit business coalition, used its $19.1 million Health Care Innovation Award (HCIA) funding to implement the Intensive Outpatient Care Program (IOCP), a care management program, in primary care practice sites affiliated with 23 participating medical groups (PMGs) in five states. The program served 15,008 participants, including those enrolled in Medicare Advantage, Medicare fee-for-service (FFS), those dually eligible for Medicare and Medicaid, and a small number of Medicaid-only beneficiaries, from May 1, 2013, to June 2015 (when HCIA-funded operations concluded). PBGH aimed to reduce total Medicare spending, hospitalizations, and emergency department (ED) visits by 5.0 percent among Medicare FFS beneficiaries and Medicare Advantage patients by the end of the three-year program. PBGH also aimed to improve the quality of care delivery to participants, as measured by improvements in quality-of-care process measures, by 2 to 4 percent by the end of the three-year program.

Objectives. This third annual report on PBGH has three objectives:

1. To describe the design and implementation of PBGH's HCIA-funded program, including the role of clinicians in the program and the extent to which anticipated changes in clinician behavior occurred
2. To assess impacts of the program on participants' outcomes and Medicare Part A and B spending during the three years of the award
3. To use both implementation and impact findings to identify possible explanations for the observed impacts

Methods. We reviewed PBGH's program documents and self-monitoring metrics, conducted interviews with PBGH's leadership and program staff, and surveyed participating clinicians and practice staff. To estimate impacts, we compared outcomes for Medicare FFS patients served by the PMGs with outcomes for Medicare FFS patients with similar characteristics served by other providers, adjusting for observed differences in outcomes between the two groups during a one-year baseline period. We did not include beneficiaries enrolled in Medicare Advantage in the impact evaluation due to limitations in available data.

Program design and implementation. The program provided care management services, whereby care managers funded and trained by the program were embedded in primary care teams to work with primary care providers (PCPs) to develop and implement personalized care plans (also known as shared action plans) for medically complex, high-risk participants. The role of the care manager was to engage in one-on-one interactions with participants to understand their medical and social needs, work with participants and their physicians to help participants manage their conditions, provide participants and their caregivers with education and emotional support, and connect participants to appropriate community resources.

Despite implementing certain aspects of the program as planned (including staffing, training, and care management services), PBGH was unable to give PMGs timely, standard risk score reports appropriate for identifying patients eligible for the program. Although PBGH provided eligibility guidelines to the PMGs, determination of eligibility was left in practice to the individual PMGs, and in some cases individual PCPs. Although clinicians might have been best suited to identify beneficiaries who could benefit most from the IOCP, this deviation from the planned, replicable risk-stratification algorithm to an undefined participant-identification method that relied on PCPs' subjective assessment likely resulted in variation in the criteria that participating providers used to identify patients eligible for the program.

Clinicians' perceptions of program effects on the care they provided. The available evidence suggests that PBGH did not engage clinicians as planned, due primarily to claims lag and data collection issues. After changing the participant-identification methodology, PBGH expected PCPs to help identify program-eligible participants. PBGH also expected PCPs to work with care managers in delivering care management services to improve the quality of care provided to participants and reduce hospitalizations and ED visits. However, in surveys we administered, only about 30 percent of clinician respondents were aware of the PBGH program. This may have resulted from PMG leadership using different terminology for the program, not labeling it as IOCP when marketing it to providers, and PMGs already having care management referral protocols in place, without designating it as IOCP. Among those who were aware of the program, respondents varied in their perceptions of program effectiveness. Of the respondents familiar with the PBGH program, the majority believed the program had a positive impact on the quality of care, ability to respond in a timely way to participants' needs, and patient-centeredness.

Impacts on patients' outcomes. Due to concerns about likely biases in the impact estimates, we were unable to draw conclusions about program impacts. The impact estimates indicate that, during the three years of the award, the intervention had substantively large and unfavorable impacts for all measures in the quality-of-care outcomes (ambulatory care-sensitive condition admissions and readmissions), service use (inpatient admissions and ED visits), and spending (Medicare Part A and B) domains, and one of the three quality-of-care processes (patients with ischemic vascular disease receiving the recommended lipid test). However, nothing in the implementation evidence explains how the IOCP intervention could have caused such large and unfavorable outcomes for patients. It is more plausible that the impact results are due to unobserved differences between the treatment and matched comparison groups than to something the care management intervention did or did not do.

Conclusion. We were unable to draw conclusions about program impacts. Results from our impact evaluation showed unfavorable impacts for treatment beneficiaries across outcomes in all four evaluation domains. However, these results were not plausible given the implementation evidence, as we found nothing about the program that could have caused such large, unfavorable effects. The treatment and comparison beneficiaries were well matched on observable characteristics at baseline, and we found no notable differences in sample attrition during the intervention period between the two groups. As a result, we believe there might have been unobservable differences between the groups, likely as a result of the process through which

PMGs identified patients eligible for the program, which influenced the results. The Center for Medicare & Medicaid Innovation and other stakeholders could consider a number of changes to the design of similar programs in the future to increase the potential to draw conclusions about program impacts on patients' outcomes. Specifically, they could consider randomization or, if this is not possible, a program could require explicit eligibility criteria that can be replicated in claims data to allow for valid comparison selection.

This page has been left blank for double-sided copying.

I. INTRODUCTION

This report presents findings from the evaluation of the Health Care Innovation Award (HCIA) received by the Pacific Business Group on Health (PBGH), with a focus on program impacts on participants' outcomes. Section II provides an overview of PBGH's HCIA-funded program and the design of the impact evaluation. Section III describes the design and implementation of the intervention, including how the program could be expected to affect evaluation outcomes through changes in participants' and providers' behavior. In Section IV, we assess the evidence on the extent to which planned changes in providers' behavior occurred. Section V describes our methods for, and results from, estimating program impacts on participants' outcomes in four domains: quality-of-care processes, quality-of-care outcomes, service use, and spending. We are unable to draw conclusions about program impact due to concerns about likely biases in the impact estimates, but we still present the data for transparency and so that readers can judge the evidence for themselves. In Section VI, we discuss ways that the Centers for Medicare & Medicaid Services (CMS) or other stakeholders could modify the program design for future tests of interventions similar to PBGH to increase the chances of drawing reliable impact conclusions.

II. OVERVIEW OF PBGH'S HCIA-FUNDED INTERVENTION AND THE IMPACT EVALUATION

A. PBGH's HCIA-funded intervention

PBGH received $19.1 million in HCIA funding to implement the Intensive Outpatient Care Program (IOCP) (Table II.1, top panel). The IOCP is a model of care focused on managing high-risk, medically complex participants using a team-based approach to deliver highly individualized and accessible primary care, based on treatment goals specific to each participant. For the HCIA program, PBGH provided funding and technical assistance to 23 participating medical groups (PMGs) in five states to help them implement the IOCP. These PMGs in turn worked with primary care practices to enroll 15,008 participants, including those enrolled in Medicare Advantage, Medicare fee-for-service (FFS), those dually eligible for Medicare and Medicaid, and a small number of Medicaid-only beneficiaries, into the program from May 1, 2013, to June 2015 (when HCIA-funded operations concluded).

PBGH is a nonprofit coalition of 50 member businesses or public organizations that purchase insurance for their employees. PBGH has a record of working with health insurance plans, physicians and consumer groups, hospitals, and the California Health and Human Services Agency to achieve its mission of improving the quality and affordability of care. PBGH applied for HCIA funding to expand a care management model that served commercially insured high-risk populations in Puget Sound, St. Louis, Northern California, and Los Angeles; PBGH and some of its member businesses (the Boeing Company, the California Public Employees' Retirement System, and Pacific Gas and Electric Company) sponsored the model. This care management model demonstrated a 20 percent reduction in total costs and significant improvements in participants' functioning, mental health status, and ratings of care. Through the implementation of the HCIA-funded IOCP, PBGH intended to expand the care management

model to Medicare beneficiaries and to create a public–private partnership to support high-risk patients, regardless of payer.

The IOCP sought to reduce hospitalizations, emergency department (ED) visits, and total health care costs by 5 percent among Medicare FFS beneficiaries and Medicare Advantage patients by the end of the three-year program; it also intended to improve the quality of care delivered to participants as measured by improvements in quality-of-care process measures by 2 to 4 percent (Table II.1). PBGH expected to achieve these outcomes through a single intervention component: providing care management services to high-risk participants, whereby embedded care managers funded and trained by the program worked with participants' primary care providers to develop and implement personalized care plans (also known as the shared action plans) for medically complex, high-risk participants. The role of the care manager was to engage in one-on-one interactions with participants to understand their medical and social needs, work with participants and their physicians to help participants manage their conditions, provide participants and their caregivers with education and emotional support, and connect participants to appropriate community resources. PBGH expected that this intervention component would reduce the need for hospitalizations and post-acute care among high-risk Medicare beneficiaries. These reductions in acute care were expected, in turn, to reduce total Medicare spending (Section III.A.3 describes the awardee's theory of action in detail).

Table II.1. Summary of PBGH's HCIA program and our evaluation for estimating its impacts on patients' outcomes

Program description	
Award amount	$19,139,861
Award start date	June 2012
Implementation date	May 1, 2013
Award end date	June 30, 2015
Awardee description	PBGH is a nonprofit business coalition of 50 members, including public and private employers that purchase insurance for their employees such as Boeing, Disney, the University of California system, and Wells Fargo. PBGH and its members work with health insurance plans, physician and consumer groups, hospitals, and the California Health and Human Services Agency to improve access to and quality of health care without increasing costs.
Intervention overview	PBGH provided technical assistance and funding to 23 PMGs in five states to implement the IOCP, a care management program. IOCP is a model of care focused on managing high-risk, medically complex participants using a team-based approach to deliver highly individualized and accessible primary care, based on treatment goals specific to each participant.
Intervention components	**Care management for high-risk participants.** PBGH's program had a single component: providing care management services to high-risk participants. Care managers embedded in PMGs worked with participants' primary care providers to develop and implement personalized care plans (also known as shared action plans) for medically complex, high-risk participants. Care managers interacted with participants one on one to learn about their medical and social needs, help them manage their conditions, provide them and their caregivers with education and emotional support, and connect them to appropriate community resources. Believing that a one-size-fits-all implementation strategy would not be appropriate given the diversity of PMGs, PBGH granted PMGs the freedom to adapt the program to their specific needs while adhering to the programs' requirements (called guardrails).

Table II.1 *(continued)*

	Participants were enrolled for a minimum of 12 months. Care managers typically met with participants in person (in the participant's home or other location) at least once during the participant's enrollment. Care managers had to contact participants at least once a month—by telephone, in person, or online—for the duration of a participant's enrollment.
Target population	The program sough to reach 15,000 predicted high-risk Medicare beneficiaries, including those enrolled in Medicare Advantage, Medicare FFS, and those dually eligible for Medicare and Medicaid. PBGH expected the model to be most successful for Medicare patients with three or more chronic conditions without functional limitations and at least one recent hospitalization. PBGH revised the enrollment target from 27,000 to 15,000 in the ninth quarter following the award date (July through September 2014) because of challenges in meeting the original enrollment target. PBGH also added one PMG that enrolled a small number of Medicaid-only beneficiaries in the seventh quarter following the award date (January through March 2014).
Target impacts on patients' outcomes	• Reduce all-cause inpatient admissions by 5 percent • Reduce outpatient ED visits by 5 percent • Reduce Medicare Part A and B spending by 5 percent • Improve quality-of-care process measures by 2 to 4 percent • Improve quality-of-care outcome measures (amount not specified)
Workforce development	As of June 30, 2015, PBGH funded a total of 267.35 FTEs across 602 clinical sites, including (1) 14.45 FTEs who were licensed independent clinical practitioners authorized to prescribe medication; (2) 58.9 FTEs who were licensed practitioners not authorized to prescribe medication; (3) 166.9 nonlicensed clinical staff; and (4) 27.1 nonclinical staff. FTEs funded by PBGH included newly hired staff and existing PMG staff whose responsibilities changed. PBGH funded training for all care managers. In addition, PBGH funded two types of training for PMG leadership: (1) California quality collaborative leadership meetings, which were quarterly meetings for PMG administrators, during which they shared best practices for implementing the program; and (2) process improvement workshops, which were on-site workshops to identify specific actions and detail processes that facilitated participants' enrollment and improved fidelity to the care management model through adherence to IOCP requirements (guardrails); all PMGs had to participate in these workshops. PBGH also funded training for direct service staff. PBGH required all IOCP direct service staff to attend a three-day Care Coordinator Academy, at which they learned about program guardrails, participants' psychosocial issues, motivational interviewing, goal setting, and participant assessments and engagement. All IOCP direct service staff hired before July 1, 2014, attended the academy. The senior manager of clinical redesign at PBGH worked with all IOCP direct service staff hired after July 1, 2014, to provide the training.
Location	Multistate (Arizona, California, Idaho, Nevada, and Washington), select areas, predominately urban
Impact evaluation	
Core design	Contemporaneous differences with matched comparison group, adjusted for differences in baseline characteristics
Treatment group	Medicare FFS beneficiaries enrolled in the IOCP by any of the 23 participating PMGs from May 1, 2013, through March 31, 2015, based on lists provided by the awardee The treatment group was further restricted to those observable in Medicare FFS claims during the 12 months before they enrolled in the PBGH program and excluded a few beneficiaries who received hospice care in the year before enrollment. Lastly, we dropped 157 treatment group members during the matching process because they were outliers. The resulting treatment group consisted of 2,996 beneficiaries.
Comparison group	The comparison group consisted of Medicare FFS beneficiaries we matched to the treatment group beneficiaries on baseline characteristics (before the intervention began).

Table II.1 *(continued)*

Intervention component(s) included in impact evaluation	The component described in the previous section—care management for high-risk participants. Although implemented independently and with some variation, all PMGs focused on the component of care management to affect outcomes.
Extent to which the treatment group reflects the awardee's target population (for the component(s) evaluated)	**Low:** Our treatment group included only 20 percent of all program enrollees. The main reasons for exclusion were Medicare Advantage enrollment and missing patient identifiers.
Study outcomes, by domain	1. **Quality-of-care processes:** Preventive care for diabetes, lipid testing for patients with IVD, and 14-day follow-up to hospitalization 2. **Quality-of-care outcomes:** 30-day unplanned readmissions and inpatient admissions for ambulatory care-sensitive conditions 3. **Service use:** All-cause inpatient admissions and outpatient ED visits 4. **Spending:** Medicare Part A and B spending

Source: Review of PBGH reports, including its original application, operational plan, and 15 quarterly narrative reports to CMS.

CMS = Centers for Medicare & Medicaid Services; ED = emergency department; FFS = fee-for-service; FTE = full-time equivalent; HCIA = Health Care Innovation Award; IOCP = Intensive Outpatient Care Program; IVD = ischemic vascular disease; PBGH = Pacific Business Group on Health; PMG = participating medical group.

B. Overview of impact evaluation

To estimate the program's impacts on patients' outcomes, we compared outcomes for Medicare FFS beneficiaries participating in the HCIA intervention (the treatment group) with outcomes for beneficiaries in a matched comparison group, adjusting for differences in outcomes between these two groups before the intervention began. The bottom panel of Table II.1 summarizes our impact evaluation design.

We selected the comparison group beneficiaries for the evaluation from a pool of potential comparison group beneficiaries in the same or similar geographic locations as treatment group beneficiaries; we further limited this pool by excluding those beneficiaries who had received services from a treatment primary care provider (PCP) or from the PMG or individual practices (identified by Taxpayer Identification Number [TIN]) participating in the intervention.

We estimated impacts on outcomes, as measured in Medicare FFS claims data, which we grouped into four domains: (1) quality-of-care processes, (2) quality-of-care outcomes, (3) service use, and (4) spending. Across the HCIA awardees in primary care redesign (PCR), we designed our impact evaluations to identify promising interventions or intervention components—consistent with evaluation goals from the Center for Medicare & Medicaid Innovation (CMMI) to find programs that could be scaled or retested as part of a future study. Before conducting analyses, we specified a series of primary tests, describing the evidence we would need to conclude that the program was effective, and the awardee and CMMI reviewed these tests. Each test specified a population, outcome, period, expected direction of effect, and threshold that we count as substantively important. The purpose of these primary tests was to focus the impact evaluation on hypotheses that would provide the most robust evidence about program effectiveness. Because we sought to identify promise, rather than only those programs with unequivocally demonstrated success, we conducted one-sided statistical tests (that is, testing

only for program benefits) and used a threshold for statistical significance of 0.1, which is not as strict as the conventional standard of 0.05.

Our impact evaluation reflects the effects of the intervention on program enrollees who were Medicare FFS beneficiaries (including those dually eligible for Medicaid), who comprise 30 to 40 percent of program enrollees. Because the evaluation's treatment group was limited to Medicare FFS beneficiaries, the evaluation design only partially aligns with PBGH's HCIA intervention; it does not capture impacts on Medicare Advantage beneficiaries (the majority of the remaining 60 to 70 percent of program enrollees) or the small number of Medicaid beneficiaries enrolled by the program at one site.

III. PROGRAM IMPLEMENTATION

This section first provides a detailed description of PBGH's HCIA-funded program, highlighting how it evolved over time and its theory of action. Second, it assesses the evidence on the extent to which the intervention was implemented as planned based on measures of program enrollment, service delivery, staffing, training, and timeliness. Third, this section summarizes the facilitators and barriers associated with implementation effectiveness.

We based our evaluation of PBGH's program implementation on a review of its quarterly reports to CMMI and self-monitoring program metrics, telephone discussions and follow-up communications with program administrators, and information collected during site visit interviews with frontline staff conducted in April 2014 and April 2015. We did not verify the quality of the performance data reported by PBGH in its self-measurement and monitoring reports.

A. Program design and adaptation

1. Target population and participant identification, recruitment, and enrollment

In this section, we describe how PBGH recruited PMGs to implement the IOCP, and how PMGs identified, recruited, and enrolled high-risk, medically complex Medicare beneficiaries eligible for the IOCP.

Participating medical groups. PBGH contracted with 23 PMGs located in five states (Arizona, California, Idaho, Nevada, and Washington State) to implement the IOCP. PMGs were either integrated health systems or independent practice associations (IPAs). Integrated health systems generally employ physicians and, therefore, exercise some authority over the care delivery process. In addition, health systems generally use a centralized electronic health record (EHR) system and are geographically near their physicians, which facilitates communication. IPAs, on the other hand, contract with independently practicing physicians affiliated under a management services organization. IPAs are less integrated because the physicians operate independently and can be geographically dispersed. Physicians affiliated with IPAs exercise more autonomy over care delivery compared with physicians working in integrated health systems, often resulting in diverse health information technology (IT) systems and varying levels of engagement with the parent entity. A breakdown of PMGs that were either integrated health systems or IPAs was not available at the time we wrote this report.

Recruitment of participating medical groups. PBGH recruited 8 medical groups in 2012 and 15 in 2013. PBGH reached out to medical groups with which it already had relationships, or who were recommended by other health plan or PMG leadership, to gauge their interest in implementing the IOCP. PBGH recruited medical groups based on (1) leadership committed to the goals of the IOCP, (2) capacity to implement the IOCP, in terms of clinical infrastructure and health IT capacity, and (3) a relatively large number of Medicare patients. PBGH did not have explicit eligibility rules by which to judge these criteria. A fourth decision factor was for the medical groups to already have financial incentives in place to target and offer care management services to patients with chronic illnesses who were at high risk of experiencing a hospitalization. PMGs that served Medicare Advantage and Medi-Cal (California dually eligible) beneficiaries had an incentive to implement the IOCP model because those PMGs could retain the savings expected to result from more intensive outpatient care leading to reduced hospital admissions and ED visits. PMGs that served Medicare FFS beneficiaries had an incentive to implement the IOCP model to prepare for accountable care organization (ACO) contracts with CMS and position themselves to contract for commercial ACO-style products. During the second year of PMG recruitment, fewer PMGs expressed interest in implementing the IOCP than during the first year due to competing initiatives resulting from the Patient Protection and Affordable Care Act. As a result, PBGH relaxed its recruitment criteria and contracted with PMGs with leadership committed to the goals of the IOCP, but without demonstrated capacity to implement the program. In the second year of recruitment, PBGH also contracted with PMGs that met the recruitment criteria but were located outside of the two states originally targeted in the first year of recruitment (Arizona and California).

Target population of participants. The IOCP, which targeted patients with chronic illness who were at a high risk of experiencing a hospitalization, initially expected to enroll 27,000 participants. PBGH developed the following eligibility guidelines for the IOCP, but allowed primary care providers to use their discretion when evaluating whether Medicare beneficiaries' risk factors made them eligible for the program. PBGH initially excluded patients with end-stage renal disease and patients under established care pathways for cancer treatment, but these limitations were removed based on PMG feedback; PMGs expected these patients to have potential unmet care coordination needs and thus benefit from the program. All Medicare beneficiaries (that is, Medicare FFS and Medicare Advantage beneficiaries, including those dually eligible for Medicare and Medicaid) were eligible for the program if they received care at a primary care practice affiliated with one of the PMGs, and had one or more of the following risk factors:

- Three or more hospitalizations and/or ED visits in the past six months
- Three or more chronic conditions
- Eight or more medications
- A combination of factors, including the following:
    - One chronic condition and one of the following: two or more ED visits or hospitalizations in the past six months, or age 65 or older and five or more medications; or

  - Clinical referral by the patient's provider and one of the following: two chronic conditions, three or more specialists, or expectation by the patient's provider that the patient might end up in the hospital or die in the next six months

- In addition to meeting at least one of these risk factors, demonstrated fragmentation of care (for example, no primary care provider or more than two primary care providers)

Identification of eligible patients. PBGH adapted its patient identification process during the course of the award. Originally, PBGH expected all PMGs to identify high-risk patients using the Milliman Advanced Risk Adjustors (Milliman) model to calculate risk scores based on PMGs' Medicare FFS claims and Medicare Advantage encounter data. (Milliman is a consulting and actuarial firm that processes Medicare claims through a proprietary algorithm to provide health care providers with risk scores for their patients.) However, the complexity of Milliman's standards for importing claims into their databases and the varied quality of the claims data submitted by PMGs delayed the development of standard risk score reports. Therefore, PBGH developed alternative methods for identifying high-risk patients:

1. **Direct referral.** PCPs within the PMGs assess that a patient would benefit from the program and directly refer the patient to the IOCP.
2. **Transfer from existing care management program.** IOCP care management staff determine that the services provided by the IOCP would be more appropriate for a patient than a care management program in which the patient is currently enrolled. For example, a patient enrolled in a disease management program might benefit from the social service aspects of the IOCP.
3. **Identification through hospital records.** IOCP care management staff review hospital records (if available through integrated health systems and partnering hospitals for IPAs) to identify patients with three or more hospitalizations or ED visits in the past six months.
4. **Identification through internal reporting.** IOCP care management staff review internal reports developed by each PMG to identify patients who have seen three or more specialists, have three or more active monitored conditions, or have been on five or more medications.
5. **Other methods.** Care management staff use other methods to support patient identification, including identifying high-risk patients during their stay in a hospital or skilled nursing facility.

When Milliman's standard risk score reports became available in early 2014, PMGs found them unhelpful because the data used to create the lists were three to six months old. Instead, PMGs continued to use the alternative methods for identifying patients. In interviews, program administrators and frontline staff said these better suited their patient populations and participating providers.

Recruitment and enrollment of IOCP participants. PMGs experienced challenges with patient recruitment, which required additional program adaptation. During the first few months of program implementation, care management staff reached out to patients identified as eligible by cold-calling them to introduce the program and describe the benefits of personalized care

management services to the patient. Patients were not receptive to this approach, likely because they were unfamiliar with the care management staff, so PMGs implemented new recruitment strategies. These new strategies included a so-called warm hand-off approach, in which PCPs introduced high-risk patients to the care manager during a visit, and other strategies, such as approaching high-risk patients in person during their stay in a hospital or skilled nursing facility. Care management staff discussed the benefits of the program with the patient, such as connecting patients to needed services, including transportation, home health visits, Meals on Wheels, behavioral health or substance abuse services, monitoring the patient's overall health, and communicating urgent needs and new developments to the patient's PCP. If the patient was interested in the program, the care manager obtained the patient's intent to participate and scheduled an initial one-on-one, face-to-face visit to enroll the patient.

2. Intervention component

PBGH's IOCP intervention had one component—care management for high-risk patients. Care management services focused on developing personalized care plans, also known as shared action plans, for the participants. Care managers were primarily clinical staff (Section III.B.4) and were trained to develop the shared action plan by reviewing the participant's chart; soliciting input from both the participant and the participant's PCP; and administering three patient assessments: (1) Patient Health Questionnaire, (2) Patient Activation Measure, and (3) Veterans RAND 12-Item Health Survey. The shared action plan contained at least one specific, measurable, and attainable goal per year, such as being able to stand without assistance or quitting smoking.

Care managers were required to communicate with participants at least monthly—by telephone, in person, or online—although some participants were contacted more frequently depending on their needs. Care managers typically met with a participant in person, in the participant's home or other location, at least once during the participant's enrollment. During these encounters, the care manager and participant jointly updated the participant's shared action plan as the participant's health progressed.

Participants' conditions typically took a few months to stabilize after program enrollment. During those first few months of enrollment, care managers often communicated with the participant more than the once-a-month minimum. After the participant's condition stabilized, the care manager's contact with the participant typically dropped to once a month at a minimum. Care managers disenrolled participants from the IOCP for any of the following three reasons:

1. **Graduation.** The participant completed 12 months of enrollment in the program and successfully completed the shared action plan or stabilized his or her condition.
2. **Drop out.** The participant died, declined to continue participation, or was lost to follow-up.
3. **Enrolled in error.** Care managers discovered the participant was ineligible after enrollment. A participant could be deemed ineligible if he or she (1) was unable to consent due to cognitive impairments (such as dementia), (2) was not a Medicare beneficiary, or (3) did not have multiple chronic conditions or other complex medical issues.

PBGH provided PMGs with the following resources to support the IOCP implementation.

1. **Funding.** In Year 1, PBGH paid PMGs $230,000 plus a per member per month payment for every participant enrolled ($10 for Medicare Advantage beneficiaries and $20 for Medicare FFS beneficiaries); in Year 2, the payment was $115,000, plus the per member per month payments. The payments were intended to cover the salary of two PMG-level staff: a program manager to coordinate IOCP implementation was funded for 12 months and a technology analyst to coordinate IT systems, claims and encounter data, and file security was funded for 18 months.
2. **Technical assistance and training.** PBGH provided collaborative technical assistance to PMGs to guide leadership through the IOCP implementation process and support peer-to-peer learning. This collaborative technical assistance included regular meetings and telephone calls with PMG leadership and site visits to support IOCP implementation. PBGH organized a care management training for care management staff, organizational leaders, and other staff. PBGH developed program requirements for the PMGs to use as a guide during program implementation, and allowed PMGs to adapt the program to meet their participants' specific needs.

3. Theory of action

Based on extensive review of PBGH's program activities and goals, we developed a theory of action to depict the mechanisms through which IOCP administrators expected the program to improve the outcomes we selected for the impact evaluation (Table II.1 includes a list of these outcomes). PBGH expected that its HCIA-funded intervention would improve outcomes for Medicare beneficiaries through one pathway, care managers, embedded in primary care practices, working with PCPs to provide personalized care management services to high-risk patients. In particular, PBGH expected to improve quality of care, reducing the frequency of acute exacerbations and therefore reducing service use and spending outcomes, through the following steps:

1. **Personalized care plans increase relevance of the services in meeting the participant's needs.** To develop each care plan, the care manager conducts a holistic participant assessment, including gathering input from the participant; the PCP; and the three patient assessment tools (Patient Health Questionnaire, Patient Activation Measure, and Veterans RAND 12-Item Health Survey). These assessment tools were selected to measure patient-reported outcomes at baseline and subsequent program periods. The care manager uses motivational interviewing techniques to identify the patients' own health-related goals, their readiness for change, and concrete, short-term actions that can be taken to meet their goals. The personalized care plan and assessment tools ensure the care manager understands the participant's needs and attitudes and can tailor services provided to the participant, thus improving the relevance of the services in meeting the participant's needs.
2. **The participants change their self-care behavior.** PBGH granted PMGs the freedom to adapt the program to their participants' specific needs while adhering to certain program requirements (guardrails). Therefore, across PMGs, care managers helped participants engage in a variety of activities to change their self-care behavior, such as increasing

physical activity, improving their diet, adhering to medications, and accessing needed community resources.

3. **Concurrent with step 2, communication between care managers and participants helps the care managers to identify gaps in care, or worsening of conditions, and to contact PCPs to intervene as appropriate.** This leads to improvements in general preventive care and more timely changes in treatment regimens when the participants' conditions worsen or change.

4. **The improvements in self-care (step 2) and clinical care (step 3) lead to improvements in participants' health overall, resulting in fewer acute events requiring hospitalization or ED visits.** By reducing the frequency of these costly events, the program reduces overall Medicare spending.

Text box III.1. Example from PBGH illustrating the program's theory of action

"…[A participant] who lives alone was referred by PCP for evaluation for home safety. Initial medication review in the home revealed that member had not been refilling medications properly. There were 5 medications he was taking but he should have been taking 12, according to the last office note. The office note stated that member verified all of the medications and was taking appropriately. Member has diabetes, CHF, and COPD. [The care manager] arranged for a Home Health Nurse to do medication set up and monitor on a regular basis. Member's blood sugar is in better control (200 instead of 275, working on that next) because he is taking the right medications. Member now has safety monitoring device that he wears at all times in case of a fall. [Physical therapist] and [occupational therapist] are working with member for strengthening and safety exercises."

Source: Awardee's quarterly reports to the Center for Medicare & Medicaid Innovation.

CHF = congestive heart failure; COPD = chronic obstructive pulmonary disease; PCP = primary care provider.

4. Intervention staff and workforce development

IOCP care management staff included a combination of licensed and non-licensed personnel, including registered nurses (RNs), social workers, pharmacists, medical assistants, and licensed practical nurses (LPNs) (Table III.1). Licensed care managers (RNs, social workers, and pharmacists) conducted the initial consultations with enrolled participants and led care teams in assessing participants' needs and developing shared action plans. Unlicensed program staff (medical assistants) were important members of the care team as well, especially because of their ability to connect participants to needed community resources. After a licensed care manager developed a participant's shared action plan, the unlicensed care manager often took the lead in executing the plan by connecting participants to community resources, regularly following-up with participants, and monitoring participants' conditions.

Table III.1. Key details about program staff

Staff members	Staff/team responsibilities	Adaptations?
Registered nurses, pharmacists, and social workers	Conduct initial consultation with enrolled participants; lead care teams in assessing participants' medical needs and developing shared action plans	No
Medical assistants, licensed practical nurses	Execute the shared action plan by connecting participants to community resources, regularly follow-up with participants, and monitor participants' conditions	No

Sources: Interviews from second site visit, April 2015; document review, March 2015.

Note: Type of staff hired varied by site.

PBGH sent all care management staff to a Care Manager Academy that PBGH funded. This three-day training covered communicating with and engaging participants, assessing participants' psychosocial issues, motivational interviewing, goal setting, using evidence-based guidelines for managing chronic conditions, coordinating care transitions, understanding health insurance coverage, and working with family caregivers and community agencies. Care managers used the skills they developed or enhanced during the Care Manager Academy to guide their encounters with participants and help them achieve their goals.

As part of the collaborative technical assistance, PBGH also funded two types of training for PMG leadership: (1) California Quality Collaborative leadership meetings and (2) process improvement workshops. Beginning in July 2013, PMG administrators attended quarterly California Quality Collaborative Leadership trainings in which they shared best practices for implementing the program.

B. Implementation effectiveness

In this section, we examine the evidence on implementation effectiveness—that is, we analyze measures of the intervention delivered and, when possible, compare those measures with the services the awardee intended to deliver. We assess the evidence on implementation effectiveness in five areas: (1) program enrollment, (2) service delivery, (3) staffing, (4) training, and (5) implementation timeliness. To conduct this analysis, we used data from interviews with program administrators and frontline staff, and self-reported metrics included in PBGH's self-monitoring and measurement reports to CMMI.

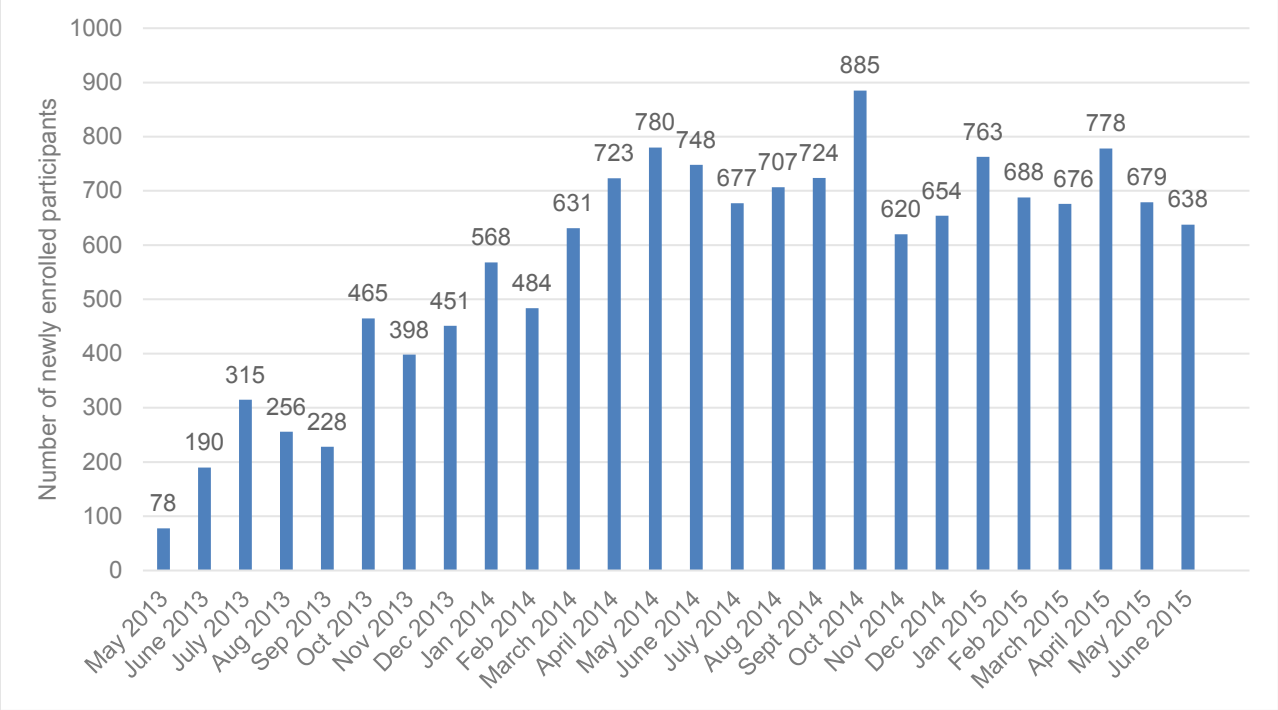
1. Program enrollment

PBGH reported that implementation started on May 1, 2013; however, it noted a few PMGs began enrolling participants in the IOCP starting in April 2013. PBGH changed its enrollment target from 27,000 to 15,000 in the ninth quarter following the award date (July through September 2014) because of challenges producing standard risk score reports for PMGs to identify patients eligible for the program (Section III.A.1). PBGH met this new target; as of June 30, 2015, PMGs had enrolled 15,008 participants. As shown in Figure III.1, the number of participants enrolled each month generally increased during the award period, particularly after the PMGs recruited in 2013 began enrolling participants in the first quarter of 2014, and as PMGs developed more effective methods of identifying, recruiting, and enrolling participants.

Figure III.1 also shows that PMGs continued to enroll participants in the last year of the program.

As of June 30, 2015, 44 percent of all enrolled participants (6,558) had been disenrolled from the IOCP. The remaining 8,450 participants remained enrolled in the program and continued receiving care management services similar to those offered through the HCIA-funded IOCP after the award ended. Of the 6,558 disenrolled participants, 34 percent (2,230) successfully completed the IOCP, meaning they were enrolled for at least 12 months and stabilized their conditions or completed their shared action plans; 56 percent (3,673) dropped out of the program due to death, loss to follow-up, or declined to continue participation; and 10 percent (656) were disenrolled because they were found to be ineligible after enrollment. Even though only 34 percent of disenrolled participants—or roughly 15 percent of all participants—successfully completed the program by July 2015, the month after HCIA-funded operations concluded, the 8,450 participants who remained enrolled in the program after the end of the HCIA period received care management services as intended throughout the award period.

Figure III.1. Total number of IOCP participants enrolled from May 2013 to June 2015



Source: Awardee's measurement and monitoring report to the Center for Medicare & Medicaid Innovation, June 2015.

Notes: Data shown include participants enrolled only from the IOCP implementation date (May 1, 2013) through the award end date (June 30, 2015). PBGH reported that the program enrolled 204 participants in April 2013, but we did not show those participants in this graph because PBGH reported May 2013 as the first month of enrollment. PBGH reported that it enrolled a total of 15,008 participants from the start (July 2012) through the end of the award.

IOCP = Intensive Outpatient Care Program; PBGH = Pacific Business Group on Health.

2. Service-related measures

PBGH reported an average of 9.6 encounters per participant by a care manager over the award period. Although PBGH's stated goal was to contact each participant at least once a month for at least 12 months, PBGH reported that when participants first enrolled in the program, care managers contacted participants weekly or biweekly, because participant's conditions were typically not stable (as determined by the PCPs and care management staff and based on the individual participant's needs). Then, the frequency of contacts dropped to monthly after the participant's condition stabilized. It is possible that the frequency of contacts might have dropped to less than monthly depending on a participant's needs and engagement in the program. Another potential explanation for the PMGs not meeting the goal of 12 encounters per participant is that PMGs continued to enroll participants until June 2015; participants enrolled in the last 11 months (August 2014 through June 2015) would not have been in the program for a year and therefore would not have had 12 encounters, if encounters, on average, occurred monthly.

PBGH aimed to have 100 percent of participants complete a shared action plan within one month after enrollment, but did not meet this target. As of June 30, 2015, PMGs reported that 92 percent of participants had a shared action plan. This is an increase since PBGH first reported the measure (62 percent) in a report covering the period April through June 2014.

3. Staffing measures

By July 2015, the PMGs had hired a total of 267.35 full-time equivalents (FTEs). The PMGs exceeded PBGH's original three-year target of hiring 211 FTE new hires. Most FTE new hires were nonlicensed clinical staff, such as patient navigators or health educators, who were staffed on IOCP care teams. Each care team had multiple nonlicensed care managers. Nonprescribing licensed care managers (RNs, social workers, and pharmacists) headed the care teams. PMGs also hired licensed independent clinical practitioners authorized to prescribe medications (physicians or mid-level providers) and nonclinical staff (management, administrative, or IT staff) (Table III.2).

Notably, although PBGH reached its FTE new hire target ahead of schedule, the ability to recruit and hire new program staff varied by site. One of the PMGs we visited could not find permanent, full-time care managers and instead relied on part-time temporary staff who often had little or no relevant care management experience. Another site we visited remained understaffed until 2015, which might have affected participant enrollment.

Table III.2. PBGH FTE measures

Staffing metrics	Initial awardee target (source)	Actual	Target met or exceeded
Cumulative FTE new hires	211	267.35	Yes
Licensed independent clinical practitioners authorized to prescribe medication[a]	n.a.	14.45	n.a.
Licensed clinical practitioners *not* authorized to prescribe medication[b]	n.a.	58.9	n.a.
Nonlicensed clinical staff[c]	n.a.	166.9	n.a.
Nonclinical staff[d]	n.a.	27.1	n.a.

Source: Awardee documents.

[a] Licensed independent clinical practitioners authorized to prescribe medication include nurse practitioners, physician assistants, and physicians.
[b] Licensed clinical practitioners *not* authorized to prescribe medication include nurses, social workers, and pharmacists.
[c] Nonlicensed clinical staff include aides/assistants/direct-care workers, behavioral/mental health workers who are not physicians, care coordinators/case managers/patient navigators, care transition specialists, clinical support staff, community health workers, health educators/health coaches, pharmacy technicians, and other types of health workers.
[d] Non-clinical staff include IT technicians/specialists and management or administrative staff.
FTE = full-time equivalent; PBGH = Pacific Business Group on Health.
n.a. = not applicable.

4. HCIA-funded training

PBGH used its HCIA to fund a variety of strategies to train care management staff. As described previously, PBGH required all frontline staff delivering direct care management services to attend a three-day Care Manager Academy. The academy aimed to teach program requirements and basic principles of care management, including communicating with and engaging participants, assessing participants' psychosocial issues, motivational interviewing, goal setting, using evidence-based guidelines for managing chronic conditions, coordinating care transitions, understanding health insurance coverage, and working with family caregivers and community agencies. Multiple webinars per year complemented the Care Manager Academy; the webinars covered topics such as managing a participant's end of life, burn-out, and motivational interviewing. Webinars were topic-driven, with both didactic and interactive components, and 50 to 100 people from several different PMGs regularly attended them. Based on feedback from PMGs, PBGH expanded training for care management staff in the last calendar quarter of 2014 to include peer clinical case conferences. At these conferences, newly hired care managers attended training and existing care managers discussed challenging clinical cases and shared success stories with their peers under the facilitation of IOCP clinical advisors.

To assess perspectives of HCIA-funded frontline staff who received this training, we administered the HCIA Primary Care Redesign Trainee Survey from January to March 2015 (21 to 23 months after the start of IOCP implementation in May 2013). For the purposes of this report, we analyzed the responses from clinical staff (nurses, care coordinators, social workers, pharmacists, and physicians), but not administrative or management staff. Stratifying our results enabled us to report on the trainings relevant to care management service delivery, rather than

leadership and other training also provided by PBGH (Section III.A.4). The stratification process is imprecise because we must rely on staff titles provided by PBGH, which sometimes were unclear. We estimate that we sent the survey to 190 frontline care management staff. Of these, 99 completed the survey (a response rate of 52 percent).

Almost all respondents (98 percent) reported receiving formal training and rated the training as good (28 percent) or excellent (67 percent) (data not shown). Respondents who reported receiving formal training (98 percent) either agreed or strongly agreed that the content of the training was relevant (94 percent), useful (93 percent), and improved the respondent's performance or helped the respondent complete his or her job responsibilities (88 percent) (Table III.3). Most respondents also agreed or strongly agreed that the training was delivered effectively in a number of domains. Most respondents also believed that the training positively affected their ability to provide care management services. Specifically, most respondents believed the training positively affected the quality (74 percent) and the patient-centeredness (73 percent) of care they provide. More than half of respondents who received training also reported that the training positively affected their ability to perform specific activities of care management, such as explaining information about patient care to participants and their families in lay terms (58 percent), relaying relevant information to the care team (55 percent), helping participants access the care they need (61 percent), and helping participants take control of their own care (70 percent).

The survey data showed that the types of care management activities delivered by clinical staff varied (Table III.4). The most common activities routinely provided by clinical staff included calling participants to check up or help coordinate care (69 percent); educating participants about managing their own care (72 percent); counseling participants on exercise, nutrition, and how to stay healthy (70 percent); following up on transitions of care (62 percent); participant coaching (63 percent); and attending team meetings and care conferences (72 percent). However, up to 19 percent of respondents reported that they did not provide these services. As expected, survey respondents were less likely to perform duties that required clinical licensing—such as executing standing orders for medication refills, ordering tests, or delivering routine preventive services—because only some clinical staff were licensed to do them. Fewer than half of survey respondents (44 percent) routinely assisted participants with accessing nonmedical services such as housing, job training, or supplemental nutrition services.

We do not have measures on the training PBGH provided to PMG leadership, but we do have qualitative data collected during site visits about PMG leaders' experiences with the training. PMG administrators we interviewed during site visits noted that the trainings helped them build relationships with PBGH leaders and leaders at other PMGs. The PMG administrators also said that they learned from other PMG administrators about best practices, new tools, and lessons learned that helped them to improve IOCP implementation at their organizations. For example, administrators we interviewed decided to use triage nurses to help identify potential participants after learning that one PMG improved efficiency in participant identification and enrollment through the use of triage nurses. PBGH also required all PMGs to participate in on-site workshops. During the workshops, PBGH identified specific actions and processes to improve participant enrollment and fidelity to the care management model.

Table III.3. Trainees' perceptions of training's content and delivery

Survey question		Percentage of 99 respondents (and number)[a]	
		Strongly agreed	Agreed
Thinking [about the training], please indicate your level of agreement with the statements listed to the right.[b]	1. The objectives of the training were clearly defined.	83% (80)	13% (13)
	2. The topics covered were relevant to me.	75% (73)	19% (18)
	3. The content was organized and easy to follow.	87% (84)	NA[c]
	4. The training experience will be useful in my work.	80% (78)	13% (13)
	5. The trainer was knowledgeable about the training topics.	89% (86)	NA[c]
	6. The trainer was well prepared.	89% (86)	NA[c]
	7. The training objectives were met.	84% (81)	13% (13)
	8. The training helped me to improve my performance or complete my new job responsibilities.	68% (66)	20% (19)
	9. The training was delivered at a comfortable pace so I could understand the content.	77% (75)	19% (18)
		Positive Impact	Negative Impact
Please indicate the impact you believe the training you received for the IOCP has had on the following aspects of care you provide to patients enrolled at your practice site.[d]	1. Quality of care	74% (73)	NA[c]
	2. Ability to respond in a timely way to patients' needs	57% (56)	NA[c]
	3. Efficiency, cost-effectiveness of care	49% (48)	NA[c]
	4. Patient-centeredness (providing care that is respectful of and responsive to an individual patient's needs, preferences, and values)	73% (72)	NA[c]
	5. Equity of care for all patients (providing care that does not vary in quality because of the demographic characteristics of the patient)	64% (63)	NA[c]
Please indicate whether the training you received has had a positive or negative effect on your ability to… [d]	1. Explain information about patient care to patients and their families in lay terms	58% (57)	NA[c]
	2. Relay relevant information to the care team	55% (54)	NA[c]
	3. Work with diverse set of patients	56% (55)	NA[c]
	4. Access the care they need	61% (60)	NA[c]
	5. Help patients access nonmedical services	49% (48)	NA[c]
	6. Help patients take control of their own care	70% (69)	NA[c]
	7. Use data to evaluate my performance to improve the services I provide to patients	49% (48)	NA[c]

Source: Mathematica's analysis of HCIA Primary Care Redesign Trainee Survey.

[a] Missing responses are excluded from the denominator when calculating percentages.

[b] Possible answers for this question included strongly agree, somewhat agree, neither agree nor disagree, somewhat disagree, strongly disagree, not applicable, or missing.

[c] Fewer than 11 respondents. Results suppressed to protect respondents' confidentiality.

[d] Possible answers for these questions included positive impact, negative impact, no impact, too soon to tell, not applicable, or missing.

HCIA = Health Care Innovation Award; IOCP = Intensive Outpatient Care Program.

NA = not available.

Table III.4. Trainees' responsibilities in managing patients' care

Survey question		Percentage of 99 respondents (and number) who reported that they:		
		Yes, routinely	Yes, occasionally	No
Please indicate whether you personally help to manage patients' care in any of the following ways:	1. Call patients to check on medications, symptoms, or help coordinate care in-between visits	69% (68)	NA[a]	19% (19)
	2. Execute standing orders for medication refills, ordering tests, or delivering routine preventive services	25% (25)	23% (23)	50% (49)
	3. Educate patients about managing their own care	72% (71)	NA[a]	18% (18)
	4. Counsel patients on exercise, nutrition, and how to stay healthy	70% (69)	NA[a]	19% (19)
	5. Assist patients with accessing nonmedical services such as housing, job training, or supplemental nutrition services (for example, SNAP benefits)	44% (44)	26% (26)	27% (27)
	6. Attend medical appointments with patients	25% (25)	25% (25)	47% (47)
	7. Conduct home visits with patients	29% (29)	20% (20)	50% (49)
	8. Follow up on transitions of care	62% (61)	18% (18)	18% (19)
	9. Patient coaching	63% (62)	13% (13)	20% (20)
	10. Attend team meetings/care conferences	72% (72)	17% (17)	NA[a]

Source: Mathematica's analysis of HCIA Primary Care Redesign Trainee Survey.

[a] Fewer than 11 respondents. Results suppressed to protect respondents' confidentiality.

HCIA = Health Care Innovation Award; SNAP = Supplemental Nutrition Assistance Program.

NA = not available.

5. Program timeline

PBGH experienced initial implementation delays due to challenges developing standard risk score reports for PMGs to identify patients eligible for the program. By the time PMGs first enrolled participants in early 2013, PBGH expected to have risk scores calculated by Milliman to help PMGs identify participants. As previously described, PMGs struggled to submit consistent, high quality claims data to Milliman because of the complex claims-reporting requirements, which led to delays in developing standard risk score reports. As a result, PMGs received the standard risk score reports in early 2014 (January through March).

C. Summary of facilitators of and barriers to implementation

Several factors facilitated implementation of PBGH's HCIA-funded program, but others posed barriers. We described those factors in detail in the second annual report (Keith et al. 2015). Here we summarize key facilitators and barriers, along with any new information since the second annual report that supports those facilitators or barriers (Table III.5).

During our two rounds of site visits in 2014 and 2015, we learned about three key factors that facilitated program implementation and two key factors that challenged it. First, program leaders attributed the program's success to the adaptability of the program design, particularly having the flexibility to develop alternative methods for identifying, recruiting, and enrolling IOCP participants to meet enrollment goals. Second, stakeholders perceived benefits of the program compared with the standard delivery of care. Specifically, administrators, PCPs, and care management staff generally agreed that the care managers were improving the quality of care delivered to high-risk participants and reducing the burden on PCPs. Third, PCPs and program staff described how having care management staff—including nurses, pharmacists, and social workers—embedded in primary care practices facilitated communication within the care team. In contrast, two factors emerged as challenges to program implementation. First, care managers struggled to meet the needs of all IOCP participants. Care managers could not adequately communicate with participants who had behavioral or mental health issues or severe cognitive impairments, and had difficulty changing the participants' attitudes about changing health behaviors. Second, physician engagement in care management varied across the PMGs. This variation was most apparent between the two organizational structures of the PMGs: integrated health systems and IPAs. Under the IPA structure, physicians have more autonomy over care delivery and less engagement with the parent entity (the PMG) compared with physicians in integrated health systems, which generally employ physicians and therefore have some authority over care delivery. PBGH confirmed in its final report to CMS that physician engagement continued to present a great challenge to program implementation throughout the entire award period.

Table III.5. Summary of key facilitators and barriers to the implementation of PBGH's program

Item	Description based on findings in the second annual report	Additional supporting data not available in the second annual report, if applicable.
	Facilitators (domain)	
Adaptability of the program	Program leaders attributed the program's success to the flexibility of its design. During the first year of participant enrollment, program leaders described how they tracked enrollment and met regularly with program staff to change and adapt the program to achieve program goals. For example, during the initial phase of implementation, some PMGs faced challenges reaching enrollment targets and implemented a variety of changes to increase enrollment, including expanding the target population, hiring part-time care managers, and recruiting participants in person instead of by telephone.	
Perceived relative advantage of the program compared with the standard delivery of care	Administrators, PCPs, and care managers perceived that the care managers improved the quality of care delivered to high-risk participants with complex conditions and reduced burden on PCPs, compared with the status quo way of delivering care to this population. PCPs described how care managers provided information that they otherwise would not have obtained, especially regarding psychosocial issues and physical barriers in participants' homes. Care managers were better positioned than PCPs to monitor participants' adherence to their care plans and check in with them regularly, thereby offering dependable emotional support and practical health education. PCPs also commented that care managers were more likely to deliver timely and patient-centered services compared with other ancillary providers, such as home health agencies.	
Embedded care managers helped improve communication between physicians and care managers	PCPs and program staff described how having care management staff embedded in primary care practices facilitated communication within the care team. Because physicians were not always linked to a central EHR system, face-to-face interactions with care coordinators was an efficient alternative. Physicians also expressed comfort communicating with care management staff via email, text message, or telephone that resulted from personally knowing care managers.	
	Barriers (domain)	
Inability of providers to meet the needs of every participant	Care management staff found it difficult to meet the needs of certain participants. For example, care managers could not adequately communicate with participants with behavioral and mental health issues and severe cognitive impairments, and could not change the attitudes of participants who were unwilling to engage in their health care.	

Table III.5 *(continued)*

Item	Description based on findings in the second annual report	Additional supporting data not available in the second annual report, if applicable.
Physician engagement varied across PMGs but was necessary to successfully implement the program	IPAs faced unique challenges to implementation, especially with regard to engaging physicians and identifying participants using health records. Engaging independent physicians proved challenging in some cases because they were often accustomed to working autonomously, had little or no in-person contact with the PMG, and were not obligated to respond to the PMG. In addition, physicians in an IPA are not part of a centralized EHR system, limiting the PMG's access to participants' data needed to identify and monitor participants.	In a report to CMMI, PBGH reported that physician engagement was a great challenge The awardee based this assessment on PMGs' responses to an IOCP survey administered by PBGH.

Note: We reviewed four domains associated with implementation experience: (1) program characteristics, (2) implementation process, (3) internal factors, and (4) external environment. Implementation research suggests that barriers and facilitators within these domains are important determinants of implementation effectiveness.

CMMI = Center for Medicare & Medicaid Innovation; EHR = electronic health record; IOCP = Intensive Outpatient Care Program; IPA = independent practice association; PBGH = Pacific Business Group on Health; PCP = primary care provider; PMG = participating medical group.

D. Conclusions about the extent to which the program, as implemented, reflects the core design

Despite PBGH implementing aspects of its HCIA-funded program as planned, the evidence from our implementation study shows that PBGH faced significant challenges in producing standard risk score reports for PMGs to identify patients eligible for the program. The challenges faced by the PMGs in submitting Medicare FFS claims and Medicare Advantage encounter data to Milliman delayed the development of standard risk score reports. When the standard risk score reports became available in early 2014, PMGs found them unhelpful because the data used to create the lists were three to six months old. Instead, PMGs used alternative methods to identify participants who better suited their participant populations and participating providers. These challenges with identifying high-risk patients eligible for the program resulted in PBGH deviating from the original program design by reducing the enrollment target by almost 50 percent (from 27,000 to 15,000 IOCP participants), more than halfway through the award period.

Aspects of the program that PBGH implemented as planned include meeting staffing and training goals for care management staff and generally providing services to participants as planned. The PMGs exceeded their target for FTE new hires to support the program; they aimed to hire 211 FTEs by June 30, 2015, and actually hired 267.35 FTEs. Based on survey results from the HCIA Primary Care Redesign Trainee Survey, clinical program staff generally reported that the training they received as part of program implementation was excellent and improved their ability to provide care that aligned with the goals of the IOCP. These survey results also indicate that most respondents provided services planned for the IOCP, including routinely calling participants to check on medications, assess symptoms, or help coordinate care between visits; educating participants about managing their own care; counseling participants on exercise, nutrition, and how to stay healthy; providing coaching to participants on how to achieve their

goals; and following up with participants during care transitions. These survey results suggest that IOCP participants were contacted regularly. However, we cannot conclude that the program met the goal of contacting each participant at least once a month for 12 months. PBGH reported an average of 9.6 encounters per participant; the percentage of participants with a shared action plan was also lower than the goal of 100 percent (92 percent). Although the two service measures seem to fall short of PBGH's targets, this might not be the case. Because PMGs enrolled participants until June 30, 2015 (6,381 in the last three quarters), it is possible that the participants who did not have a shared action plan had not yet been enrolled in the IOCP for one month and the participants who had fewer than 12 encounters had not yet been enrolled in the IOCP for 12 months. PBGH did meet its targets for participants enrolled early enough to be followed for 12 months. At the time of PBGH's final report to CMMI, 42.5 percent of the participants had not yet been enrolled in the program for 12 months.

IV. CLINICIANS' PERCEPTIONS OF PROGRAM EFFECTS ON THEIR CARE

This section describes the available evidence on the extent to which PBGH's intervention had its intended effects on changing PCPs' behavior as a way to achieve desired impacts on participants' outcomes. As described in Section III.A.3, the program's theory of action required that PCPs (1) provide input on the participant's shared action plan and (2) intervene in a participant's care if deemed necessary through communication with a care manager. We use data from two rounds of the HCIA Primary Care Redesign Clinician Survey to describe changes in providers' behaviors and conclude whether the anticipated changes occurred. Both surveys rely on self-reported responses and reflect clinicians' perceptions of the program, rather than measuring quantitatively direct program effects on the care they provide.

A. Clinician survey

Survey methods. We administered the clinician survey in two rounds (fall 2014 and summer 2015). We sent the survey to PCPs (physicians and mid-level providers) working in the 23 PMGs in the IOCP. A total of 312 and 90 clinicians participating in PBGH's HCIA program responded to the survey during the first and second rounds, respectively (a response rate of 56 percent in Round 1 and 45 percent in Round 2) (Table IV.1). In the first round, we sent surveys to all IOCP clinicians (555 people). In the second round, for any awardee with more than 200 clinicians, such as PBGH, we sampled and the sent the survey to 200 clinicians who represented the total clinician population.

Table IV.1 Survey methods and timing

Survey	Who we surveyed	Field period (months relative to program implementation)	Number of surveys sent	Number of respondents	Response rate (excluding ineligible respondents)
Clinician, Round 1	Primary care providers (physicians and mid-level) in 23 PMGs	9/2014–11/2014 (17–19 months)	555	312	56%
Clinician, Round 2	Same	5/2015–8/2015 (24–27 months)	200	90	45%

Source: HCIA Primary Care Redesign Clinician Survey: Round 1 (field period September 2014 through November 2014) and Round 2 (field period May 2015 through July 2015).
Note: Response rate = Number of respondents divided by/ [number of surveys sent minus number ineligible].
PMG = participating medical group.

Survey results. Fewer than a third of respondents to the clinician survey in both rounds reported being somewhat or very familiar with the HCIA program (31 percent in Round 1 and 32 percent in Round 2). Unfortunately, we cannot distinguish which survey respondents are part of IPAs versus integrated health systems. We learned during our site visits that IPAs often have unique barriers to communicating with clinicians, because in an IPA clinicians operate independently and therefore might be less likely to be familiar with a new initiative such as the IOCP compared with clinicians employed by an integrated health system. Clinicians who are unaware of the IOCP are unlikely to refer patients to the program and unlikely to communicate with care managers to provide care management services as intended. We learned during our site visits that, in some cases, the lack of familiarity might have been caused by program leadership (from both IPAs and integrated health systems) choosing to integrate the IOCP into existing clinical workflows without identifying it as a new program among clinicians. This may have resulted from PMG leadership using different terminology for the program, not labeling it as IOCP when marketing it to providers. In addition, PBGH reported that many of the PMGs held risk-based Medicare Advantage contracts; therefore some physicians may have been accustomed to referring patients for care management without designating it as part of IOCP. Clinicians did not receive training as part of IOCP implementation, although interview respondents during our site visits noted that PMGs usually conducted formal or informal outreach to clinicians—through informational meetings, emails, or fliers—to educate clinicians about the IOCP.

As shown in Table IV.2, the small proportion of clinicians who were familiar with the HCIA program had mixed feelings about whether it had its intended effects. Specifically, during the second round of the survey, 55 to 59 percent of these respondents thought the HCIA program improved the quality, timeliness, safety, and patient-centeredness of care they provided to participants in the previous year. In contrast to the generally positive perceived effects on quality, timeliness, and safety, only 33 to 41 percent of these respondents thought the program improved the efficiency of care, equity of care, and information available for clinical decision making, with the remaining respondents reporting that the program had no effect on these dimensions of care or that it was too early to tell. Clinicians' perceptions of program effects were similar across the two survey rounds, although the change in sample size in the second round makes it difficult to draw conclusions about changes in clinicians' perceptions of the IOCP over time.

Table IV.2. PCPs' perceptions of the effects of the program on the care they provided to patients, from the clinician surveys (both rounds)

Dimension of care	Percentage (and number) of PCPs reporting that the HCIA had the following effect on the care they provided to patients enrolled in their practice in the past year			
	First round of survey (17 to 19 months after program implementation) N = 98		Second round of survey (24 to 27 months after program implementation) N = 29	
	Positive impact	No impact or too soon to tell	Positive impact	No impact or too soon to tell
Quality	53% (52)	43% (42)	59% (17)	NA[a]
Ability to respond in a timely way to patients' needs	54% (53)	41% (40)	55% (16)	41% (12)
Efficiency	42% (41)	51% (50)	41% (12)	48% (14)
Safety	45% (44)	49% (48)	55% (16)	42% (12)
Patient-centerdeness	54% (53)	40% (39)	55% (16)	NA[a]
Equity	33% (32)	62% (60)	NA[a]	52% (15)
Information available for clinical decision making	NA[b]	NA[b]	41% (12)	55% (16)

Source: HCIA Primary Care Redesign Clinician Survey: Round 1 (field period September 2014 through November 2014), Round 2 (field period May 2015 through July 2015).

Note: The percentages (and number) are limited to PCPs who reported that they were at least somewhat familiar with the HCIA program.

[a] Fewer than 11 respondents. Results suppressed to protect respondents' confidentiality.
[b] This question was asked only in the second round of the survey.

HCIA = Health Care Innovation Award; PCP = primary care provider.

NA = not available.

B. Conclusions about intermediate program effects on clinicians' behavior

For most of PBGH's program goals, the program design did not require clinicians to change the way they provide care. However, PCPs were expected to provide input into participants' shared action plans and to communicate with care managers about participants' worsening conditions. In addition, following adaptations made to improve program participation among eligible patients, PCPs were expected to provide a warm hand-off, introducing potential program participants to a care manager. The available evidence suggests that PBGH did not engage clinicians as planned. Only a small proportion of clinicians surveyed (roughly one-third) were aware of the program, and PCPs cannot, or are unlikely to, refer patients to the program if they do not know it exists. The findings from the clinician survey are consistent with PBGH's reports stating that engaging physicians was a key challenge throughout the three years of program implementation—especially engaging IPA physicians. However, over the course of our evaluation, we learned that PMG leadership used different terminology to describe the program, not always using the IOCP terminology, and that PMGs may already have had protocols in place

through which physicians referred patients for care management without designating it as IOCP, which may explain some of the disconnect between physician and administrator perceptions about the initiative. Moreover, PCPs who were aware of the program varied in whether they believed that the program improved key dimensions of care. For example, more than half believed the program improved the quality, timeliness of care, safety, and patient-centeredness, but fewer than half believed that the program improved the efficiency of care, equity of care, and information available for clinical decision making.

V. PROGRAM IMPACTS ON PATIENTS' OUTCOMES

In this section of the report, we present results for the quantitative analysis that aimed to draw conclusions about the impacts of PBGH's HCIA program on patients' outcomes in four domains: quality-of-care processes, quality-of-care outcomes, service use, and spending. We first describe the methods for estimating impacts (Section V.A.) and then the characteristics of the 2,996 treatment Medicare FFS beneficiaries at the start of the intervention (Section V.B). We next demonstrate that the treatment beneficiaries were similar at the start of the intervention to the beneficiaries we selected as a comparison group (Section V.C), and that there were similar patterns of attrition (Section V.D), both of which are important for limiting potential bias in impact estimates. Finally, in Section V.E, we describe the quantitative impact estimates, their plausibility given implementation findings, and why we were unable to draw conclusions in any of the study domains.

A. Methods

1. Overview

We estimated program impacts as the difference in outcomes for 2,996 treatment beneficiaries and outcomes for 6,665 matched comparison beneficiaries, adjusting for differences between these groups before PBGH's HCIA intervention began. We prespecified primary tests, describing the evidence we would need to conclude that the program was effective, and the awardee and CMMI reviewed these. Each test specified a population, outcome, time period, expected direction of effect, and threshold that we count as substantively important. The purpose of these primary tests was to focus the impact evaluation on hypotheses that would provide the most robust evidence about program effectiveness. We planned to use the results from the primary tests, in conjunction with the implementation evidence, to draw conclusions about program impacts in each of the four evaluation domains.

2. Treatment group definition

The treatment group is composed of Medicare FFS beneficiaries who, based on lists from PBGH, enrolled in the IOCP from May 1, 2013, to March 31, 2015. Although PBGH worked to improve data quality and completeness, we were still missing health insurance claim numbers and Social Security numbers for 22 percent of program participants, making it impossible for us to match all participants to their claims data and calculate outcomes. These beneficiaries are therefore not included in the treatment group. Using a cutoff of March 31, 2015, rather than the program end date of June 30, 2015, ensures that all beneficiaries could potentially be exposed to the intervention for at least one full quarter. We did not include beneficiaries exposed for less

time because we and the awardee did not expect the program to have immediate effects. We chose three months as the minimum potential exposure, rather than a longer period, for two reasons. First, program staff typically found that participants' conditions stabilized after the first few months of enrollment (Keith et al. 2015). Second, requiring a one-year exposure period would have significantly decreased the treatment group size (because, as shown in Figure III.1, about half of all participants enrolled during the last year of program operations) and would have excluded participants from several PMGs that did not begin enrolling participants until mid-2014.

We imposed several inclusion criteria for the treatment group used in the impact evaluation, limiting the Medicare population included in our analysis. First, we limited the analysis sample to those continuously enrolled in FFS Medicare and observable in Medicare data during the four quarters before their program enrollment (the baseline period). We did this to make it easier to match treatment beneficiaries to potential comparison beneficiaries. Continuous enrollment ensured that we had a complete record of beneficiaries' service use in the year before program enrollment. We further restricted the treatment group by excluding a small number of participants who received hospice care in the year before enrollment (less than 20 participants). Finally, 157 beneficiaries were dropped from the treatment group during the matching process, described in Section V.A.3.

Additional sample restrictions in each quarter. To be included in the analytic sample in any given quarter, each treatment group member had to meet two additional criteria. First, because we defined our evaluation outcomes quarterly (described in Section IV.A.4), and with the intervention quarters specified relative to each beneficiary's enrollment date (we set the day following the beneficiary's enrollment date as the first day of the intervention period), we required that a beneficiary's last full intervention quarter ended no later than June 30, 2015, the last day of the HCIA program. Second, the beneficiary's outcomes had to be observable in Medicare claims for at least one day during the quarter. Outcomes were observable for beneficiaries if they were alive, enrolled in Medicare FFS (Part A and B), and had Medicare as their primary payer (including beneficiaries dually eligible for Medicare and Medicaid).

The treatment group included participants from 18 of the 23 PMGs participating in the PBGH program. We did not include participants from 5 PMGs because all participants at these sites were enrolled in Medicare Advantage. One of the 5 PMGs that dropped out of the sample was the sole PMG located in Washington State.

3. Comparison group definition

The comparison group consists of 6,665 Medicare FFS beneficiaries we matched to the 2,996 treatment group beneficiaries on baseline characteristics (that is, characteristics observed during the four quarters before enrollment or pseudo-enrollment, defined below). This section describes how we constructed the matched comparison group; Section V.C shows the balance we achieved between the two groups on the matching variables.

Although PBGH first selected provider organizations and then beneficiaries within those PMGs for enrollment into the program, because we did not have a means to identify potential

comparison PMGs we were unable to identify and match at the PMG level. As a result, our unit of matching is the beneficiary. We included several characteristics in matching to account for group- and practice-level characteristics that might be correlated with the outcomes.

We constructed the comparison group in five steps.

First, we identified a pool of potential comparison members among Medicare FFS beneficiaries whose zip codes in the Medicare Enrollment Database indicated residence in a defined market area in and/or near the areas where the treatment group beneficiaries lived. Specifically, depending on the PMG, we included (1) counties where most treatment beneficiaries resided, (2) counties identified from lists provided by PBGH as part of the service area of a participating PMG, and/or (3) contiguous counties similar in size and population composition to geographic areas where the treatment group resided. In all cases, we balanced the need for a large pool of comparison beneficiaries to ensure a sufficient sample of well-matched comparison beneficiaries against the desire to restrict the pool to potential comparison beneficiaries located in areas similar to those of treatment beneficiaries, to ensure the face validity of our approach.

Second, we used Medicare claims and enrollment data to construct a person-month file for the potential comparison beneficiaries—meaning that the file contained one record for each month that a beneficiary was observable and living in a designated geographic area—for the period from May 2012 to March 2015. To approximate the date the potential comparison beneficiary would have enrolled in the intervention if he or she had been in the treatment group, for each person-month observation we defined a *pseudo-enrollment date* as the 15th of each month (as a result, each potential comparison beneficiary could have multiple pseudo-enrollment dates, one in each month). We then restricted this beneficiary-month pool to those beneficiary-months during which the beneficiary was enrolled in Medicare FFS continuously during the prior 12 months. This restriction mirrored that used in developing the treatment group and facilitated matching by ensuring that our claims-based matching variables were comparable between the two groups.

Third, we developed matching variables for all treatment group members (defined as the beneficiary's date of enrollment in the program) and all potential comparison beneficiary person-months (that is, for each of a potential comparison beneficiary's pseudo-enrollment dates that met the inclusion criteria defined previously). These matching variables included the following:

- Indicator variables to account for beneficiaries' demographic characteristics at the time of enrollment or pseudo-enrollment, including age, sex, race or ethnicity, and socioeconomic characteristics of the zip code in which the beneficiary resided
- Indicators of health care use and risk in the four-quarter baseline period, including Medicare and Medicaid dual eligibility; Hierarchical Condition Category (HCC) scores (which reflect projected spending in the coming year); chronic conditions (discussed in Section V.B); original reason for Medicare eligibility (either old age, disability, end-stage renal disease [ESRD], or both disability and ESRD); ED visits and hospitalizations; Medicare spending; and use of primary care, home health care, and skilled nursing care

- Three clinical quality-of-care process measures to assess the quality of health care received during the four-quarter baseline period: (1) whether a beneficiary had all hospital discharges in a quarter followed up with a primary care or ambulatory specialist visit within 14 days; (2) whether beneficiaries ages 18 to 75 with diabetes received four recommended preventive services in the year; and (3) whether beneficiaries 18 and older with ischemic vascular disease (IVD) received a complete lipid test in the year
- Characteristics of the physician group or practice that the beneficiary is attributed to, as this accounts for implicit selection criteria that stem from the fact that the awardee did not choose the PMGs at random; characteristics include whether the provider the beneficiary saw most often for primary care (the attributed provider) received payments from CMS for using EHRs in a meaningful way, and the size and type of the physician group practice in which the attributed provider worked

To identify physician group practices, we first attributed each treatment and potential comparison group member to the PCP who, based on Medicare FFS claims, provided the plurality of primary care services in the 12 months before enrollment or pseudo-enrollment. If the beneficiary did not have any primary care services in the past 12 months, we attributed him or her to the PCP who provided the plurality of care in the past 24 months. If there was a tie, we attributed to the PCP who provided the most recent service. For potential comparison group members, for whom we have multiple pseudo-enrollment dates, we attribute the potential comparison beneficiary to a National Provider Identifier for each of those pseudo-enrollment months. Next, we used the 2013 Medicare Data on Physician Practice and Specialty (MD-PPAS) to match attributed providers to their primary TIN. Following Welch et al. (2013), we then defined practice size as the number of physicians assigned to a TIN, and a physician group as single specialty if at least 90 percent of its physicians were in only one of the six main physician specialties in the MD-PPAS (all other groups were considered to be multispecialty).

Fourth, we narrowed the pool of potential comparison beneficiaries by excluding any potential members who had received an evaluation and management service from a treatment PCP or from the PMG or practice associated with the treatment PCP (identified by TIN) in the 24 months before the beneficiary's pseudo-enrollment month. (We defined our universe of treatment providers as those PCPs who had at least one treatment beneficiary attributed to them. Likewise, we defined the universe of organizations participating in the program as the set of TINs to which providers attributed at least one treatment beneficiary bill [as identified in the MD-PPAS].) Doing so limited the possibility we matched to beneficiaries receiving care from the same practice as treatment group beneficiaries, ensuring against (1) intervention spillover; (2) matching to beneficiaries invited to participate in the intervention but who declined; or (3) accidently matching to beneficiaries who received treatment but were not included in our final treatment sample due to missing patient identifiers, as described in Section III.A.2. As with the treatment group, we further restricted the pool of potential comparison group members by excluding those beneficiary months' observations if the beneficiary had received hospice care in the year before pseudo-enrollment dates.

Finally, we executed matching. Given the large size of the potential comparison pool (more than 1,000,000 unique potential comparison beneficiaries), we used a two-stage matching approach:

- In the **first stage,** we used the nearest-neighbor matching approach to narrow the comparison pool. We began by organizing treatment and potential comparison beneficiaries into exact-match strata, meaning that beneficiaries in each stratum had exactly the same values on two key characteristics: (1) market area and (2) enrollment or pseudo-enrollment month (23 monthly cohorts over the program period May 2013 to March 2015).

  Next, we used a combination of Mahalanobis distances—a distance matrix based on the ranks of covariate values—between the treatment beneficiaries and all the potential comparison beneficiaries and calipers (that is, forcing each matched comparison beneficiary to have a value within a specified range of the treatment beneficiary's value) applied to HCC score, ED visits, inpatient admissions, and Medicare spending to select up to 20 potential comparison beneficiaries for each treatment beneficiary. We set caliper values based on the distribution of a given characteristic among treatment beneficiaries, requiring tighter matches (a smaller caliper value) among treatment beneficiaries who had many potential comparison matches, but expanding the calipers (allowing less-close matches) for beneficiaries with outlier values on the characteristic. We allowed potential comparison beneficiaries to be matched to no more than one treatment beneficiary, meaning that a potential comparison beneficiary selected as the nearest match for one treatment beneficiary could not be selected as the nearest match for another treatment beneficiary—whether enrolled in the same or a different month. We used an iterative approach to the first-stage matching, ordering strata according to the mean difference in spending between treatment and potential comparison beneficiaries, starting with the stratum with the largest difference in mean spending (that is, the hardest-to-match stratum). During this stage, 157 (of 3,153) treatment beneficiaries were dropped because they could not be matched to any potential comparison beneficiaries due to the caliper-matching restriction.

- In the **second stage,** we pooled matches from the first stage across all strata and used a combination of exact-matching and propensity scores to select the final comparison group. As described previously, exact-matching means that we forced a treatment beneficiary to have an identical value for a given variable to his or her matched comparisons. We used the same exact-match characteristics used in the first stage as well as Medicare–Medicaid dual eligibility. For all other variables, we matched using propensity scores. A propensity score is the predicted probability, based on all of a beneficiary's matching variables, that a given beneficiary was selected for treatment (Stuart 2010). In other words, it collapses all of the matching variables into a single number for each beneficiary that can be used to assess how similar beneficiaries are to one another. By matching each treatment beneficiary to one or more comparison beneficiaries with similar propensity scores, we generated a comparison group that was similar, on average, to the treatment group. Each treatment beneficiary was matched to up to three beneficiaries from the potential comparison pool to increase the precision in the impact estimates (relative to 1:1 matching).

  Because of concerns about selection into the comparison group and the potential for differential attrition, we required equivalence between the treatment and comparison groups

on the distribution (for example, 25th, 50th, and 75th percentiles) of several key baseline characteristics, such as a patient's HCC score and prior year Medicare spending (results not shown). We also excluded several hard-to–match participants from the treatment groups during the matching process.

Additional sample restrictions in each quarter. To be included in the analytic sample, a matched comparison beneficiary had to meet the same additional criteria as the treatment group members—that is, the end of the last intervention quarter had to be no later than June 30, 2015, and the beneficiary had to be observable in Medicare claims for at least one day of the quarter.

4. Construction of outcomes and covariates

We used Medicare claims from January 1, 2009, to June 30, 2015, for beneficiaries assigned to the treatment and comparison practices to develop two types of variables: (1) outcomes, defined for each beneficiary in each baseline or intervention quarter (that is, in each three-month period, either before [baseline] or after [intervention] enrollment or pseudo-enrollment); and (2) covariates, which describe key characteristics of the beneficiary and the beneficiary's PCP during four baseline quarters (before enrollment or pseudo-enrollment) for use as control variables in the regression models. Control variables were measured during the baseline period to avoid the potential bias that could occur if the intervention affected both control variables and outcomes. For example, the intervention could result in greater contact with the health system and earlier diagnoses of diseases and conditions, which could affect both health-related characteristics and outcomes. If we adjusted for changes in health-related status during the intervention period—which we define for each beneficiary as the four quarters after enrollment or pseudo-enrollment—we would adjust away part of the impact of the intervention. Appendix 1 provides details on the methods we used to construct these variables.

Outcomes. For each beneficiary, we calculated eight outcomes that we grouped into four domains:

1. Domain: Quality-of-care processes
    a. Diabetes quality-of-care composite (binary variable for each beneficiary); calculated as whether a beneficiary with diabetes had had all four recommended tests—lipid profile, hemoglobin A1c test, dilated eye exam, and nephropathy screening—during the previous 12 months
    b. IVD lipid profile (binary variable for each beneficiary); calculated as whether a beneficiary with IVD had a complete lipid profile during the previous 12 months
    c. Ambulatory-care follow-up visit within 14 days of a hospital discharge (binary variable for each beneficiary); calculated as whether all of an individual's discharges in a quarter were followed by an ambulatory visit with a primary care or specialist physician within 14 days of the discharge
2. Domain: Quality-of-care outcomes
    a. Inpatient admissions for ambulatory care-sensitive conditions (ACSCs) (number/quarter)

   b. 30-day unplanned hospital readmission rate (number/quarter)

3. Domain: Service use
   a. All-cause inpatient admissions (number/quarter)
   b. Outpatient ED visit rate (number/quarter); outpatient ED visits are defined as ED visits or observational stays that do not end in a hospital admission
4. Domain: Spending
   a. Total Medicare Part A and B spending (dollars/month)

Four of these outcomes—30-day unplanned readmissions, all-cause inpatient admissions, outpatient ED visits, and total Medicare Part A and B spending—are outcomes that CMMI has specified as core for the evaluations of all HCIA programs. Our definition of the readmission measure, however, differs from CMMI's standard definition. CMMI typically defines readmissions as the proportion of inpatient admissions that end in an unplanned readmission. Instead, we analyzed impacts on the *number* of these unplanned readmissions across all beneficiaries per quarter, because this enables us to look at the total impact on readmissions across the treatment group, rather than readmissions contingent on an inpatient admission. We made this decision, in consultation with CMMI, because the intervention might also affect the number of and type of admissions.

All outcomes are quarter-specific—meaning that we calculated them for each baseline and intervention quarter separately—except for the two quality-of-care process measures for IVD and diabetes. Because these two measures assess whether a beneficiary received recommended preventive care services over a year-long period, we calculated these measures over full years rather than quarters: for example, over the baseline year (that is, the period corresponding to the four baseline quarters) and over the full year of the intervention period (corresponding to the first four intervention quarters). We avoided calculating these measures for overlapping periods, meaning that no measurement year included services provided in another measurement year. As a result, these two measures are defined only for beneficiaries that enrolled by June 30, 2014 (those that could be exposed to the intervention for at least one full year).

Finally, we defined all outcomes for all treatment and comparison group members, except for the three measures of quality-of-care processes. We calculated the measure of 14-day follow-up after discharge among only those beneficiaries with at least one hospital discharge in the relevant quarter. We calculated the diabetes composite measure among beneficiaries ages 18 to 75 with diabetes at the beginning of the period (baseline or intervention period), and calculated the measure of lipid screening among beneficiaries ages 18 or older with IVD at the beginning of the period.

Covariates. The covariates, defined at the enrollment (treatment group) or pseudo-enrollment date (comparison group) include (1) measures of chronic conditions created by applying Chronic Condition Warehouse algorithms to claims in the 12 to 36 months (depending on the condition) before the beneficiary's enrollment or pseudo-enrollment date, including the number of major chronic conditions (among 25 mostly physical health conditions) and 5 specific

chronic conditions (Alzheimer's disease, chronic kidney disease, chronic obstructive pulmonary disease, congestive heart failure, and diabetes); (2) the number of mental health conditions (out of 6); (3) HCC score; (4) outpatient ED visits, inpatient admissions, readmissions, ACSC admissions, Part A and B Medicare spending, and inpatient spending in the baseline year; (5) other service use (primary care spending and number of visits, and use of skilled nursing facilities or home health care) in the baseline year; (6) the three quality-of-care process measures for diabetes care, IVD care, and ambulatory-care follow-up after a hospital discharge; (7) an indicator for dual Medicare and Medicaid enrollment; (8) beneficiary-level demographics (age, gender, and race and ethnicity); (9) characteristics of the PCP to whom each beneficiary was attributed (size of medical group, whether the physician belonged to a single or multispecialty group, and if the physician demonstrated Meaningful Use of EHRs); and (10) 2012 ZIP code-level characteristics (poverty rate, unemployment rate, percentage of adults with a college degree, and whether the ZIP code is urban or rural) of the ZIP code where the beneficiary lived.

5. Regression model

We used a regression model to implement a contemporaneous differences analysis. We used this model, rather than a difference-in-differences model, because we could not construct a beneficiary population similar to the treatment group before the intervention start date. For each outcome, the model estimates the relationship between the outcome and a series of predictor variables, assuming that each of the predictor variables has a linear (additive) relationship with the outcome. The predictor variables include the patient-level covariates (defined in Section V.A.4); whether the patient is assigned to the treatment or the comparison group; indicators for each post-intervention quarter (or, for the diabetes and IVD measures, for the final post-intervention quarter of the year-long measurement period); and an interaction of a beneficiary's treatment status with each post-intervention quarter (or, for the diabetes and IVD measures, the final post-intervention quarter of the year-long measurement period).

The estimated relationship between the interaction term and the outcome in a given quarter is the impact estimate for that quarter (or, for the diabetes and IVD measures, for the year ending with that quarter), while controlling for any differences in outcomes associated with differences in the covariates. By providing separate impact estimates for each intervention quarter (or year, for the diabetes and IVD measures), the model enables the program's impacts to change the longer the beneficiaries are enrolled in the program. We can also test impacts over discrete sets of quarters or years, which is needed to implement the primary tests discussed in the next section. Finally, the model quantifies the uncertainty in the impact estimates, allowing for statistical tests that determine whether observed differences in outcomes between the treatment and comparison groups are likely due to chance. Appendix 2 provides details on the regression methods, including descriptions of the weights each beneficiary receives in the model.

6. Primary tests

Table V.1 shows the primary tests for PBGH, by domain. Each test specifies a population, outcome, time period, expected direction of effect, and threshold that we count as substantively important. The purpose of these primary tests is to focus the impact evaluation on hypotheses that will provide the most robust evidence about program effectiveness (see Appendix 3 for

detail and a description of how we selected each test). We provided both the awardee and CMMI an opportunity to comment on the primary tests.

Our rationale for selecting these primary tests is as follows:

- **Outcomes.** PBGH's central goal was to reduce hospitalizations, ED visits, and Medicare Part A and B spending, so our primary tests address these three outcomes. In addition, the primary tests address two quality-of-care outcomes the intervention is expected to affect: hospitalizations for ACSCs and 30-day unplanned hospital readmissions. Finally, we include three quality-of-care process measures that, based on PBGH's theory of action, we think the program could improve: (1) a composite measure for whether a beneficiary with diabetes received all of four recommended processes of care during the year (HbA1c test, lipid profile, dilated eye exam, and nephropathy screening); (2) receipt of a complete lipid profile for people with IVD; and (3) receipt of a follow-up ambulatory care visit with a primary care or specialist provider within 14 days of hospital discharge. Although PBGH did not set explicit targets for these particular quality-of-care process measures, the one-on-one interactions between care managers and participants could be expected to improve rates of recommended care.
- **Time period.** We expect reductions in outcomes across all domains to be largest during program participation and perhaps harder to identify as the health of the treatment and comparison group beneficiaries evolved beyond the participation period. The intervention was expected to last at least 12 months. We chose to specify our primary tests based on outcomes in the four quarters following a participant's enrollment date (that is, intervention quarters 1 to 4 [I1 to I4]). To implement each primary test, we take the average of the regression-adjusted estimates across the four intervention quarters for that outcome for the quarterly outcomes, or the regression-adjusted estimate for the intervention year for the diabetes and IVD measures (which are measured over a full year).
- **Population.** PBGH expected to have impacts for the population of beneficiaries enrolled in the IOCP. Therefore, the primary tests include all (observable) Medicare FFS beneficiaries who enrolled in the PBGH program during the period from May 2013 through March 2015.
- **Direction (sign) of the impact estimate.** For the quality-of-care process measures, we expect the impact estimate to be positive, signaling an increase in the percentage of people receiving recommended care. For all other outcomes, we expect the impact estimates to be negative, indicating a reduction in service use or overall expenditures.
- **Substantive thresholds.** Some impact estimates could be large enough to be policy relevant (to CMMI and other stakeholders) even if they are not statistically significant; for this reason, we have prespecified thresholds for what we call substantive importance. We express the threshold as a percentage change from the counterfactual—that is, the outcomes that beneficiaries in the treatment group would have had if they had not received the HCIA-funded intervention. The threshold of 3.75 percent that we chose for substantive importance for outpatient ED visits, all-cause admissions, and total Medicare spending is 75 percent of PBGH's anticipated 5 percent impact on these outcomes. (We used 75 percent, recognizing that PBGH could still be considered successful if it approached, but did not achieve, its fully

anticipated effects.) The thresholds of 3.0 percent used for the quality-of-care process measures are based on PBGH's anticipated 4.0 percent improvement in health care quality.

The 15 percent threshold for the quality-of-care outcomes—unplanned readmissions and hospitalizations for ACSCs—is extrapolated from the literature (Peikes et al. 2011; Rosenthal et al. 2016) because PBGH did not specify by how much it expected to improve these outcomes.

Table V.1. Specification of the primary tests for PBGH

Domain (number of tests in the domain)[a]	Outcome (units)	Time period for impacts (controlling for baseline differences)[b, c]	Population	Substantive threshold (expected direction of effect)[d, e]
Quality-of-care processes (3)	Received all four recommended diabetes processes of care in the year (binary [yes or no]/beneficiary/year)	Full intervention year (corresponding to intervention quarters 1 through 4)	Medicare FFS beneficiaries ages 18 to 75 with diabetes	3.0% (+)
	Received lipid profile in the year (binary [yes or no]/beneficiary/year)	Full intervention year (corresponding to intervention quarters 1 through 4)	Medicare FFS beneficiaries ages 18 or older with IVD	3.0% (+)
	All inpatient admissions within a quarter were followed by an ambulatory care visit with a primary care or specialist provider within 14 days (binary [yes or no]/beneficiary/year)	Average over intervention quarters 1 through 4	Medicare FFS beneficiaries with at least one hospital stay in the quarter	3.0% (+)
Quality-of-care outcomes (2)	Inpatient admissions for ambulatory care-sensitive conditions (#/beneficiary/quarter)	Average over intervention quarters 1 through 4	Medicare FFS beneficiaries	15.0% (-)
	30-day unplanned hospital readmission rate (#/beneficiary/quarter)	Average over intervention quarters 1 through 4	Medicare FFS beneficiaries	15.0% (-)
Service use (2)	All-cause inpatient admissions (#/beneficiary/quarter)	Average over intervention quarters 1 through 4	Medicare FFS beneficiaries	3.75% (-)
	Outpatient ED visit rate (#/beneficiary/quarter)	Average over intervention quarters 1 through 4	Medicare FFS beneficiaries	3.75% (-)
Spending (1)	Medicare Part A and B and Medicaid spending ($/beneficiary/month)	Average over intervention quarters 1 through 4	Medicare FFS beneficiaries	3.75% (-)

Note: The intervention quarters were measured relative to the date that an individual enrolled in the IOCP (or pseudo-enrolled, if a comparison beneficiary). For example, the first intervention quarter for a beneficiary who enrolled on January 10, 2014, is from January 10, 2014, to April 9, 2014.

[a] We adjusted the *p*-values from the primary test results for the multiple comparisons made within each domain, but not across domains.

[b] The regression models controlled for differences between the treatment and comparison groups during the baseline year when estimating program impacts.

[c] For all but the diabetes and IVD quality-of-care process measures, we took the average of the regression-adjusted estimates for intervention quarters 1 through 4. For the diabetes and IVD measures, which are defined annually, we took the impact estimates for the 12-month period following enrollment (or pseudo-enrollment).

[d] The substantive threshold is the impact as a percentage of the counterfactual. The counterfactual is the outcome the treatment group would have had in the absence of the HCIA-funded intervention.

[e] For the quality-of-care process measures, the populations were further restricted by the measures' respective condition. Thus, primary care follow-up visits were measured among those beneficiaries with an index hospital stay, composite diabetes quality of care was measured among those with diabetes, and lipid testing was measured among those with ischemic vascular disease.

ED = emergency department; FFS = fee-for-service; IOCP = Intensive Outpatient Care Program; IVD = ischemic vascular disease; PBGH = Pacific Business Group on Health.

7. Secondary tests (robustness checks)

We also conducted secondary quantitative tests to help corroborate the findings from the primary tests. This is important because some of the differences observed between the treatment and comparison groups in the primary test results could reflect limitations of the non-experimental impact evaluation design or random fluctuations in the data. We have greater confidence in the primary results if they are generally consistent with the expected broader pattern of results from the secondary tests.

We conducted two sets of secondary tests for PBGH:

1. After matching, we assessed sample attrition between the treatment and matched comparison group during the intervention period. Demonstrating that the treatment and comparison groups had similar attrition patterns over the study period is important for the impact estimation. Differential attrition between treatment and comparison groups could lead us to confound potential program effects with changes in the composition of the treatment and comparison groups.
2. As a sensitivity check, we examined the patterns of service use, quality-of-care outcomes, and spending for treatment group beneficiaries (and their matched comparison beneficiaries) who enrolled by June 30, 2014, to allow at least 12 months of potential exposure to the intervention. As described in Section III, we included program enrollees through March 2015, three months before the end of the IOCP. One potential concern of this cutoff is that impact estimates might have been attenuated because we included people who would have been exposed to the intervention for fewer than 12 months. That is, by including enrollees through March 2015, our main estimates included people who were potentially exposed to as few as 3 months of the intervention.

8. Synthesizing evidence to draw conclusions

Within each domain, we planned to draw one of five conclusions about program effectiveness based on the primary test results, the results of secondary tests, and the plausibility of those findings given the implementation evidence:

1. Statistically significant favorable effect (the highest level of evidence)
2. Substantively important (but not statistically significant) favorable effect
3. Substantively important (but not statistically significant) unfavorable effect
4. No substantively large effect
5. Indeterminate effect

We could not conclude that a program has a statistically significant unfavorable effect because, in consultation with CMMI, we decided to use one-sided statistical tests (which do not test for evidence of unfavorable effects). We used one-sided tests to increase the probability that, if a program truly did have impacts, we would be able to detect them.

Appendix 3 describes our decision rules for each of the five possible conclusions. In short, we concluded that a program had a statistically significant favorable effect in a domain if (1) at least one primary test result in the domain was favorable and statistically significant, after adjusting the statistical tests to account for multiple tests (if applicable) within a domain; or (2) the average impact estimate across all primary tests in the domain was favorable and statistically significant. In both cases, we also had to determine that the primary test results were plausible given the results of the secondary test and implementation evidence. We concluded that a program had a substantively important favorable effect if the average impact estimate in the domain was substantively important but not statistically significant, and if the result was plausible given the secondary tests and implementation evidence. In contrast, if the average impact estimate was unfavorable (opposite the hypothesized direction), larger than the substantive threshold, and unfavorable effects were plausible given the other evidence, we concluded the program had a substantively important unfavorable effect. If the tests in a domain did not meet any of these criteria, we instead used the following rules. First, if the tests for at least one outcome in the domain (or all outcomes in the domain together) had sufficient statistical power to detect an impact of the size of the substantive threshold with at least 75 percent probability, we concluded there was not a substantively large effect because we are reasonably confident that we would have detected such an effect had there been one. Second, if the power was not sufficient (less than 75 percent) to detect this type of impact, we concluded the impact in the domain was indeterminate. Indeterminate means either that the program truly did not have effects that were substantively large, or that it did, but our statistical tests were unable to detect them. Finally, if the results for the primary tests in a domain were not plausible given the implementation evidence, we did not draw any conclusions about program impacts in that domain.

B. Characteristics of the treatment group at baseline

This section describes the characteristics of the 2,996 treatment group beneficiaries at their enrollment dates into the intervention. We also show this information in the second column of Table V.2. (Table V.2 serves a second purpose—to show the equivalence of the treatment and comparison beneficiaries at the start of the intervention—which we describe in Section V.C). For benchmarking purposes, the last column shows the values of relevant variables for the national Medicare FFS population, when available.

Table V.2 indicates the treatment group had greater health care needs than the general Medicare population, consistent with PBGH's target population of chronically ill beneficiaries. The HCC risk score for the treatment group is 2.38, indicating that the group could be expected to have Medicare spending that is 2.38 times higher than the national average (1.00) over the next year. The treatment group members had an average of 6.6 multiple chronic conditions. A high percentage had chronic kidney disease (48.2 percent), diabetes (47.5 percent), congestive heart failure (35.0 percent), and chronic obstructive pulmonary disease (27.8 percent). These condition-specific rates are each about two to three times the national averages.

Treatment group members also had high service use (inpatient admissions and outpatient ED visits) and spending in the 12 months before program enrollment, relative to national Medicare

FFS averages. For example, the treatment group beneficiaries had on average 418 inpatient admissions per 1,000 beneficiaries in the quarter directly before their enrollment dates (of which 83 were ACSC-related) and 162 admissions per 1,000 beneficiaries per quarter in the period 4 to 12 months before their enrollment dates (including 30 ACSC-related admissions), compared with a national average of 74 admissions per 1,000 beneficiaries per quarter. In addition, the ED visit rate among the treatment group was nearly four times the national average (380 visits per beneficiary per quarter versus a benchmark of 105) in the quarter before enrollment, and almost 2.5 times as many (235 visits per beneficiary per quarter) in the previous three quarters. These hospitalizations and ED visits, and perhaps other service use, drove up treatment group beneficiaries' Medicare spending as well, to $27,281 per person in the year before enrollment, nearly three times the national average.

Table V.2. Characteristics at baseline of treatment and comparison beneficiaries for PBGH

Characteristic	Treatment group (n = 2,996)	Unmatched comparison pool (after first stage restrictions) (n = 52,401)	Matched comparison group (n = 6,665)	Absolute difference[a]	Standard-ized difference[b]	Medicare FFS average
Exact-match variables[c]						
Market area (percentages)						
Phoenix, Arizona	30.1	30.4	30.1	0.0	0.000	n.a.
Las Vegas, Nevada	3.0	3.1	3.0	0.0	0.000	n.a.
Boise, Idaho	10.4	7.9	10.4	0.0	0.000	n.a.
San Francisco, California	22.5	24.2	22.5	0.0	0.000	n.a.
Sacramento, California	7.4	7.9	7.4	0.0	0.000	n.a.
Los Angeles, California	3.1	3.3	3.1	0.0	0.000	n.a.
San Diego, California	23.5	23.0	23.5	0.0	0.000	n.a.
Medicaid dual eligibility status at enrollment (%)	19.5	17.7	19.5	0.0	0.000	22[d]
Propensity-matched variables[e]						
Demographic characteristics						
Age (years)	76.2	74.6	76.1	0.1	0.013	71[f]
Male (%)	37.2	44.3	36.4	0.8	0.016	45.3[g]
Race						
Black (%)	6.4	4.1	6.5	-0.2	-0.007	10.4[g]
Hispanic (%)	3.2	3.1	3.6	-0.4	-0.021	2.7[g]
Other nonwhite (%)	6.4	7.6	5.5	1.0	0.042	5.5[g]
Medicare-related characteristics						
Original reason for entitlement (%)						
Disability only	20.5	19.1	21.0	-0.5	-0.013	16.7[g]
ESRD only	0.3	0.2	0.2	0.0	0.001	0.1[g]
Disability and ESRD	0.8	0.5	0.9	-0.1	-0.009	NA

Table V.2 *(continued)*

Characteristic	Treatment group (n = 2,996)	Unmatched comparison pool (after first stage restrictions) (n = 52,401)	Matched comparison group (n = 6,665)	Absolute difference[a]	Standard-ized difference[b]	Medicare FFS average
Health status and chronic conditions						
HCC risk score (at enrollment)	2.38	1.89	2.37	0.0	0.009	1.0
Chronic conditions (# out of 25)[h]	6.6	5.3	6.7	-0.1	-0.030	NA
Mental health conditions (# out of 6)[i]	0.7	0.5	0.7	-0.0	-0.026	NA
Chronic conditions (%)						
Alzheimer's	5.5	4.9	6.4	-0.9	-0.039	4.9[j]
CHF	35.0	24.6	36.2	-1.2	-0.026	15.3[j]
CKD	48.2	31.5	48.3	-0.1	-0.002	16.2[j]
COPD	27.8	19.0	28.7	-0.9	-0.020	11.8[j]
Diabetes	47.5	34.7	48.8	-1.3	-0.025	28.0[j]
Service use and spending 3 months before enrollment						
All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	417.6	215.5	374.6	43.0	0.072	74[k]
Ambulatory care-sensitive condition-related inpatient admissions (#/1,000 beneficiaries/quarter)	82.8	27.0	65.1	17.7	0.069	NA
Outpatient ED visits (#/1,000 beneficiaries/quarter)	379.5	244.4	364.1	15.4	0.022	105[l]
Medicare Part A and B spending ($/beneficiary/month)	3,557	2,031	3,337	220	0.042	860[m]
Service use and spending 4 to 12 months before enrollment						
All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	161.7	133.1	150.9	10.7	0.040	74[k]
Ambulatory care-sensitive condition-related inpatient admissions (#/1,000 beneficiaries/quarter)	30.3	16.5	27.7	2.6	0.025	NA
Outpatient ED visits (#/1,000 beneficiaries/quarter)	234.6	181.4	229.1	5.6	0.015	105[l]
Medicare Part A and B spending ($/beneficiary/month)	1,846	1,554	1,781	65	0.024	860[m]
Service use and spending in the 6 or 12 months before enrollment						
More than 3 ED or hospital admissions in 6 months before enrollment (%)	16.2	6.6	14.2	1.9	0.057	NA
All-cause inpatient admissions (#/1,000 beneficiaries/year)	902.5	614.9	827.4	75.2	0.072	296[k]
Ambulatory care-sensitive condition-related inpatient admissions (#/1,000 beneficiaries/year)	173.6	76.4	148.1	25.5	0.060	NA
Outpatient ED visits (#/1,000 beneficiaries/year)	1,083.6	788.8	1,051.5	32.1	0.022	420[l]
Medicare Part A and B spending ($/beneficiary/year)	27,281	20,081	26,037	1,244	0.040	10,320[m]
30-day unplanned hospital readmissions (#/1,000/year)	100.8	41.1	82.8	18.0	0.053	NA

Table V.2 *(continued)*

Characteristic	Treatment group (n = 2,996)	Unmatched comparison pool (after first stage restrictions) (n = 52,401)	Matched comparison group (n = 6,665)	Absolute difference[a]	Standard-ized difference[b]	Medicare FFS average
Other service use						
Primary care spending (Medicare) in 12 months before enrollment ($/beneficiary/year)	1,205	788	1,145	61	0.040	NA
Number of primary care visits						
Up to 3 months before enrollment (#)	1.7	1.0	1.6	0.2	0.092	NA
4 to 12 months before enrollment (#)	3.9	2.7	3.6	0.3	0.096	NA
Any primary care visit in month before enrollment (%)	56.7	31.2	53.0	3.7	0.074	NA
Any use of SNF care in 12 months before enrollment (%)	16.3	8.5	16.1	0.3	0.007	NA
Any use of home health care in 12 months before enrollment (%)	35.5	21.8	34.7	0.8	0.018	NA
Clinical quality-of-care process measures						
Met denominator criteria for diabetes care measure (%)	44.7	31.5	46.1	-1.4	-0.028	NA
Received all four recommended diabetes tests in 12 months before enrollment[n]	45.9	39.7	44.2	1.7	0.033	NA
Met denominator criteria for ischemic vascular disease care measure (%)	48.8	39.0	50.0	-1.2	-0.023	NA
Complete lipid profile conducted in 12 months before enrollment[o] (%)	65.3	72.0	66.9	-1.6	-0.035	NA
Ambulatory-care follow-up occurred after hospital stays[p]						NA
No index stay (%)	61.2	63.5	61.6	-0.4	-0.009	NA
Fewer than one-third of stays followed up (%)	4.9	8.8	5.6	-0.6	-0.027	NA
One-third to two-thirds of stays followed up (%)	4.0	2.1	4.1	-0.1	-0.005	NA
More than two-thirds of stays followed up (%)	29.9	25.6	28.8	1.1	0.025	NA
Characteristics of the PCP to whom the beneficiary is attributed						
Beneficiary was attributed to PCP (%)	94.6	75.8	92.3	0.0	0.082	NA
Size of medical group that the PCP is in[q]						
1 provider (%)	16.3	43.4	19.2	-2.9	-0.070	NA
2–9 providers (%)	27.7	29.8	33.3	-5.5	-0.117	NA
10–49 providers (%)	7.9	7.1	10.5	-2.6	-0.089	NA
50–99 providers (%)	2.5	1.9	3.0	-0.5	-0.032	NA
100 or more providers (%)	45.1	17.5	33.7	11.4	0.246	NA
Unknown (%)	0.5	0.3	0.4	0.0	0.008	NA
PCP is in multispecialty medical group[q] (%)	52.2	26.9	42.2	9.9	0.204	NA
Meaningful use of EHR by the PCP[q,r] (%)	59.3	43.3	55.3	4.0	0.081	NA

Table V.2 *(continued)*

Characteristic	Treatment group (n = 2,996)	Unmatched comparison pool (after first stage restrictions) (n = 52,401)	Matched comparison group (n = 6,665)	Absolute difference[a]	Standard-ized difference[b]	Medicare FFS average
Characteristics of zip code (2012)						
Poverty rate (%)	12.4	12.5	12.8	-0.4	-0.056	NA
Unemployment rate (%)	9.1	9.4	8.8	0.3	0.108	NA
Adults with college degree (%)	35.4	33.7	34.7	0.8	0.045	NA
Urban (%)	97.7	96.5	96.8	0.9	0.053	NA

Sources: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at CMS. Zip code poverty rate merged from the 2012 Five-Year American Community Survey Zip Code Characteristics.

Notes: Characteristics were measured at the time of a beneficiary's enrollment (for the treatment group) or pseudo-enrollment (for the potential and matched comparison groups). The matched comparison group means were weighted based on the number of matched comparisons per treatment beneficiary. For example, if four comparison beneficiaries were matched to one treatment beneficiary, the four comparison beneficiaries each had a matching weight of 0.25.

The unmatched comparison group shown was the group that came out of the first stage of matching, which used nearest-neighbor matching to narrow the pool and make it much more similar to the treatment group than the initial pool of potential comparisons.

The chronic condition flags were calculated using one to three years of claims before the enrollment or pseudo-enrollment date (depending on the condition), using the Chronic Condition Warehouse definitions.

[a] The absolute difference is the difference in means between the treatment and matched comparison groups. Absolute differences might not be exact due to rounding.

[b] The standardized difference is the difference in means between the treatment and comparison groups divided by the standard deviation of the variable, which is pooled across the treatment and comparison groups.

[c] Variables for which we required treatment and comparison beneficiaries to match on exactly.

[d] Health Indicators Warehouse (2014c).

[e] Variables that we matched on through a propensity score, which captures the relationship between a beneficiary's characteristics and his or her likelihood of being in the treatment group.

[f] Health Indicators Warehouse (2014a).

[g] Chronic Conditions Data Warehouse (2014a, Table A.1).

[h] We use 25 of the 27 chronic condition categories defined by the Chronic Conditions Warehouse (see https://www.ccwdata.org/web/guest/condition-categories). We exclude the Alzheimer's disease and the acute myocardial infarction flags because other flags include these conditions.

[i] The six mental health conditions are conduct disorders and hyperkinetic syndrome, anxiety disorder, bipolar disorder, personality disorders, schizophrenia and other psychotic disorders, and depressive disorders, as defined by the Chronic Conditions Warehouse

[j] Chronic Conditions Warehouse (2014b, Table B.2).

[k] Health Indicators Warehouse (2014b).

[l] Gerhardt et al. (2014).

[m] Boards of Trustees (2013).

[n] Measured among those with diabetes.

[o] Measured among those with ischemic vascular disease.

[p] Measured among those with an index hospital stay.

[q] Measured among those with an attributed PCP.

[r] This measure examines EHR meaningful use payments to an individual PCP, rather than the organization to which the PCP belongs.

CHF = congestive heart failure; CKD = chronic kidney disease; CMS = Centers for Medicare & Medicaid Services; COPD = chronic obstructive pulmonary disease; ED = emergency department; EHR = electronic health record; ESRD = end-stage renal disease; FFS = fee-for-service; HCC = Hierarchical Condition Category; PBGH = Pacific Business Group on Health; PCP = primary care provider; SNF = skilled nursing facility.

NA = not available.

n.a. = not applicable.

Among other characteristics, more than half (56.7 percent) of treatment group members had a visit to their PCP in the month before enrollment, and about one in six treatment group members (16.3 percent) had some use of skilled nursing facility care in the 12 months before enrollment. This is consistent with the population PBGH targeted, as PCP referrals and recent use of skilled nursing facility care were among the ways that PMGs identified potential program participants. Almost half (45.1 percent) of treatment group members were attributed to PCPs who were a part of medical groups of 100 or more physicians, with a similar proportion attributed to providers working in groups smaller than 10 physicians. This bimodal distribution of the treatment group across group size is consistent with the selection of PMGs that embodied two different organizational structures: integrated health systems or IPAs, in which practices in the latter can have their own unique TINs. About half (52.2 percent) of treatment group members were attributed to PCPs who belonged to multispecialty groups, and more than half (59.3 percent) were attributed to providers who had received payment from CMS for using EHRs in a meaningful way.

C. Equivalence of the treatment and comparison groups at the start of the intervention

Demonstrating that the treatment and comparison groups were similar at the start of the intervention is critical for the evaluation design. This similarity increases the credibility of a key assumption underlying contemporaneous differences models—that the comparison group outcomes are the same as the outcomes the treatment group would have had in the absence of the intervention.

Table V.2 shows that the 2,996 treatment beneficiaries and the 6,665 selected comparison beneficiaries were similar at the start of the intervention on variables used in matching. By construction, there were no differences between the two groups on the market area in which beneficiaries received care, or on Medicare–Medicaid dual eligibility status at enrollment. There were some differences between the treatment and matched comparison beneficiaries on the variables we matched through propensity scores, but the standardized differences across the propensity-score matching variables were within our target of 0.25 standardized differences, and nearly all were actually within 0.15 standardized differences (the 0.25 target is an industry standard; see Institute of Education Sciences [2014]).

Matching substantially improved the balance for most variables compared with the full, unmatched comparison pool (results not shown). This improvement was important given how different the treatment population was compared with the national Medicare FFS population, as discussed previously. The unmatched comparison group shown in Table V.2 is the group that came out of the first stage of matching, which used nearest-neighbor matching to narrow the pool and make it much more similar to the treatment group than the initial pool of potential comparisons used at the start of the matching process. Although this first stage of matching substantially improved the balance for most variables compared with the unmatched comparison pool, the unmatched (but restricted) comparison pool was still quite different from the treatment group, necessitating a second stage of matching to select the comparison group used to estimate impacts.

D. Beneficiaries' outcomes and intervention impacts

In this section, we first present sample sizes and mean outcomes, by quarter, for the treatment and comparison groups. These mean outcomes provide context for understanding the contemporaneous differences estimates that follow; however, the differences in mean outcomes are not regression-adjusted and not impact estimates by themselves. Next, we present the results of the primary tests, by domain. Then, we present the results of the secondary tests (robustness checks) and assess whether the primary test results are plausible given the secondary test results and the implementation evidence. We end with a discussion of results, including why we did not draw conclusions about program impacts in any domain.

1. Sample sizes

The sample sizes for impact estimation differ depending on the outcome. We present sample sizes by domain.

Quality-of-care processes (Table V.3)

- The **diabetes preventive care composite measure** is defined among Medicare FFS beneficiaries ages 18 to 75 with diabetes. The sample size for the treatment group and the weighted comparison group ranged from 329 to 686 across the baseline year and the intervention year. This population accounted for 11 to 22 percent of the total Medicare FFS sample in the treatment and comparison groups.
- The **lipid profile measure for people with IVD** is defined among Medicare FFS beneficiaries ages 18 or older with IVD. The sample size for the treatment group and the weighted comparison group ranged from 770 to 1,492 across the baseline year and the intervention year. This population accounted for about 26 to 50 percent of the total Medicare FFS sample in the treatment and comparison groups. This percentage is higher than for the diabetes measure because (1) IVD (which is a broad disease category) is more common than diabetes among the treatment and comparison beneficiaries; and (2) the diabetes measure excludes beneficiaries older than 75 (about half our sample), whereas the IVD measure does not.
- The **14-day follow-up measure** is defined among Medicare FFS beneficiaries who had at least one hospital stay in the quarter. For the treatment group, the sample size ranged from 268 to 896 beneficiaries across the baseline and intervention quarters (accounting for 30 to 51 percent of all treatment beneficiaries in each quarter). For the weighted comparison group, the sample ranged from 193 to 834 across the baseline and intervention quarters (accounting for a similar proportion of the total comparison group).

Table V.3. Sample sizes and unadjusted means for quality-of-care process measures for Medicare FFS beneficiaries in the treatment and comparison groups for PBGH, by quarter

		Number of Medicare FFS beneficiaries			Mean outcomes		
Period	Quarter	T	C (not weighted)	C (weighted)	T	C	Difference (%)
Among those with diabetes and ages 18 to 75 , received all four recommended diabetes processes of care in the year (binary [yes or no]/beneficiary/year)							
Baseline	B1-B4[a]	686	1,409	665	44.5	43.8	0.6 (1.4%)
Intervention	I1-I4[a]	365	668	329	43.3	38.6	4.7 (12.2%)
Among those with ischemic vascular disease and ages 18 or older, received complete lipid profile in the year (binary [yes or no]/beneficiary/year)							
Baseline	B1-B4[a]	1,458	3,161	1,492	65.3	66.9	-1.6 (-2.4%)
Intervention	I1-I4[a]	770	1,586	783	56.8	65.9	-9.1 (-13.9%)
Among those with at least one inpatient admission in the quarter, all inpatient admissions in the quarter were followed by an ambulatory care visit with a primary care or specialist provider within 14 days of discharge (binary [yes or no]/beneficiary/year)							
Baseline	B1	357	747	342	77.9	74.2	3.7 (5.0%)
	B2	386	784	369	76.2	75.5	0.7 (0.9%)
	B3	432	831	409	76.4	75.5	0.9 (1.2%)
	B4	896	1,557	834	80.1	74.5	5.6 (7.6%)
Intervention	I1	563	898	459	75.8	68.6	7.3 (10.6%)
	I2	381	707	353	70.9	65.6	5.3 (8.1%)
	I3	324	520	253	67.6	61.8	5.8 (9.4%)
	I4	268	376	193	65.7	65.5	0.2 (0.2%)

Sources: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at the Centers for Medicare & Medicaid Services.

Notes: The baseline and intervention quarters are measured relative to the date that an individual enrolled in the IOCP (or pseudo-enrolled, if a comparison beneficiary). For example, the first baseline quarter for a beneficiary who enrolled on January 10, 2014, is from October 10, 2013, to January 9, 2014. For the same beneficiary, the first intervention quarter was from January 10, 2014, to April 9, 2014. In each period (baseline or intervention), the treatment and comparison groups in each quarter included all beneficiaries in the start of the quarter and who met other sample criteria—that is, they were alive and enrolled in FFS Medicare.

Table V.3 *(continued)*

The outcome means were weighted such that (1) each treatment beneficiary gets a weight of 1; and (2) each comparison beneficiary gets a matching weight equal to the reciprocal of the total number of comparison beneficiaries matched to the same treatment beneficiary. The difference between the treatment and comparison groups in a quarter was calculated by subtracting the mean outcome for the comparison group from the mean outcome for the treatment group. The percentage difference equaled that difference divided by the mean outcome for the comparison group.

[a] The quality-of-care process measures for diabetes and IVD were calculated over year-long periods, corresponding to the baseline and intervention quarters shown in the table.

B = baseline; C = control; FFS = fee-for-service; I = intervention; IOCP = Intensive Outpatient Care Program; IVD = ischemic vascular disease; PBGH = Pacific Business Group on Health; T = treatment.

For all three of these quality-of-care process measures, sample sizes were typically smaller in the intervention period than the baseline period. This reflects the fact that all treatment and comparison group members were observed for the full baseline period (because we constructed the groups this way), whereas beneficiaries were not necessarily observed for the full intervention period—either because they died, became unobservable in FFS claims (for example, by switching into managed care), or were not enrolled early enough to be followed up for the entire year.

Quality-of-care outcomes, service use and spending (all beneficiaries). The sample sizes for all outcomes in these three domains included the full treatment and comparison groups. In each baseline quarter and the first intervention quarter (I1), the treatment group included all 2,996 treatment group beneficiaries and 6,665 comparison group beneficiaries (Table V.4). The samples then decreased in each subsequent intervention quarter, as expected, because (1) some beneficiaries did not enroll or pseudo-enroll early enough to be followed for a second, third, or fourth intervention quarter before the end of our intervention period (March 31, 2015); and (2) some treatment or comparison group members exited the sample due to death or becoming unobservable. The net decrease in sample during the intervention period, and from quarter to quarter, was similar for the treatment and comparison groups: about 58 percent of treatment beneficiaries and 54 percent of comparison beneficiaries in I1 remained in the sample in I4, with the rate of attrition accelerating in successive intervention quarters.

Table V.4. Sample sizes and unadjusted mean outcomes (quality-of-care outcomes, service use, and spending) for Medicare FFS beneficiaries in the treatment and comparison groups for PBGH, by quarter

	Number of Medicare FFS beneficiaries			Inpatient admissions for ambulatory care-sensitive conditions (#/1,000 beneficiaries/quarter)			30-day unplanned hospital readmissions (#/1,000 beneficiaries/quarter)			All-cause inpatient admissions (#/1,000 beneficiaries/quarter)			Outpatient ED visit rate (#/1,000 beneficiaries/quarter)			Medicare Part A and B spending ($/beneficiary/month)		
Q	T	C (no wgt)	C (wgt)	T	C	Diff (%)	T	C	Diff (%)	T	C	Diff (%)	T	C	Diff (%)	T	C	Diff (%)
Baseline period																		
B1	2,995	6,665	2,996	27.4	25.6	1.8 (7.0%)	20.4	12.8	7.6 (59.2%)	146.9	140.4	6.5 (4.6%)	213.4	209.4	4.0 (1.9%)	$1,680	$1,636	$44 (2.7%)
B2	2,995	6,665	2,996	29.0	24.1	4.9 (20.3%)	20.0	16.4	3.7 (22.5%)	158.9	149.4	9.6 (6.4%)	235.4	218.9	16.5 (7.5%)	$1,782	$1,729	$53 (3.1%)
B3	2,996	6,665	2,996	34.4	33.3	1.1 (3.3%)	29.4	18.7	10.7 (57.1%)	179.2	164.3	14.9 (9.1%)	255.3	260.8	-5.5 (-2.1%)	$2,076	$1,991	$85 (4.3%)
B4	2,996	6,665	2,996	82.8	65.1	17.7 (27.2%)	64.8	60.1	4.7 (7.8%)	418.2	373.4	44.8 (12.0%)	380.5	362.3	18.1 (5.0%)	$3,560	$3,325	$235 (7.1%)
Intervention period																		
I1	2,996	6,665	2,996	50.1	39.3	10.8 (27.5%)	41.7	30.5	11.2 (36.6%)	212.6	186.9	25.8 (13.8%)	284.7	247.3	37.4 (15.1%)	$2,573	$2,465	$109 (4.4%)
I2	2,778	6,082	2,737	43.9	36.8	7.1 (19.2%)	30.2	28.8	1.4 (5.0%)	187.2	170.9	16.3 (9.5%)	253.1	225.3	27.8 (12.3%)	$2,267	$2,131	$137 (6.4%)
I3	2,315	4,916	2,260	38.9	30.5	8.4 (27.6%)	40.2	28.5	11.7 (41.1%)	193.1	149.9	43.2 (28.8%)	258.1	241.4	16.7 (6.9%)	$2,129	$1,988	$141 (7.1%)
I4	1,745	3,597	1,709	43.0	26.7	16.3 (60.8%)	35.0	25.4	9.6 (37.9%)	200.6	145.7	54.9 (37.7%)	237.1	223.0	14.1 (6.3%)	$2,338	$1,917	$422 (22.0%)

Sources: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at the Centers for Medicare & Medicaid Services.

Notes: The baseline and intervention quarters are measured relative to the date that an individual enrolled in the IOCP (or pseudo-enrolled, if a comparison beneficiary). For example, the first baseline quarter for a beneficiary who enrolled on January 10, 2014, is from October 10, 2013, to January 9, 2014. For the same beneficiary, the first intervention quarter was from January 10, 2014, to April 9, 2014. In each period (baseline or intervention), the treatment group each quarter included all beneficiaries in the start of the quarter and who met other sample criteria—that is, they were alive and enrolled in FFS Medicare.

The outcome means were weighted such that (1) each treatment beneficiary receives a weight of 1; and (2) each comparison beneficiary receives a matching weight equal to the reciprocal of the total number of comparison beneficiaries matched to the same treatment beneficiary. The difference between the treatment and comparison groups in a quarter was calculated by subtracting the mean outcome for the comparison group from the mean outcome for the treatment group. The percentage difference equaled that difference divided by the mean outcome for the comparison group.

B = baseline; C = comparison; Diff = difference; ED = emergency department; FFS = fee-for-service; I = intervention; IOCP = Intensive Outpatient Care Program; Q = quarter; T = treatment; wgt = weights.

NA = not available.

n.a. = not applicable.

2. Mean outcomes for the treatment and comparison groups, by domain and quarter

Quality-of-care processes. During the baseline year, 44.5 percent of treatment and 43.8 percent of comparison beneficiaries with diabetes and ages 18 to 75 received all four recommended processes of care (Table V.3). These rates decreased slightly in the first intervention year for the treatment group (to 43.3 percent), and more markedly for the comparison group (to 38.6 percent).

During the baseline year, 65.3 and 66.9 percent of the treatment and comparison beneficiaries, respectively, ages 18 or older with IVD received the recommended lipid test. These rates fell to 56.8 and 65.9 percent, respectively, in the intervention year.

From 76.2 to 80.1 percent of treatment group beneficiaries and 74.2 to 75.5 percent of comparison group beneficiaries who had any hospital stay in a baseline quarter had all of those stays followed by an ambulatory care visit within 14 days of discharge. These percentages decreased during the intervention period, so that by the fourth intervention quarter the value was 65.7 percent for the treatment group and 65.5 percent for the comparison group.

Quality-of-care outcomes. Inpatient admissions for ACSCs spiked in the fourth baseline quarter for both the treatment and comparison groups, increasing from 34.4 and 33.3 admissions per 1,000 beneficiaries in the third baseline quarter (B3) for treatment and comparison beneficiaries, respectively, to 82.8 and 65.1 admissions per 1,000 beneficiaries, respectively, in B4 (Table V.4). From B4 to I1, admissions dropped to 50.1 and 39.3 admissions per 1,000 beneficiaries, respectively, and over the following three quarters of the intervention period, declined to 43.0 and 26.7 admissions, respectively, by I4. During the intervention period, treatment beneficiaries had a higher ACSC admission rate than the comparison beneficiaries, with the difference ranging from 7.1 to 16.3 percent.

As with inpatient admissions for ACSCs, rates of 30-day unplanned readmissions spiked in the fourth baseline quarter relative to previous baseline quarters for both groups: treatment and comparison group members had 29.4 and 18.7 readmissions per 1,000 beneficiaries in B3, and 64.8 and 60.1 readmissions per 1,000 beneficiaries in B4 (Table V.4). During the intervention period, rates for 30-day unplanned readmissions ranged from 30.2 to 41.7 per 1,000 beneficiaries per quarter for the treatment group. This rate was substantially higher than the rate among the comparison group in each quarter, with the difference between the rates ranging from 1.4 to 11.7 percent of the comparison group rate in each quarter.

Service use. As in the quality-of-care outcomes domain, we observed a spike in B4 among outcomes in the service use domain, followed by a drop in each of the outcomes in the intervention period relative to B4 (Table V.4). The pattern of inpatient admissions and outpatient ED visits spiking in the fourth baseline quarter is consistent with the program's targeting, because staff at the PMGs could use recent hospitalizations or ED visits to identify high-risk patients for enrollment into the PBGH program (see Section III.A.1).

During the intervention period, service use rates for the treatment group were higher than for the comparison group. The mean admission rate for the treatment group was from 9.5 to 37.7

percent higher than for the comparison group and the outpatient ED visit rate was from 6.3 to 15.1 percent higher than for the comparison group.

Spending. Aligning with the other domains, total Medicare Part A and B spending per month increased during the baseline period for both treatment and comparison beneficiaries, with the largest jump from B3 to B4. Spending then dropped to $2,573 and $2,465 per beneficiary per month, respectively, from B4 to I1, and to $2,338 and $1,917 per beneficiary per month in I4. During the intervention period, spending was 4.4 to 22.0 percent higher among treatment beneficiaries than comparison beneficiaries.

3. Results for primary tests, by domain

Overview. In the quality-of-care processes domain, the regression-adjusted differences between the treatment and comparison groups were substantively large for two of the three measures, although one of these differences was favorable and the other unfavorable, and the combined effect estimate across all three measures was neither statistically significant nor larger than the substantive threshold. In contrast, in the other three study domains—quality-of-care outcomes, service use, and spending—we found substantively large and *unfavorable* differences between the treatment and comparison groups in all seven primary tests. Table V.5 summarizes these results.

Quality-of-care processes. The likelihood of receiving recommended processes of care for diabetes was 9.9 percent higher for the treatment group than its estimated counterfactual. (Our estimate of the counterfactual is the treatment group mean minus the regression-adjusted contemporaneous differences estimate.) This favorable difference was larger than the substantive threshold, but not statistically significant after accounting for multiple comparisons in the domain ($p = 0.29$). The likelihood of receiving a complete lipid profile among people with IVD was 8.7 percent lower for the treatment group than its estimated counterfactual, an unfavorable estimate that is larger than the substantive threshold for this outcome (3.0 percent). The likelihood of receiving an ambulatory care visit within 14 days of hospital discharge was 2.0 percent higher in the treatment group than its estimated counterfactual, a (favorable) difference that was neither substantively large nor statistically significant.

The combined estimate across the three measures in the quality-of-care processes domain was 1.1 percent, a favorable point estimate that was not larger than the substantive threshold of 3.0 percent. The statistical power to detect substantively large effects, however, was poor (17.5 to 41.1 percent) for all three quality-of-care process measures individually and, in addition, combined across the measures (19.9 percent).

Table V.5. Results of primary tests for PBGH

Primary test definition					Statistical power to detect an effect that is[a]		Results			
Domain (# of tests in domain)	Outcome (units)	Time period for impacts	Population	Substantive threshold (expected direction of effect)[b]	Size of the substantive threshold	Twice the substantive threshold[c]	Treatment group mean	Regression-adjusted difference between the treatment and estimated counterfactual (standard error)[b]	Percentage difference[d]	*p*-value[e]
Quality-of-care processes (3)	Received all four recommended diabetes processes of care in the year (binary [yes or no]/beneficiary/year)	Average over quarters I1–I4[f]	Medicare FFS beneficiaries ages 18 to 75 with diabetes	3.0% (+)	17.5	27.8	43.3	3.9 (3.4)	9.9%	0.29
	Received complete lipid profile in the year (binary [yes or no]/beneficiary/year)	Average over quarters I1–I4[f]	Medicare FFS beneficiaries ages 18 or older with ischemic vascular disease	3.0% (+)	33.6	66.8	56.8	-5.4 (2.2)	-8.7%	0.98
	All inpatient admissions within a quarter were followed by an ambulatory care visit with a primary care or specialist provider within 14 days (binary [yes or no]/beneficiary/year)	Average over quarters I1–I4	Medicare FFS beneficiaries with at least one hospital stay in the quarter	3.0% (+)	41.1	79.8	70.0	1.4 (1.9)	2.0%	0.43
	Combined	Average over quarters I1–I4	Varies by outcome	3.0% (+)	19.9	34.0	n.a.	n.a.	1.1%	0.38
Quality-of-care outcomes (2)	Inpatient admissions for ambulatory care-sensitive conditions (#/1,000 beneficiaries/quarter)	Average over quarters I1–I4	All Medicare FFS beneficiaries	15.0% (-)	63.6	97.6	44.0	9.5 (3.2)	27.6%	>0.99
	30-day unplanned readmissions (#/1,000 beneficiaries/quarter)	Average over quarters I1–I4	All Medicare FFS beneficiaries	15.0% (-)	50.1	90.1	36.8	7.8 (3.4)	26.9%	0.98
	Combined (%)	Average over quarters I1–I4	All Medicare FFS beneficiaries	15.0% (-)	60.3	96.4	n.a.	n.a.	27.2%	>0.99

Table V.5 *(continued)*

Primary test definition					Statistical power to detect an effect that is[a]		Results			
Domain (# of tests in domain)	Outcome (units)	Time period for impacts	Population	Substantive threshold (expected direction of effect)[b]	Size of the substantive threshold	Twice the substantive threshold[c]	Treatment group mean	Regression-adjusted difference between the treatment and estimated counterfactual (standard error)[b]	Percentage difference[d]	*p*-value[e]
Service use (2)	All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	Average over quarters I1–I4	All Medicare FFS beneficiaries	3.75% (-)	32.0	63.6	198.4	31.8 (7.7)	19.1%	>0.99
	Outpatient ED visits (#/1,000 beneficiaries/quarter)	Average over quarters I1–I4	All Medicare FFS beneficiaries	3.75% (-)	35.4	70.3	258.3	15.0 (10.1)	6.2%	0.88
	Combined (%)	Average over quarters I1–I4	All Medicare FFS beneficiaries	3.75% (-)	40.6	78.9	n.a.	n.a.	12.6%	>0.99
Spending (1)	Medicare Part A and B spending ($/beneficiary/month)	Average over quarters I1–I4	All Medicare FFS beneficiaries	3.75% (-)	40.5	78.9	$2,327	$174 ($77.5)	8.1%	0.98

Sources: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at the Centers for Medicare & Medicaid Services.

Note: The results for each outcome are based on a contemporaneous differences regression model, as described in the text. Estimates are calculated for Medicare beneficiaries who are observable in the relevant time period: that is, beneficiaries who are enrolled in Medicare FFS (Part A and B), are alive, and have Medicare as their primary payer. Additional sample restrictions apply to the quality-of-care process measures; see text for details.

[a] The power calculation is based on actual standard errors from the analysis. For example, in the last row, an 8.1 percent effect on Medicare Part A and B spending (from the counterfactual of $2,327 + $174 = $2,501) would be a change of $203. Given the standard error of $77.50 from the regression model, we would be able to detect a statistically significant result 40.5 percent of the time if the impact was truly $203, assuming a one-sided statistical test at the $p = 0.10$ significance level.

[b] The substantive threshold is the impact as a percentage of the counterfactual. The counterfactual is the outcome the treatment group would have had in the absence of the HCIA-funded intervention. Our estimate of the counterfactual is the treatment group mean minus the regression-adjusted contemporaneous differences estimate.

[c] We show statistical power to detect a very large effect (twice the size of the substantive threshold) because this provides additional information about the likelihood that we will find effects if the program is indeed effective. If power to detect effects is less than 75 percent even for a very large effect, then the evaluation is extremely poorly powered for that outcome.

[d] Percentage difference is calculated as the regression-adjusted difference between the treatment and comparison groups, divided by the adjusted comparison group mean.

Table V.5 *(continued)*

[e] *p*-values test the null hypothesis that the regression-adjusted estimate is less than or equal to zero for outcomes in the quality-of-care processes domain, or greater than or equal to zero in all other domains (a one-sided test). Because it is a one-sided test, as the contemporaneous differences estimate approaches infinity in an unfavorable direction (negative for process-of–care measures and positive for all other measures), the *p*-value approaches 1, whereas it would approach 0 in a two-sided test. We adjusted the *p*-values from the primary test results for the multiple (three) comparisons made within the quality-of-care processes domain, and (separately) for the two comparisons made within the quality-of-care outcomes domain and for the two comparisons made within the service use domain.

[f] We estimated impacts as the average across intervention quarters 1 through 4 for all outcomes but two: the quality-of-care process measures for diabetes and ischemic vascular disease. For those two measures, we calculated outcomes instead over year-long periods (rather than quarters). The impact estimates apply to the same time period—that is, the year that corresponds to intervention quarters 1 through 4—but the estimate is not an average of quarterly estimates.

ED = emergency department; FFS = fee-for-service; HCIA = Health Care Innovation Award; PBGH = Pacific Business Group on Health.

n.a. = not applicable.

Quality-of-care outcomes. The rate of ACSC admissions for the treatment group during the primary test period was 27.6 percent higher than our estimate of the counterfactual, and the rate of unplanned readmissions was 26.9 percent higher. These higher rates for the treatment group are in the unfavorable direction (indicating an increase in ACSC admissions and readmissions) and larger than the substantive threshold for each measure of 15.0 percent. After combining results across the two outcomes in this domain, the combined effect was 27.2 percent, larger than the substantive threshold of 15.0 percent and in the unfavorable direction. We cannot assess whether this unfavorable result is statistically significant because our one-sided statistical tests assess only improvements in outcomes.

The statistical power to detect effects the size of the substantive threshold was marginal for both ACSC admissions (63.6 percent) and 30-day unplanned readmissions (50.1 percent). Power was also marginal (60.3 percent) for the combined effect in the domain.

Service use. Treatment group admission and outpatient ED visit rates were 19.1 and 6.2 percent higher, respectively, than their estimated counterfactuals, both of which are substantively large and unfavorable. When combining results across the two outcomes in this domain, the combined effect was 12.6 percent, larger than the substantive threshold of 3.75 percent and in the unfavorable direction. Power to detect effects that were the size of the substantive thresholds was poor for the admissions and outpatient ED visit measures individually (32.0 and 35.4 percent, respectively) and poor (40.6 percent) for the two outcomes combined.

Spending. The treatment group averaged $2,327 per beneficiary per month in Part A and B spending during the primary test period, a value 8.1 percent (or $174) higher than its estimated counterfactual. This unfavorable estimate is larger than the substantive threshold for this outcome (3.75 percent). Statistical power to detect an effect the size of the substantive threshold was poor (40.5 percent).

4. Results for secondary tests

Sample attrition. PBGH's treatment and matched comparison groups had slightly different mortality rates during the intervention period, with treatment beneficiaries more likely to survive. By the start of their second intervention quarter, 2.6 percent of the treatment and 3.5 percent of the comparison beneficiaries had died; 5.3 percent of the treatment and 6.7 percent of the comparison beneficiaries had died by the start of the third intervention quarter, and 7.7 percent of the treatment group and 9.0 percent of the comparison group had died by the fourth intervention quarter (results not shown). This differential attrition could indicate that, despite baseline equivalence on observable characteristics, the comparison group beneficiaries were less healthy than the treatment group beneficiaries in ways that increased the risk of death but were not observable in Medicare claims. Alternatively, it is possible that the treatment and comparison beneficiaries had similar health at enrollment, in which case this differential attrition could make the treatment group appear less healthy than the comparison group over time (because sicker beneficiaries were disproportionately more likely to survive in the treatment group and continue to be observed than in the comparison group). In either case, if the differential mortality is related to (or correlated with) study outcomes, it might bias estimates of program impacts. However, the magnitude of the differential attrition was not large.

Results limiting the sample to beneficiaries with at least 12 months of potential exposure to the intervention. The secondary test results limited to beneficiaries who enrolled by June 30, 2014—which allows at least 12 months of potential exposure to the intervention—are generally consistent with those presented in this report for the larger sample (results not shown). The results show treatment beneficiaries had much higher rates of inpatient admissions for ACSCs, inpatient admissions, and spending than the comparison beneficiaries in the intervention period compared to difference in the baseline period. Given these findings, we concluded the substantively large unfavorable effects observed in the primary tests were not likely a result of the treatment group criteria but were more in line with limitations of the comparison group.

5. Consistency of quantitative estimates with implementation findings

The impact estimates in the primary tests were implausible given the implementation findings. The primary test results showed that the differences between the treatment and comparison groups were substantively large and unfavorable for all measures in the quality-of-care outcomes (ACSC admissions and readmissions), service use (inpatient admissions and ED visits), and spending (Medicare part A and B spending) domains, and one of the three quality-of-care processes (patients with IVD receiving the recommended lipid test).

There is nothing in the implementation evidence that explains how the IOCP intervention could have caused the large observed increase in these six outcomes. Given the magnitude of the unfavorable findings, we do not believe the intervention itself could have driven these results. Moreover, it is difficult to imagine how care managers could have encouraged, for example, higher rates of hospital admissions.

Other evidence from our implementation study shows that, despite implementing aspects of the IOCP intervention as planned, including meeting FTE and staff training goals and providing care management services to enrollees, PBGH experienced barriers to obtaining claims data from PMGs to generate risk score reports that could be used to identify and enroll beneficiaries, and to recuiting PMGs to participate. Of these barriers, the most significant resulted in a delayed implementation of PBGH's planned approach for identifying high-risk patients using a risk-stratification algorithm from Milliman. PBGH was unable to give PMGs timely, standard risk score reports appropriate for identifying beneficiaries eligible for the program. Although PBGH provided eligibility guidelines to the PMGs, in practice, determination of eligibility was left to the individual PMGs, and in some cases to individual PCPs. Despite matching on a number of characteristics that capture beneficiaries' prior utilization, health status, and risk—all characteristics that were part of PBGH's revised eligibility guidelines—it is likely that unobserved differences remain between the treatment and matched comparison groups. It is more plausible that the impact results are due to these differences than to something the care management intervention did or did not do.

6. Conclusions about program impacts

Based on all evidence currently available, we are unable to draw conclusions about program impacts during the primary test period (Table V.6). We could not fully replicate the process used

to identify and enroll participants into PBGH's HCIA program using claims data. As a result, having found implausible impact estimates given the implementation evidence, we believe that we could not define a valid comparison group necessary to draw conclusions about program impacts on patients' outcomes, despite achieving good matches between treatment and comparison beneficiaries on baseline characteristics observed in Medicare claims.

Table V.6 Conclusions about the impacts of PBGH's HCIA program on patients' outcomes, by domain

		Evidence	
Domain	Conclusion	Primary test result(s)	Primary test result(s) plausible given implementation evidence?
Quality-of-care processes	None	• Combined effect across the three outcomes in the domain was neither substantively large nor statistically significant; power was poor	No
Quality-of-care outcomes	None	• Combined effect across the two measures in the domain was unfavorable and larger than the substantive threshold	No
Service use	None	• Combined effect across the two measures in the domain was unfavorable and larger than the substantive threshold	No
Spending	None	• The single test in the domain was unfavorable and larger than the substantive threshold	No

Source: Table V.5.

HCIA = Health Care Innovation Award; PBGH = Pacific Business Group on Health.

VI. DISCUSSION AND CONCLUSIONS

PBGH used its $19.1 million HCIA to provide care management services to high-risk participants. Care managers funded and trained by the program were embedded in primary care practices and worked with participants' PCPs to develop and implement personalized care plans for program participants. Care managers interacted with participants one on one to understand their medical and social needs, help participants manage their conditions, connect them to appropriate clinical services, and work with them to overcome socioeconomic obstacles to care. Through the program, PBGH aimed to reduce total Medicare spending by reducing participants' need for acute care—such as inpatient admissions and ED visits—and by increasing use of appropriate primary and specialty care.

We were unable to draw conclusions about program impacts. The primary test results from our impact evaluation showed that the differences between the treatment and comparison groups were substantively large and unfavorable in the quality-of-care outcomes, service use, and spending domains, driven by large unfavorable point estimates for all five outcomes in these three domains: ACSC admissions, 30-day unplanned readmissions, inpatient admissions, outpatient ED visits, and Medicare Part A and B spending. However, these results were not

plausible given the implementation evidence, as nothing about the program could plausibly have caused such large, unfavorable effects. The treatment and comparison beneficiaries were well matched on observable characteristics at baseline, and we found no notable differences in sample attrition during the intervention period between the two groups. We matched on a number of characteristics that capture beneficiaries' prior utilization, health status, and risk (characteristics that were part of PBGH's revised eligibility guidelines), and added several steps to the matching process—such as dropping hard-to-match treatment group members and adding post-matching diagnostics on several indicators of health care use and risk—to ensure we achieved balance not only on averages but on the distribution of values across the two groups. As a result, we believe there might have been unobservable differences between the groups, likely as a result of the process through which PMGs identified beneficiaries eligible for the program, that influenced the results.

The implementation findings support this conclusion. PBGH experienced barriers to implementing its planned approach for identifying high-risk patients for enrollment into the program. In response, PBGH provided eligibility guidelines to the PMGs for IOCP participation concerning beneficiaries' recent service use for acute care, number of chronic conditions, and medication use. However, in practice, determination of eligibility was left to the individual PMGs, and in some cases to individual PCPs. During the site visits to PMGs, we learned that one of the alternate methods that PMGs had developed to recruit participants was the direct referral by a PCP, often through a warm handoff to a care management team member. This method might have relied less on utilization review, instead focusing on clinical judgement and information about the beneficiaries' social context and need for social services; as well as whether patients' acuity level was trending up versus remaining stable or trending down. Although we matched on a number of characteristics such as ED visits, hospital discharges and Medicare spending in the quarter prior to enrollment in an attempt to capture changes in acuity levels, it is possible that these claims-based measures were insufficient to fully capture beneficiaries' prospective outcomes. The substantively large and unfavorable findings and implementation findings about the program's identification process confirmed our reservations about using observable (claims-based) characteristics to select a comparison group that could serve as a valid counterfactual.

CMMI and other stakeholders could consider changes to the design of similar programs in the future to increase the potential to draw conclusions about program impacts on participants' outcomes. First, it is important to acknowledge that clinicians might be best suited to identify the beneficiaries who could benefit most from a particular intervention. In future programs in which awardees believe clinician judgment is important, it might be necessary to require patient-level random assignment to obtain unbiased estimates of the interventions' effects.

Second, although PBGH collected and shared with evaluators the PMGs' information on program participants, such as patient identifiers and dates of enrollment, the awardee did not require PMGs to collect similar information on beneficiaries identified as eligible to participate in the intervention but who declined to enroll. For voluntary programs such as the IOCP, prospective enrollees' decisions about whether to enroll will always introduce potential for bias if the analysis is limited to those who actually enrolled, as in the PBGH evaluation. For example,

beneficiaries who consented to participate in the IOCP might have done so due to expectations of a decline in future health status and/or a subsequent increased need for services. The substantively large and unfavorable differences we observed in the primary test results could be due, in part, to differences in the construction of the treatment and comparison groups—with the treatment group limited to beneficiaries who decided to enroll, whereas the comparison group contains those who would have participated if offered the program and those who would not have enrolled had they been given the chance.

Third, in the future, if a program proposes an objective method for determining targeting—and can replicate the method in claims—any deviations should be documented and based on criteria that can be readily measured and, therefore, replicated. Here, lessons can be learned from the Rutgers Center for State Health Policy (CSHP), another awardee in the HCIA Primary Care Redesign portfolio presented in Chapter 7. Similar to PBGH, CSHP used its award to implement a care management/care coordination program at multiple provider organizations. The participating sites differed in how they defined their target population, but site-specific eligibility criteria were well defined and based on information available in claims data, facilitating the use of the same criteria when defining the pool of potential comparison group members. The findings from CSHP's HCIA suggest that there is potential for drawing conclusions for beneficiary-level care management interventions even when randomization is not used, provided the program consistently uses well-defined eligibility criteria that can be replicated using available data sources.

This page has been left blank for double-sided copying.

REFERENCES

Boards of Trustees, Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds. "2013 Annual Report of the Boards of Trustees of the Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds." Table V.D1. Washington, DC: Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds, 2013. Available at http://downloads.cms.gov/files/TR2013.pdf. Accessed August 13, 2014.

Chronic Conditions Data Warehouse. "Table A.1. Medicare Beneficiary Counts for 2003–2012." Baltimore, MD: Centers for Medicare & Medicaid Services, 2014a. Available at https://www.ccwdata.org/cs/groups/public/documents/document/ccw_website_table_a1.pdf. Accessed November 19, 2014.

Chronic Conditions Data Warehouse. "Table B.2. Medicare Beneficiary Prevalence for Chronic Conditions for 2003 Through 2012." Baltimore, MD: Centers for Medicare & Medicaid Services, 2014b. Available at https://www.ccwdata.org/cs/groups/public/documents/document/ccw_website_table_b2.pdf. Accessed November 19, 2014.

Gerhardt, Geoffrey, Alshadye Yemane, Keri Apostle, Allison Oelschlaeger, Eric Rollins, and Niall Brennan. "Evaluating Whether Changes in Utilization of Hospital Outpatient Services Contributed to Lower Medicare Readmission Rate." *Medicare & Medicaid Research Review,* vol. 4, no. 1, 2014, pp. E1–E13.

Keith, Rosalind, Rumin Sarwar, Boyd Gilman, Catherine DesRoches, and Lorenzo Moreno. "Findings for Pacific Business Group on Health." In Moreno, Lorenzo, Boyd Gilman, Greg Peterson, Catherine DesRoches, Sheila Hoag, Linda Barterian, Laura Blue, Katherine Bradley, Emily Ehrlich, Kristin Geonnotti, Lauren Hula, Keith Kranker, Rumin Sarwar, Rachel Shapiro, KeriAnn Wells, Joseph Zickafoose, Sandi Nelson, Frank Yoon with the Implementation Team, Impact Team, Data Processing Team, Surveys Team, and Production Coordination and Editorial Team. "Evaluation of the Health Care Innovation Awards (HCIAs): Primary Care Redesign Programs. Second Annual Report, Volumes I and II." Princeton, NJ: Mathematica Policy Research, December 11, 2015.

Health Indicators Warehouse. "Average Age of Medicare Beneficiaries (mean)." Hyattsville, MD: National Center for Health Statistics, HIW, 2014a. Available at http://www.healthindicators.gov/Indicators/Average-age-of-Medicare-beneficiaries-mean_308/Profile/ClassicData. Accessed November 19, 2014.

Health Indicators Warehouse. "Hospital Inpatient Medicare Admissions (per 1,000 beneficiaries)." Hyattsville, MD: National Center for Health Statistics, HIW, 2014b. Available at http://www.healthindicators.gov/Indicators/Hospital-inpatient-Medicare-admissions-per-1000-beneficiaries_2001/Profile/ClassicData. Accessed August 13, 2014.

Health Indicators Warehouse. "Medicare Beneficiaries Who Are Also Eligible for Medicaid (percent)." Hyattsville, MD: National Center for Health Statistics, HIW, 2014c. Available at http://www.healthindicators.gov/Indicators/Medicare-beneficiaries-eligible-for-Medicaid-percent_317/Profile/ClassicData. Accessed August 4, 2015.

Institute of Education Sciences. "What Works Clearinghouse: Procedures and Standards Handbook, Version 3.0." Washington, DC: U.S. Department of Education, IES, 2014. Available at http://ies.ed.gov/ncee/wwc/DocumentSum.aspx?sid=19. Accessed September 15, 2014.

Peikes, Deborah, Stacy Dale, Eric Lundquist, Janice Genevro, and David Myers. "Building the Evidence Base for the Medical Home: What Sample and Sample Size Do Studies Need? White Paper." AHRQ Publication No.11-0100-EF. Rockville, MD: Agency for Healthcare Research and Quality, October 2011.

Rosenthal, M.B., S. Alidina, M. Friedberg, S. Singer, D. Eastman, Z. Li, and E. Schneider. "A Difference-in-Differences Analysis of Changes in Quality, Utilization, and Cost Following the Colorado Multi-Payer Patient-Centered Medical Home Pilot." *Journal of General Internal Medicine,* vol. 31, no. 3, 2016, pp. 289–296.

Stuart, Elizabeth A. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science,* vol. 25, no. 1, 2010, pp. 1–21.

Welch, W. Pete, Alison Evans Cuellar, Sally C. Stearns, and Andrew B. Bindman. "Proportion of Physicians in Large Group Practices Continued to Grow in 2009–11." *Health Affairs,* vol. 32, no. 9, 2013, pp. 1659–1666.

CHAPTER 6

PEACEHEALTH KETCHIKAN MEDICAL CENTER

Purvi Sevak, Boyd Gilman, Victoria Peebles, Greg Peterson, Laura Blue, Keith Kranker, Kate Stewart, Sheila Hoag, and Lorenzo Moreno

This page has been left blank for double-sided copying.

PEACEHEALTH KETCHIKAN MEDICAL CENTER

CHAPTER SUMMARY

Introduction. PeaceHealth Ketchikan Medical Center (PeaceHealth) used its $3.2 million Health Care Innovation Award to implement the Better Health Through Coordinated Care—A Plan for Southeast Alaska program (hereafter referred to as the coordinated care program). The program involved four interrelated components: (1) general transitional care services for all patients discharged from the PeaceHealth Ketchikan Medical Center and intensive transitional care services for patients discharged with a diagnosis of congestive heart failure (CHF); (2) short-term outpatient care management for patients with a temporary medical or social hurdle; (3) longer-term outpatient case management for patients requiring ongoing assistance to effectively manage their chronic conditions; and (4) population health management, including refinement of the scrub-and-huddle process, outreach to paneled patients to improve preventive care, and deployment of a nurse practitioner to fill demand for same-day appointments. The scrubbing process involved reviewing patients' medical records to identify outstanding care needs, and the huddling process involved regularly scheduled team meetings to review and discuss patients' needs before their visits. Over three years, PeaceHealth expected to reduce 30-day hospital readmission rates for patients with CHF by 20 percent, emergency department costs for patients with chronic conditions by 75 percent, and total costs for patients with chronic conditions by 15 percent.

Objectives. This report aims to (1) describe the design and implementation of PeaceHealth's intervention, including the role of primary care providers (PCPs) (including physicians, nurse practitioners, and physician assistants) and the extent to which anticipated changes in providers' behavior occurred; (2) assess impacts of the intervention on patients' outcomes and Medicare Part A and B spending during the three-year award period; and (3) use both implementation and impact findings to identify possible explanations for the observed impacts.

Methods. We reviewed PeaceHealth's program documents and self-monitoring metrics; conducted interviews with PeaceHealth leadership, program, and frontline staff; and surveyed trainees and participating clinicians. To estimate impacts, we compared outcomes for Medicare fee-for-service patients served by the two treatment clinics with outcomes for Medicare patients served by 57 comparison practices in rural Alaska, adjusting for any differences in outcomes between the two groups during a one-year baseline period.

Program implementation. The available evidence indicates that, after an initial one-year delay, the intervention was implemented largely as planned. PeaceHealth hired 9.5 full-time equivalent staff and served 3,881 unique participants through Year 3. However, PeaceHealth was unable to provide a clear description of its service delivery protocols, which made it difficult to distinguish between the delivery of short-term care management versus longer-term case management services and to evaluate whether the awardee delivered the intervention services as intended. A high degree of adaptation and flexibility led to a lack of standardization of the model (different care coordinators undertook different activities). In addition, PeaceHealth did not

report metrics on the types of intervention services the program provided to patients that we could use to evaluate the intervention.

Clinicians' perceptions of the intervention's effects on the care they provided to patients. PeaceHealth's program design required PCPs to refer patients for care coordination, to work collaboratively with care coordinators and medical office assistants (MOAs) to address patients' needs, and to participate in regular scrub-and-huddle meetings. Qualitative and survey data suggest that PeaceHealth engaged PCPs as planned: virtually all surveyed PCPs reported being aware of the program. In addition, most PCPs reported that they believed the intervention improved the quality, timeliness, and patient-centeredness of care they provided to patients.

Impacts on patients' outcomes. We found substantively large and favorable impacts in the quality-of-care process domain; this was driven by estimates of program impacts on implementing process-of-care guidelines for patients with diabetes. We were unable to draw conclusions in the other three study domains (quality-of-care outcomes, service use, and spending) because baseline comparisons between the treatment and comparison groups and the results from robustness checks suggest that the two groups differed in important ways that could have led to biased conclusions on program impacts for these outcomes. The difficulty in finding valid comparison practices stems from the fact that the program was implemented in two distinct practices, one unusually large (for the region) and one particularly small practice on a remote island. Further, because the treatment group was small (due to only two participating practices and the fact that we had to limit the sample to Medicare beneficiaries), the impact estimates were statistically imprecise. Despite the fact that we cannot draw conclusions in these three domains, this report includes their results for transparency, and to enable policymakers to review the evidence and draw their own conclusions.

Conclusion. The evidence indicates that PeaceHealth's program improved quality-of-care processes—particularly implementing the process-of-care guidelines for patients with diabetes. This is consistent with PeaceHealth's emphasis on high-risk patients. However, we are unable to draw conclusions on program impacts in the quality-of-care outcomes, service use, and spending domains. The Center for Medicare & Medicaid Innovation and other stakeholders could consider a number of changes to the design of similar programs in the future to increase the potential to draw conclusions about program impacts on patients' outcomes. These include randomization of patients who receive the new program services and use of more timely, high quality Medicaid data for nondual Medicaid beneficiaries to improve the statistical power and the relevance of the impact estimates to the intervention.

Summary of intervention and impact results for PeaceHealth

Intervention description			
Awardee description		Medical center (with a 25-bed critical access hospital) and two affiliated primary care clinics	
Award amount ($ millions)		$3.2 million	
Award extended beyond June 2015?		No	
Location		Remote island communities in southeastern Alaska	
Target population		All patients served by 2 primary care clinics, with some intervention components targeted to specific patients within those clinics	
Interventions		Conducted population health and disease management activities through several program components: • Transitional care, in which nurses (1) called each patient (once only) to review discharge instructions and medications, and assess need for further support; and (2) made additional calls to patients with CHF to assess signs of excess fluid and encourage follow-up with a PCP • Individualized care management for patients with specific conditions,[a] provided by 6 HCIA-funded nurses and a social worker • Expanded use of population health IT and scrub-and-huddle process[b]	
Metrics of intervention delivered		12,599 direct encounters with 3,881 unique patients 60 to 80 percent of targeted patients (depending on month) received transitional care	
Impact evaluation methods			
Core design		Difference-in-differences model with comparison group (unmatched)[c]	
Treatment group	Definition	Medicare FFS beneficiaries attributed to 2 PeaceHealth treatment clinics	
	# of beneficiaries during primary test period[d]	996 to 1,101	
Comparison group definition		Medicare FFS beneficiaries attributed to 57 (unmatched) comparison practices	
Impact results: Quality-of-care processes domain			
Ambulatory care visit within 14 days of discharge (% of beneficiaries/quarter)		Comparison mean[e]	40.8%
		Impact estimate (% difference)	-14.7 pp (-36.0%)
Received all four recommended diabetes processes of care (% of beneficiaries/year)		Comparison mean[e]	20.1%
		Impact estimate (% difference)	+11.5 pp (+57.2%)**
Combined impact estimate[f]		-5.4%[g]	
Impact conclusion[h]		**Statistically significant favorable effect**	

Note: See this chapter for details on the intervention, impact methods, and impact results. As explained in the chapter, we did not draw impact conclusions in the three other outcome domains: quality-of-care outcomes, service use or spending.

[a] Program staff initially targeted those with CHF or diabetes, then expanded to those with hypertension and high-risk pregnancies.

[b] Scrubbing involved reviewing a patient's medical records to identify outstanding care needs, such as colorectal screenings, immunizations, laboratory tests, or mammograms. The huddling process involved a team meeting to review a patient's needs before a regularly scheduled visit.

[c] The comparison group was unmatched because statistical matching did not meaningfully improve balance on prespecified matching variables relative to the full pool of potential comparison practices. We relied on the difference-in-differences model to account for any differences in outcomes that stemmed from persistent (time-invariant) differences between the treatment and comparison practices.

[d] For some outcome measures the sample is limited to a relevant subset of beneficiaries.

[e] The comparison mean is the estimate of the outcome the treatment group beneficiaries would have had if they had not received the intervention. It is equal to the mean for the treatment group over the intervention quarters (in the primary test period) minus the impact estimate.

[f] The combined estimate is the average across all the individual estimates in each domain, in which the impact estimate for each individual outcome is expressed as percentage change relative to the comparison group.

[g] The combined impact includes the two estimates in this table plus one test not shown here (14-day ambulatory care follow-up visits for Medicare FFS beneficiaries with CHF, diabetes, and/or hypertension) but that is reported in the PeaceHealth chapter.

[h] We drew conclusions at the domain level based on the results of prespecified primary tests, secondary tests (robustness checks), and consistency with implementation evidence. For each domain, we planned to draw one of five conclusions: (1) Statistically significant favorable effect (the highest level of evidence), (2) Substantively important (but not statistically significant) favorable effect, (3) Substantively important (but not statistically significant) unfavorable effect, (4), No substantively large effect, and (5) Indeterminate effect. Section V.A.8 of this report describes the decision rules we used to reach each of these possible conclusions.

*Significantly different from zero at the .10 level, one-tailed test.

**Significantly different from zero at the .05 level, one-tailed test.

***Significantly different from zero at the .01 level, one-tailed test.

CHF = congestive heart failure; FFS = fee-for-service; HCIA = Health Care Innovation Award; PCP = primary care provider; pp = percentage point.

This page has been left blank for double-sided copying.

I. INTRODUCTION

This report presents findings from the evaluation of the Health Care Innovation Award (HCIA) received by PeaceHealth Ketchikan Medical Center (PeaceHealth), with a focus on program impacts on patients' outcomes. Section II provides an overview of PeaceHealth's HCIA-funded intervention and the design of the impact evaluation. Section III describes the design and implementation of the intervention, including how the program could be expected to affect study outcomes through changes in patients' and providers' behaviors. In Section IV, we assess the evidence of the extent to which planned changes in providers' behavior occurred. Section V describes our methods for, and results and conclusions from, estimating program impacts on patients' outcomes in four domains: (1) quality-of-care processes, (2) quality-of-care outcomes, (3) service use, and (4) Medicare spending. As we will describe, we are unable to draw conclusions about program impacts in three of the four domains (all but the quality-of-care process domain) due to concerns about potential biases in the impact estimates. We still present the data for transparency and so that readers can judge the evidence for themselves. Section VI concludes, including a discussion about ways that the Centers for Medicare & Medicaid Services (CMS) or other stakeholders could modify the program design for future tests of interventions like PeaceHealth's to increase the chances of drawing unbiased impact conclusions.

II. OVERVIEW OF PEACEHEALTH'S HCIA-FUNDED INTERVENTION AND THE IMPACT EVALUATION

A. PeaceHealth's HCIA-funded intervention

PeaceHealth received a three-year, $3.2 million award to implement the Better Health Through Coordinated Care—A Plan for Southeast Alaska program (hereafter referred to as the coordinated care program). Table II.1 summarizes key features of the program. PeaceHealth's goals were to reduce 30-day hospital readmission rates for patients with congestive heart failure (CHF) by 20 percent, emergency department (ED) costs for patients with chronic conditions by 75 percent, and total costs for patients with chronic conditions by 15 percent. PeaceHealth expected to achieve these outcomes through four interrelated intervention components: (1) transitional care for all patients discharged from PeaceHealth Ketchikan's ED or hospital who were on a PeaceHealth provider panel, with more intensive services for patients with CHF; (2) short-term care management for patients with social and behavioral health needs or chronic conditions; (3) longer-term case management for patients with chronic conditions who needed three or more encounters to manage their conditions; and (4) population health management for patients with uncontrolled chronic conditions and patients in need of routine screenings, with a special focus on patients with diabetes. PeaceHealth expected that these intervention components would help patients better manage their own conditions and help avoid unnecessary care (Section III.A.3 describes the awardee's theory of action in detail).

Table II.1. Summary of PeaceHealth Ketchikan Medical Center HCIA program and our evaluation for estimating its impacts on patients' outcomes

Program description	
Award amount	$3.2 million
Award start date	July 1, 2012
Implementation date	October 18, 2012
Award end date	June 30, 2015
Awardee description	PeaceHealth Ketchikan Medical Center is a hospital in rural southeastern Alaska. Two clinics are associated with the medical center and the PCR program, one in Ketchikan and one in Craig, both of which serve remote island communities.
Intervention overview	The intervention provided care coordination and care management services to help patients better manage their acute and chronic conditions in two clinics.
Intervention components	1. **Transitional care.** Care coordinators provided general transitional care services to all patients discharged from the PeaceHealth Ketchikan Medical Center and intensive transitional care services to patients discharged with a diagnosis of CHF. 2. **Short term-care management.** Care coordinators provided short-term outpatient care management to all patients with social and behavioral health needs or with a chronic condition. 3. **Long-term case management.** Care coordinators provided longer-term outpatient case management to all patients requiring ongoing assistance to effectively manage their chronic conditions. 4. **Population health management.** Care coordinators identified and contacted patients with uncontrolled chronic conditions (initially limited to diabetes, CHF, and hypertension, and later expanded to include high-risk pregnancies) and scheduled appointments for patients who needed a routine screening or test (including mammograms, colorectal screenings, immunizations, and diabetic tests). A nurse practitioner was hired to offer same-day appointments to provide patients with more timely access to care.
Target population	1. The transitional care component was targeted at all patients discharged from the PeaceHealth Ketchikan Medical Center ED or hospital. 2. The short-term care management and longer-term case management components were targeted at all patients with chronic conditions, including diabetes, CHF, hypertension, and high-risk pregnancies. 3. The population health management component was targeted at all patients.
Target impacts on patients' outcomes	Over three years: • 20 percent reduction in 30-day hospital readmission rate for patients with CHF • 75 percent reduction in ED costs for patients with diabetes, CHF, and hypertension • 15 percent reduction in total costs for patients with patients with diabetes, CHF, and hypertension
Workforce development	1. Created six new care coordinator positions with HCIA funding; trained the four care coordinators hired in the first year of the program through a course offered at the Oregon Health and Sciences University 2. Provided ongoing informal training to existing MOAs 3. Hired one full-time nurse practitioner to provide same-day appointments
Location	Ketchikan and Craig, Alaska (remote island communities in southeastern Alaska)

Table II.1 *(continued)*

Impact evaluation	
Core design	Difference-in-differences with comparison group (not matched)
Treatment group	The treatment group consisted of Medicare FFS beneficiaries we attributed to the two PeaceHealth treatment clinics, based on receipt of a plurality of services over a 24-month period by PeaceHealth providers. For some outcomes, we limited the treatment group to the subgroup of beneficiaries with diabetes, CHF, or hypertension because the awardee targeted these groups in particular for some components of the intervention, and as a result, expected to have a greater impact on them.
Comparison group	The comparison group consisted of Medicare FFS beneficiaries we attributed to 57 non-FQHC practices in geographically isolated parts of southeastern and southern Alaska. For some outcomes, we limited the comparison group to a subgroup of beneficiaries with CHF, diabetes, or hypertension. Although we attempted to use statistical matching techniques, the small comparison pool and uniqueness of the treatment practices resulted in poor balance on key characteristics between the treatment and comparison groups. For this reason, we identified comparison practices by filtering those that met certain characteristics shared by the treatment practices but did not match on all prespecified matching variables.
Intervention component(s) included in impact evaluation	The impact evaluation estimated impacts of all four program components.
Extent to which the treatment group reflected the awardee's target population (for the component[s] evaluated)	**Low.** The program's target population included all patients assigned to participating providers' panel regardless of payer. In the last quarter, Medicare FFS beneficiaries accounted for 22 percent of the direct participants (defined as those receiving services from an HCIA-funded position). Other direct participants included patients dually eligible for Medicare and Medicaid (10 percent), nondual Medicaid beneficiaries (15 percent), patients with private health insurance (35 percent) uninsured patients (14 percent), and a combination of CHIP, TRICARE (Armed Forces), or Indian Health Services (4 percent).
Study outcomes, by domain	1. **Quality-of-care outcomes:** 30-day unplanned readmissions 2. **Service use:** All-cause inpatient admissions and outpatient ED visits 3. **Spending:** Medicare Part A and B spending 4. **Quality-of-care processes:** – Preventive care for diabetes – 14-day follow-up to hospitalization All outcomes, except diabetes preventive care, were estimated for (1) all attributed Medicare FFS beneficiaries; and (2) the subset of Medicare FFS beneficiaries with CHF, diabetes, or hypertension. The diabetes measure was calculated for all attributed Medicare FFS beneficiaries with diabetes.

Source: Review of PeaceHealth reports, including its original application, operational plan, and 15 quarterly narrative reports to the Centers for Medicare & Medicaid Services.

CHF = congestive heart failure; CHIP = Children's Health Insurance Program; ED = emergency department; FFS = fee-for-service; FQHC = federally qualified health center; HCIA = Health Care Innovation Award; MOA = medical office assistant; PCR = primary care redesign.

B. Overview of impact evaluation

To estimate the program's impacts on patients' outcomes, we compared changes in outcomes for Medicare fee-for-service (FFS) beneficiaries served by the two PeaceHealth clinics (treatment clinics) with changes in outcomes for beneficiaries served by 57 comparison practices over the same period, adjusting for any differences in characteristics between these two groups before the intervention began. Table II.1, bottom panel, summarizes our impact evaluation design.

We selected the 57 comparison practices for the evaluation from a pool of 239 potential comparison practices in Alaska. The comparison practices had to meet two criteria that were common to the two PeaceHealth treatment clinics: (1) they had to be located in a geographically isolated part of southern or southeastern of Alaska and (2) they could not be federally qualified health centers (FQHCs).

We estimated impacts on outcomes using Medicare FFS claims. We grouped the outcomes into four domains: (1) quality-of-care processes, (2) quality-of-care outcomes, (3) service use, and (4) Medicare spending. Across the 14 HCIA awardees in the primary care redesign (PCR) group, we designed our impact evaluations to identify promising interventions or intervention components—consistent with the evaluation goals from the Center for Medicare & Medicaid Innovation (CMMI) to find programs that could be scaled or retested in the future. Before conducting the analysis, we specified a series of primary tests, describing the evidence we would need to conclude that the program was effective; the awardee and CMMI reviewed and approved these tests. Each test specifies a population, outcome, period, expected direction of effect, and threshold that we consider substantively important. The purpose of these primary tests was to focus the impact evaluation on hypotheses that would provide the most robust evidence about program effectiveness. Because we were looking to identify promising interventions, rather than only those with unequivocally demonstrated success, we conducted one-sided statistical tests (that is, testing only for program benefits) and used a threshold for statistical significance of 0.10, which is not as strict as the conventional standard of 0.05.

We had originally planned to use the results from the primary tests and robustness checks to draw conclusions about program impacts in each of the four outcome domains. However, as described in Section V, after we began applying the decision rules, we determined that it was not possible to draw impact conclusions in three of the four domains. We drew conclusions only in the quality-of-care process domain. Nevertheless, we present the full set of quantitative results for transparency and to enable readers to judge the evidence for themselves.

Our impact evaluation design aligns reasonably well with PeaceHealth's HCIA program, meaning the evaluation should reflect the effects of all four intervention components among PeaceHealth's full HCIA target population. PeaceHealth expected the four intervention components—transitional care, short-term care management, longer-term case management, and population health management—to work in combination to affect outcomes for all patients served by treatment clinics. The awardee provided certain services to high-risk patients only—patients with CHF, diabetes, or hypertension—and we conducted some primary tests on this subpopulation of as a result. However, a limitation of the evaluation is that the treatment group is limited to Medicare FFS beneficiaries (including those who were dually eligible for Medicaid) who received care at the two treatment clinics, whereas the program served other patients as well (such as patients with private insurance, Medicaid beneficiaries, and uninsured patients). Our exclusion of non-Medicare patients limits the generalizability of the impact findings.

III. PROGRAM IMPLEMENTATION

In this section, we first provide a detailed description of PeaceHealth's HCIA-funded intervention, highlighting how it evolved over time and its theory of action. Second, we assess

the evidence on the extent to which PeaceHealth implemented the intervention as planned based on measures of program enrollment, service delivery, staffing, training, and timeliness. Third, we summarize the facilitators and barriers associated with implementation effectiveness.

We based our evaluation of PeaceHealth's program implementation on a review of the awardee's quarterly reports to CMMI and self-monitoring program metrics, telephone discussions and follow-up communications with program administrators, and information collected during site visit interviews with frontline staff conducted in May 2014 and April 2015. We did not verify the quality of the performance data reported by the awardee in its self-measurement and monitoring reports.

A. Program design and adaptation

1. Target population and patient identification, recruitment, and enrollment

Target population. PeaceHealth's HCIA-funded program target population varied by component. The transitional care component targeted all patients, regardless of health status or condition, discharged from PeaceHealth Ketchikan Medical Center hospital or its ED who were identified on the discharge list as having a primary care provider (PCP) at one of the two affiliated ambulatory clinics. Short-term care management and longer-term case management primarily focused on patients with diabetes, CHF, and hypertension who were part of a provider's panel at the two treatment clinics. During the third program year, the awardee began offering intervention services to women with high-risk pregnancies because of a lack of resources in the community for this population. The population health component was available to all patients assigned to a panel at the two treatment clinics.

Identification of patients for participation. Patient identification strategies for the PeaceHealth program varied by component. For the transitional care component, program staff used hospital discharge data to identify patients discharged in the previous 24 hours from the PeaceHealth Ketchikan Medical Center inpatient department and ED. Although program staff called all patients on the list, they stratified discharges into three groups (red, yellow, and green), indicating their risk of rehospitalization based on demographic and diagnostic information available from their medical records. They used the risk score to prioritize patients and to gain an understanding of their conditions and health status before placing the post-discharge telephone call. Stratification characteristics in the risk assessment model included the following:

- **Demographics.** Age and race
- **Prior hospital admission.** Two admissions in the past year, one admission in the past 180 days with a length of stay of three or more days, or one admission in the past 30 days
- **Diagnosis.** Diabetes, chronic obstructive pulmonary disease, end-stage renal disease, and mental health
- **Medications.** At least five active prescription medications
- **Charlson Comorbidity Index.** 10-year mortality rate based on several comorbid conditions

- **Receipt of charity care.** Eligibility for Bridge Assistance, a PeaceHealth financial assistance program
- **ED visit.** Co-occurring visit to the hospital ED

PeaceHealth developed the risk score, which it reported on the hospital discharge form, based on a risk-stratification analysis completed by Whatcom Alliance for Health Advancement, using claims and enrollment data from August 1, 2012, to July 31, 2013.

To identify patients for the short-term care management and longer-term case management components of the HCIA program, staff focused initially on the patients with diabetes and CHF, and later expanded the focus to patients with hypertension and high-risk pregnancies. Physicians also used their clinical judgement when deciding whether to refer a patient to case management for psychosocial issues. Patients referred to case management for psychosocial issues did not need to also have a co-occurring chronic condition. Patients were identified for the population health management component through the daily or weekly scrub-and-huddle process, in which patients with outstanding care needs were identified, contacted, and encouraged to schedule an appointment. They were also identified through the diabetic outreach report, which identified patients without diabetic follow-up appointments.

Recruitment and enrollment. PeaceHealth did not formally recruit or enroll patients into the program. Rather, program staff considered all patients in the target population for each component eligible for and enrolled in the program. Patients often knew they were referred for additional services or contacted about making follow-up visits, but they typically would not have been aware that they were enrolled in a particular program to receive special services.

2. Intervention components

Transitional care. Care coordinators provided general transitional care services for all patients discharged from the PeaceHealth Ketchikan Medical Center and intensive transitional care services for patients with CHF. Program staff identified patients listed on the PeaceHealth Ketchikan Medical Center hospital discharge form in the previous 24 hours. A transitional care protocol instructed program staff on how to conduct a follow-up telephone call after hospital or ED discharge. During the call, staff reviewed the discharge instructions with the patient, assessed the patient's need for additional support or education, and ensured the patient understood his or her medications. If care coordinators thought it was medically necessary, they would encourage patients to seek immediate follow-up care. Patients with CHF received additional telephone calls 14 and 28 days after discharge to ensure they made follow-up appointments with their PCPs and to review their medications, signs of fluid volume excess, and other clinical red flags. All other patients received only one post-discharge telephone call.

Short-term care management. Care coordinators provided short-term care management for clinic patients with a temporary medical or social hurdle. Three medical care coordinators (registered nurses) and one social worker care coordinator in the primary care department of the two outpatient clinics provided short-term care management services. (During the final year of its award, PeaceHealth added a fifth care coordinator in the obstetrics and gynecology clinic and

a sixth care coordinator in the administrative office focused exclusively on helping patients resolve their billing issues.) Short-term medical care management initially targeted patients with diabetes and CHF; later the awardee added hypertension and high-risk pregnancies. No formal documented protocols were in place for this component of the program. The medical care coordinators typically responded to the instructions provided by the clinicians and developed care plans based on an individual patient's needs and the specific skill set of the care coordinator. Care coordinators also provided education to patients, coordinated visits with specialists, and obtained free diabetic testing supplies for patients. Patients with diabetes, CHF, or hypertension who required more than one or two encounters with the medical care coordinator continued to receive longer-term case management, as deemed necessary by the clinician and the care coordinator. Most encounters with care coordinators occurred at the clinics or by telephone; some care coordinators occasionally visited clients in their homes, particularly to help patients use medical equipment.

Short-term social work care management services were also available to any patient with an identified psychosocial need (even those without diabetes, CHF, or hypertension); the social worker care coordinator provided these servcies. The social worker care coordinator worked to reduce patients' hurdles to ongoing medical care and healthy living by, for example, obtaining city bus passes for transportation to clinic appointments and providing patients with referrals and coordination to community and state resources.

Longer-term case management. Care coordinators provided longer-term case management services for patients requiring three or more encounters to effectively manage their chronic conditions. Program staff targeted patients with diabetes, CHF, and later those with hypertension and high-risk pregnancies. Care coordinators had flexibility in how they managed their panel of clinically high-risk and medically complex patients, as no formal documented protocols were in place for this component of the program. Care coordinators reported continuing to create, implement, and adjust care plans during ongoing monitoring of patients' chronic conditions. They also continued to provide education and help patients manage their diseases, assist with obtaining medical supplies, link patients to home and community resources, and coordinate visits with specialists. Care coordinators scheduled their own appointments or met with patients during PCP visits. Patients remained in longer-term case management as long as the patient and the care coordinator thought there was a need for these services.

Population health management. The population health management component centered on improving the scrub-and-huddle process and outreach to patients assigned to a panel to improve preventive care. Before the HCIA, the scrubbing and huddling was a loosely defined process, conducted on a case-by-case basis depending on a provider's preference and a patient's need, and completed by medical office assistants (MOAs) without formal training in the process. Under the HCIA, processes were standardized and used for all patients. Scrubbing generally involves reviewing a patient's medical records to identify outstanding care needs, and documenting outstanding needs on a patient's health maintenance worksheet. The huddling process generally involves a team meeting to review a patient's needs before a regularly scheduled visit. Some care teams did a less formal huddle because they frequently discussed patients' needs throughout the day; others went through a formal scrub-and-huddle at the

beginning of each week or day before seeing their patients. This component initially focused on overdue laboratory tests, mammograms, immunizations, colorectal cancer screenings, uncontrolled high blood pressure, and a positive tobacco status with no counseling. The program later expanded to other conditions and screenings. The population health management component included the hiring of a full-time nurse practitioner to be available for same-day appointments (which are newly available to support all patients).

3. Theory of action

Based on a review of PeaceHealth's program activities and goals, we developed a theory of action to depict the mechanisms through which program administrators expected the program to improve the outcomes we selected for the impact evaluation (Table II.1 lists these outcomes). PeaceHealth expected that its HCIA-funded intervention would improve outcomes for Medicare patients through three pathways.

Transitional care pathway to improved outcomes. By providing telephone calls within 24 hours after discharge, care coordinators ensured that patients had their medications reviewed and reconciled, understood their medications and discharge instructions, recognized early warning signs, and were not experiencing pain or discomfort. Through these follow-up calls, the program expected to help patients manage their conditions, avoid preventable complications, and improve patients' adherence to post-discharge treatment regimens, and—through timely visits to PCPs—make any necessary adjustments to treatment plans. Improvements in self-management and adherence to discharge instructions, in turn, were expected to increase the rate of 14-day post-discharge follow-up visits, reduce the likelihood of exacerbations of the illness that prompted the initial hospital stay or ED visit, and reduce the likelihood of needing to return to the ED or the hospital. This should have reduced 30-day readmissions rates and, in doing so, reduced overall medical costs.

Care management pathway to improved outcomes. By identifying, coordinating, and case managing a panel of clinically high-risk and medically complex patients, care coordinators ensured that patients had the information and skills they needed to self-manage their conditions, knew how to access the home and community resources needed to supplement their medical care, obtained the medical testing supplies needed to monitor their conditions, and remained compliant with their overall treatment regimens. The availability of same-day appointments also helped ensure patients could access care when they needed it in an appropriate setting. These short- and long-term care management services (and same-day appointments) should have reduced the risk of preventable complications from chronic conditions and improved the overall health of patients. Better management of chronic conditions and improved overall health should, in turn, have reduced ED visits and inpatient admissions, all-cause rehospitalizations, and overall costs of care.

Text Box III.1. Example from PeaceHealth illustrating the program's theory of action

"Man with uncontrolled diabetes has been working with a care coordinator (a registered nurse and a social worker); came to us with DM-related blindness and now has access to regular appointments with PCP, had a trip to Seattle to see an eye specialist, free diabetes supplies and insulin, and better outcomes. Trip to Seattle was coordinated by the social worker and included charity airfare, no-cost-to-him lodging and transportation, as well as assistance to be approved for charity care at the Harborview eye clinic."

Source: PeaceHealth's 11th quarterly report to the Centers for Medicare & Medicaid Services.
DM = diabetes mellitus; PCP = primary care provider.

Population health pathway to improve outcomes. Increasing outreach to paneled patients with outstanding health care needs and introducing the availability of same-day appointments should have led to an increase in the number of patients who visited the clinic on a regular basis for necessary medical screenings, tests, and check-ups, including the four recommended tests for diabetes (lipid profile, hemoglobin A1c test, dilated eye exam, and nephropathy screening). Increasing the availability of same-day appointments should also have increased the number of patients with an unexpected health event who sought treatment in the clinic rather than in the ED. This, in turn, should have (1) increased the number of patients whose chronic conditions were under control and (2) reduced complications associated with their conditions. More regular (and same-day) primary care appointments and better management of chronic conditions should have reduced ED visits, hospitalizations, and inpatient readmissions. If the savings associated with averted ED and inpatient services exceeded the costs associated with the additional primary and preventive services, the population management component should also have reduced the total costs of care.

4. Intervention staff and workforce development

Table III.1 provides key details about the staff hired or trained for the HCIA-funded intervention. Program services were administered primarily through two positions: care coordinators and MOAs. The award created and paid for three full-time positions for medical care coordinators and one full-time position for a social work care coordinator. The program later added two new care coordinators focused on patients with high-risk pregnancies and on helping patients resolve their billing issues. In addition, PeaceHealth hired a nurse practitioner, clinical educator, nurse, and program coordinator to support the program's objectives, all paid for with HCIA funding. The nurse practitioner was hired to increase access for same-day appointments. The clinical educator was hired to facilitate training among the MOAs, and the program coordinator was hired to administer the overall program. Although neither the nurse practioner nor the clinical educator provided intervention services, they supported the intervention by training staff and expanding access to clinical care.

PeaceHealth also used the HCIA funding to provide training for care coordinators and MOAs. All medical care coordinators completed a multiday training course on care coordination at the Oregon Health and Sciences University. The course included training on motivational interviewing, chronic illness, nursing assessment planning, and communication, among other

aspects of care coordination. The award also provided funding to increase the competencies of existing MOAs and help the care team members work at the top of their licensure. As part of the program's population health management component, MOAs were trained to execute the scrub-and-huddle process with the provider team. Training for the MOAs focused on six competencies: (1) point-of-care testing, (2) facilitating patients' visits, (3) infection control, (4) medication administration, (5) exam room preparation, and (6) patients' safety. The clinical educator also conducted monthly brown bag sessions for MOAs on topics such as immunizations and cardiovascular care and sent them daily facts on other educational topics, such as the definition of cholesterol.

Table III.1. Key details about intervention staff

Program component	Staff	Staff responsibilities	Adaptations?
Transitional care	Care coordinators	Care coordinators provided transitional care services through a single telephone call 24 hours after discharge to all patients discharged from the PeaceHealth Ketchikan Medical Center inpatient and EDs and two additional telephone calls 14 and 28 days after discharge for patients with CHF.	No
Short-term care management	Care coordinators	These care coordinators identified, coordinated, and case managed a panel of clinically high-risk patients with complex care needsand provided oversight for caregivers who managed patients at moderate or low risk, including creating care plans, providing education, or offering diabetic supplies. The social work care coordinator helped patients apply for Medicaid, understand their insurance, coordinate transportation, connect them with other mental health resources, and assist them with other psychosocial issues. Short-term case management was defined as only one or two encounters to address a patient's issues.	The program later expanded by adding two new care coordinators focused on patients with high-risk pregnancies and on helping patients resolve their billing issues.
Longer-term case management	Care coordinators	These care coordinators worked with patients with CHF, diabetes, and hypertension and their families to create, implement, and oversee a care plan that met identified needs and incorporated services that improved outcomes. Care coordinators worked within an interdisciplinary care team to assess a patient's physical, cognitive, emotional, and behavioral status, as well as social and financial support systems. Longer-term case management was defined as three or more encounters to address a patient's issues. Patients remained in longer-term case management as long as the patient and the care coordinator thought there was a need for these services.	No
Population health	Care coordinators	These care coordinators were responsible for making sure that patients whose records indicated they were out of compliance for certain health tests came in for regular check-ups. Care coordinators also provided assistance scheduling appointments in the clinic.	Care coordinators assumed responsibility for reviewing the medical charts for patients with diabetes and identified those who needed to schedule an appointment.

Table III.1 *(continued)*

Program component	Staff	Staff responsibilities	Adaptations?
Population health	MOAs	PeaceHealth provided funding to increase the competencies of existing MOAs and help the care team members work at the top of their licensure. MOAs previously were responsible for preparing patients for their visit with their provider. As part of the program's population health management component, MOAs were trained to execute the scrub-and-huddle process with the provider team.	MOAs were initially responsible for scrubbing all patients' charts. In an effort to improve process measures for diabetic patients, care coordinators took over scrubbing for diabetic patients during the second year of the program.
Population health	Nurse practitioner	The nurse practitioner was a new position created so that PeaceHealth could expand the availability of same-day appointments.	No
Population health	Clinical educator	The clinical educator provided on-site, ongoing, informal trainings to the MOAs until a local MOA certificate program could be established.	No

Sources: Interviews and document review.

CHF = congestive heart failure; ED = emergency department; MOA = medical office assistant.

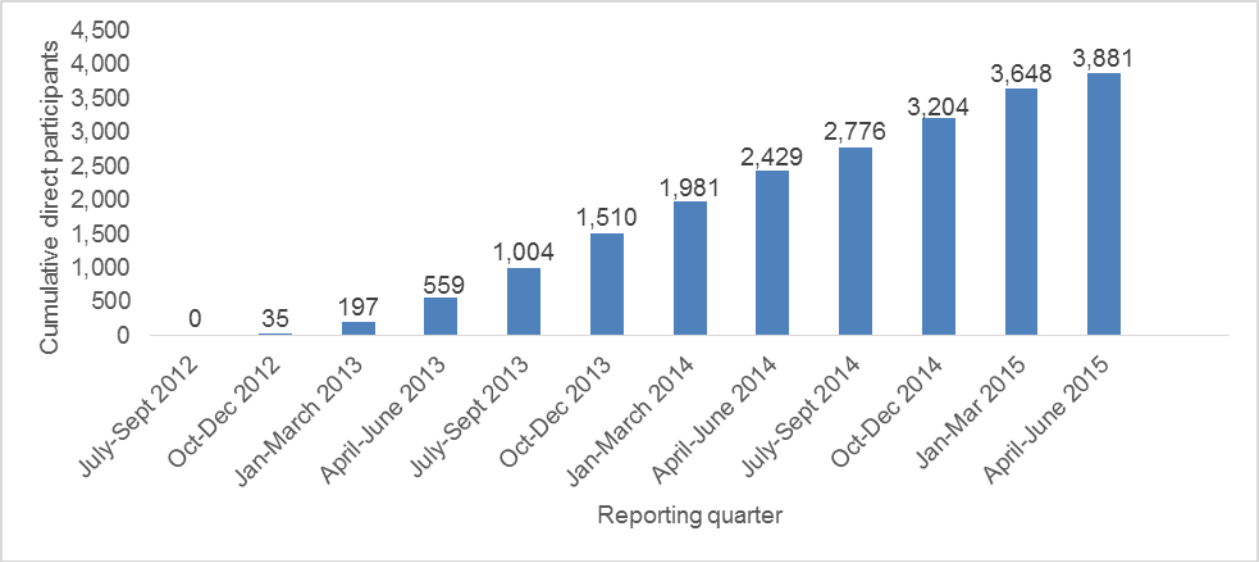
B. Implementation effectiveness

In this section, we examine the evidence on implementation effectiveness—that is, we analyze measures of the intervention delivered and, when possible, compare those measures with the services the awardee intended to deliver. We assess the evidence on implementation effectiveness in five areas: (1) program enrollment, (2) service delivery, (3) staffing, (4) training, and (5) implementation timeliness. To conduct this analysis, we used data from interviews with program administrators and frontline staff, self-reported metrics included in PeaceHealth's self-monitoring and measurement reports, and data from PeaceHealth on patients it enrolled in care coordination.

1. Program enrollment

The awardee projected directly serving 3,500 unique participants through Year 3. PeaceHealth served 3,881 patients, achieving 111 percent of its target for the three-year award (Figure III.1). These include patients who received HCIA-funded services through at least one of the intervention's four components—transitional care, care management (short- or longer-term), or population health management. In the last program quarter, 22 percent of these patients were Medicare FFS beneficiaries, 10 percent were dually eligible for Medicare and Medicaid, 15 percent were nondual Medicaid beneficiaries, 35 percent had private health insurance,14 percent were uninsured patients, and 4 percent were covered under the Children's Health Insurance Program (CHIP), TRICARE (Armed Forces), or Indian Health Services. PeaceHealth steadily increased the number of new patients who received intervention services by about 400 patients each quarter.

Figure III.1. Cumulative number of unique direct participants, by program quarter



Source: Review of PeaceHealth's program reports as of June 2015.

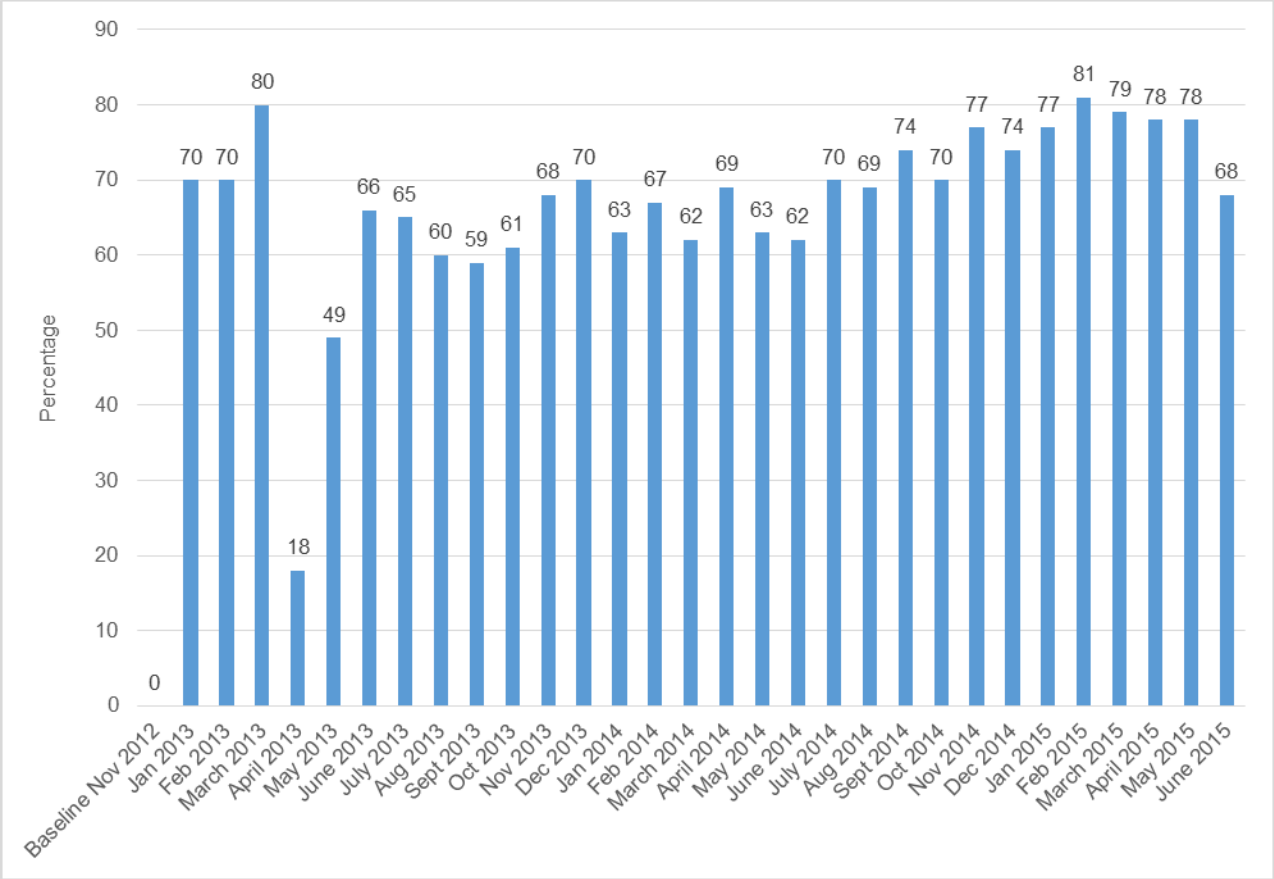
2. Service-related measures

Incomplete data and lack of clear specifications on service-related metrics make it difficult to assess implementation performance in this domain. The only metric available in PeaceHealth's measurement and monitoring report pertinent to the implementation evaluation relates to the transitional care component. According to PeaceHealth's final program quarter measurement and monitoring report, care coordinators contacted 60 to 80 percent of all patients discharged from the Ketchikan Medical Center inpatient department and ED during the program, depending on the quarter. PeaceHealth reported that no patients received transitional care services following discharge before the intervention was introduced (Figure III.2). In addition, the awardee reported providing 12,599 direct participant encounters by the end of the program, representing on average 3.25 encounters per unique patient served. During the final quarter of program activities (April through June 2015), the awardee provided 1,385 direct participant encounters. Of these, 86 percent were via telephone and the remaining 14 percent were in-person visits.

3. Staffing measures

PeaceHealth hired a total of 9.5 full-time-equivalent (FTE) staff with HCIA funding, nearly meeting its goal of 11 FTE staff. Of these new hires, six full-time staff were medical or social work care coordinators (Section III.A.4). The other staff include a full-time program manager, a full-time clinical educator, a full-time nurse practitioner, and a half-time nurse. By September 2013 (about one year after the award), PeaceHealth had hired all of the care coordinators needed to implement the intervention as planned.

Figure III.2. Percentage of patients followed up with by a care coordinator after discharge, by month



Source: Analysis of PeaceHealth's HCIA quarterly reports, December 2012 through July 2015.

Note: PeaceHealth reported 0 percent at baseline (November 2012), and then reported approximate estimates in January through March 2013.

HCIA = Health Care Innovation Award.

4. Perceived effectiveness of training

To learn more about the effectiveness of training in meeting the goals of the program, we administered the HCIA Primary Care Redesign Trainee Survey from January to March 2015 (roughly two years after the start of implementation). We mailed the survey to the 26 staff members who PeaceHealth had identified as having received training. However, after administering the survey, we learned that the list included none of the 6 care coordinators hired under HCIA funding (and who were essential to program operations); thus, they did not receive the survey. Of the 26 MOAs, 18 responded to the survey and 3 were ineligible because they responded no to one or both of the screener questions, resulting in a 65 percent response rate. The omission of the care coordinators from the survey limits our ability assess the effect of training on the program.

Of the 13 MOAs who reported receiving training, all rated the training they received as good or excellent. When asked about the effect of the training on a multiple dimensions related to care,

only five respondents answered the question. All five said the training had a positive effect on the quality, efficiency, patient-centeredness, and equity of care they provided to patients. All five also reported that the training had a positive effect on their ability to work in teams, particularly in (1) explaining information about patients' care to patients and their families in lay terms, (2) relaying relevant information to the care team, (3) working with diverse set of patients, and (4) helping accessing the care the patients needed. However, CMS does not allow us to report results with fewer than 11 responses and, given the low number of respondents for these questions, it is impossible to draw conclusions about the perceived effect of training on care delivery.

5. Program timeline

Program administrators reported initial delays in implementing the program, and the timing varied by component. It took more than a year for all program components to become operational, only after the awardee had hired staff, transitioned to the new electronic health record (EHR) system (unrelated to the program), and developed service delivery protocols. The program became partially operational by January 2013 (six months after the award). However, it took another year to fully implement the transitional care component of the project, according to program administrators. Similarly, the full implementation of the short-term care management and longer-term case management components was delayed when one of the medical care coordinators resigned. By September 2013 (about one year after award), all of the program staff were in place and providing intervention services.

C. Summary of facilitators of and barriers to implementation

Several factors facilitated implementation of PeaceHealth's HCIA-funded intervention, but others hindered it. We described those factors in detail in the second annual report (Gilman et al. 2015). Here, we summarize key facilitators and barriers, along with any new information since the second annual report that supports those facilitators or barriers (Table III.2).

Four factors were particularly important in facilitating program implementation, and one factor was a barrier. First, flexibility in delivering care coordination services enabled frontline staff to adapt the care coordination model to meet an individual patient's needs and to address providers' preferences. Second, the ability to adapt the overall design of the program facilitated implementation because PeaceHealth could focus on specific populations and realign staff roles as the program learned more about workflows that were effective for particular groups of patients. Third, PeaceHealth's investment of resources facilitated implementation because PeaceHealth invested in areas itbelieved would have the biggest impact, including hiring new staff and training MOAs and care coordinators. Fourth, the intervention was consistent with PeaceHealth's mission and its overall approach to care. The alignment of goals between the program and the corporate office facilitated implementation, despite the potential loss of hospital revenue from lower inpatient and ED service use. Finally, staff engagement and buy-in were initially barriers to implementation because providers did not understand the purpose or role of care coordinators and were not sure how to use them. The awardee overcame these barriers as providers learned more about the purpose of the intervention and the role of the care coordinators and began to see the benefits of their services.

Table III.2. Summary of key facilitators of and barriers to the implementation of PeaceHealth's HCIA-funded initiative

Item	Description based on findings in the second annual report	Additional supporting data not available in the second annual report, if applicable
	Facilitators	
Frontline users' flexibility in implementing the program	PeaceHealth's program gave frontline staff flexibility in implementing the care coordination model to meet an individual patient's needs and to address providers' preferences. Program administrators recognized the need for flexibility in administering protocols for care management and case management and were willing to allow the team to innovate.	PeaceHealth created a "Care Coordinator Training Resource Document" with resources such as motivational interviewing and depression screening. The document included protocols for transitional care and outreach calls. Care management activities did not have a protocol, and care coordinators felt protocols would not be helpful because they had to tailor their interventions to meet individual's specific needs.
Adaptation of the program to meet patients' and providers' needs	PeaceHealth adapted its program to focus on specific populations and realign staff roles as the program learned more about workflows that were effective for particular groups of patients. The program initially focused its transitional care component on all discharges from the PeaceHealth Medical Center. It later narrowed its focus to include only those patients with CHF and diabetes because program leadership believed those patients could benefit the most from transitional care services. Later, the program shifted again to provide transitional care to all patients on a PeaceHealth panel who were discharged from the local hospital. The short-term care management component also originally focused on smoking cessation, but shifted its focus to patients with diabetes, and then added CHF, hypertension, and high-risk pregnancies. Program administrators, working with providers, determined that these high-risk conditions were expensive for the PeaceHealth system, and there was a need in the community for these services.	
Dedicating resources to support the program	Program leaders invested and focused HCIA resources toward areas they believed could have the biggest impact, using HCIA funding to hire new staff to provide care coordination and social work services, and to train MOAs and care coordinators.	
Culture of the organization	PeaceHealth's corporate culture was a factor in deciding to apply and helped facilitate the program's implementation. Program staff said the intervention was consistent with PeaceHealth's mission and its overall approach to care. The alignment of goals between the program and the corporate office facilitated implementation, despite the potential loss of hospital revenue from lower inpatient and ED service use.	In its closeout report, PeaceHealth wrote, "PeaceHealth's system of 10 hospital and 45 medical group sites have made a corporate commitment to Population Health, influenced by the benefits derived from the Ketchikan CMS Innovation program. In addition, the Ketchikan CMS Innovation program's learnings, best practices and practice models have become an essential element in developing our System's strategic approach and plan for Population Health and Coordinated Care."
	Barriers	
Engagement of and buy-in from staff	Initially, staff engagement and buy-in was a barrier to implementation. For example, providers at first did not understand the purpose or role of care coordinators and were not sure how to use them. Some providers also hesitated to buy into the scrub-and-huddle process. Providers found that the appointment could be inefficient if the necessary chart preparation was not conducted beforehand or they had to rework the chart preparation themselves if the MOA was not adequately trained and the scrub-and-huddle was not conducted properly.	

Table III.2 *(continued)*

Note: We reviewed four CFIR domains associated with implementation experience: (1) program characteristics, (2) implementation process, (3) internal factors, and (4) external environment. Implementation research suggested that barriers and facilitators within these domains were important determinants of implementation effectiveness.

CFIR = Consolidated Framework for Implementation Research; CHF = congestive heart failure; CMS = Centers for Medicare & Medicaid Services; ED = emergency department; HCIA = Health Care Innovation Award; MOA = medical office assistant.

D. Conclusions about the extent to which the program, as implemented, reflects the core design

PeaceHealth was unable to provide a clear description of its service protocols for short- and longer-term case management, making it difficult to evaluate whether the awardee delivered those services in the intended way. A high degree of adaptation and flexibility also led to a lack of standardization in the service protocols, with different care coordinators providing different services in different ways, depending on patients' needs and the unique skills and interests of the care coordinator.

Despite these limitations, we conclude that the implementation of the program was at least minimally successful based on the following observations:

- All care coordinators completed an intensive multiday training course at the Oregon Health and Science University.
- MOAs participated in a series of informal training modules designed to increase their core competencies.
- Most MOAs who responded to the trainee survey rated the training good or excellent in terms of improving their ability to perform the responsibilities of their role.
- Of the 11 planned-for FTE positions, 10 were filled.
- PeaceHealth experienced minimal staff turnover after the two care coordinators left in the first year.
- The program overcame initial challenges to obtaining buy-in from clinicians, with clinicians reporting that they increased the number of patients they referred to care coordination.
- The program enrolled 111 percent of its 3,500 targeted direct program participants.
- The awardee reported a high number of direct patient encounters, totaling 12,599 over the full three years of the program and representing on average 3.25 encounters per unique patient served.
- During the third year of the program, care coordinators consistently made follow-up calls to more than 70 percent of all patients discharged from the Ketchikan Medical Center hospital or ED, depending on the quarter.

Despite the initial implementation delays and a lack of clear and standardized protocols for the care and case management components, the program was fully staffed and already providing intervention services to an above-target number of patients by the end of its second year. The

awardee continued to operate at this level throughout the third year of the award. We therefore have no reason to assume we should not see effects on patients by the end of the program.

IV. CLINICIANS' PERCEPTIONS OF THE PROGRAM'S EFFECTS ON THE CARE THEY PROVIDED TO PATIENTS

This section describes the available evidence on the extent to which PeaceHealth's intervention had its intended effects on changing providers' behavior as a way to achieve desired impacts on patients' outcomes. As described in Section III.A.3, the program's theory of action expected that program services would be delivered mainly through the care coordinators. However, providers were important in referring patients, working collaboratively with the care coordinators and MOAs to address patients' needs, and participating in the scrub-and-huddle process. We use data from two rounds of the HCIA Primary Care Redesign Clinician Survey to assess changes in providers' behavior and to conclude whether the anticipated changes occurred. The survey relies on self-reported responses and reflects clinicians' perceptions of the program, rather than quantitatively measuring direct program effects on the care they provided.

A. Clinician survey

Survey methods. We administered the clinician survey in two rounds (fall 2014 and summer 2015). We sent the survey to the eight providers working in either of the two clinics, and received a response rate of 75 percent on both rounds. Because our survey frame was so small and CMS does not allow us to report results with fewer than 11 responses, we are unable to report the actual number of responses.

Survey results. Almost all providers reported that they were aware of the program. Of those who reported they were at least somewhat familiar with the program, all or most said that that the program had a positive impact on quality, their ability to respond in a timely way to patients' needs, patients' safety, and the patient-centeredness of care they provided. However, only about half of the respondents thought the program improved the efficiency or equity of care, or the information available for clinical decision making. Providers' perceptions of the impact of the program on quality of care and patients' safety increased between the two rounds, but the number of respondents was too small to draw firm conclusions.

B. Conclusions about the program's effects on clinicians' behavior

It is impossible to draw conclusions based on the small number of clinicians and respondents. However, based on the information available from the clinician survey, we have no reason to assume that the program did not have its intended effect on the care that most PCPs provided. Nearly all PCPs surveyed were aware of the program and, on dimensions of care related to care coordination, the program appeared to have its intended effects for most respondents.

V. PROGRAM IMPACTS ON PATIENTS' OUTCOMES

This section of the report presents results for the quantitative analysis that aimed to draw conclusion, based on available evidence, about the impacts of PeaceHealth's HCIA program on

patients' outcomes in four domains: quality-of-care processes, quality-of-care outcomes, service use, and spending. We first describe the methods for estimating impacts (Section V.A) and then the characteristics of the two HCIA treatment clinics at the start of the intervention (Section V.B). We next describe the similarities and differences between treatment clinics and comparison practices at the start of the intervention, which is important for limiting potential bias in impact estimates (Section V.C). Finally, in Section V.D, we describe the quantitative impact estimates, their plausibility given implementation findings, and why we were unable to draw conclusions in three of the four study domains. The findings in this report update the impact results from the second annual report for PeaceHealth (Gilman et al. 2015), extending the outcome period by 6 months and adding new outcomes.

A. Methods

1. Overview

We estimated program impacts on patients' outcomes as the differences in outcomes for Medicare FFS patients served by the two treatment clinics and those served by 57 comparison practices, adjusting for any differences in outcomes between these groups during the year before the intervention began. We prespecified primary tests, describing the evidence we would need to conclude that the program was effective, and the awardee and CMMI reviewed these. Each test specified a population, outcome, time period, expected direction of effect, and threshold that we count as substantively important. The purpose of these primary tests was to focus the impact evaluation on hypotheses that would provide the most robust evidence about program effectiveness.

We had originally planned to use the results from the primary and secondary tests (robustness checks) to draw conclusions about program impacts in each of the four outcome domains. However, as described in Section V, after applying the decision rules, we determined that it was not possible to draw impact conclusions in three of the four domains. We drew conclusions only in the quality-of-care process domain. Nevertheless, we present the full set of quantitative results for transparency and to enable the readers to judge the evidence for themselves.

2. Treatment group definition

The treatment group consisted of Medicare FFS beneficiaries served by the two treatment clinics in four baseline quarters before the intervention began (January 1, 2012, to December 31, 2012) and eight intervention quarters (January 1, 2013, to June 30, 2015).

We constructed the treatment group in three steps.

1. First, we attributed beneficiaries to clinics using the same decision rule that CMMI uses for its Comprehensive Primary Care (CPC) initiative. Specifically, in each baseline and intervention month, we attributed beneficiaries to the primary care clinic whose providers (physicians, nurse practitioners, or physician assistants) delivered the plurality of primary care services in the past 24 months. If there was a tie, we attributed beneficiaries to the

clinic they visited most recently. PeaceHealth provided data on which providers worked in the treatment practices and when.

2. Second, in each baseline and intervention period, we assigned each patient to the first treatment clinic he or she was attributed to in that period, and continued to assign him or her to that clinic for all quarters in the period. This assignment rule—which is distinct from the attribution method—ensures that, during the intervention period, patients did not exit the treatment group solely because the intervention succeeded in reducing their service use (including visits at treatment clinics). The definition for the baseline period corresponds to that of the intervention period so that, across the two periods, interpretation of the population changes over time should be comparable.
3. Third, we applied additional restrictions to refine the analysis sample in each quarter. A patient assigned to a treatment clinic in a quarter was included in the analysis sample for that quarter if he or she (1) had observable outcomes for at least one day in the quarter; and (2) lived in Alaska, for at least one day of the quarter. For this sample, outcomes were observable for beneficiaries who were enrolled in Medicare FFS (Part A and B), were alive, and had Medicare as their primary payer.

In addition to this full treatment sample, we defined a subset of high-risk patients who had diabetes, CHF, and/or hypertension. This high-risk subgroup enabled us to examine whether any observed effects were concentrated among high-risk members. This would be expected from the program's theory of action (Section III.A.3), given that PeaceHealth targeted some of its services at beneficiaries with one or more of these three conditions. We identified this high-risk subgroup in each quarter by applying Chronic Condition Warehouse algorithms for these conditions to claims in the 12 to 36 months (depending on the condition) before the start of the baseline or intervention periods. (We did not examine high-risk pregnancies, although this was also a focus of the PeaceHealth program, because pregnancy is rare among Medicare beneficiaries, most of whom are elderly.) As with assignment to the treatment group, a Medicare FFS beneficiary who had previously been identified as having one of these conditions in either period will *remain* a member of this subgroup for the rest of the relevant period (baseline or intervention).

3. Comparison group definition

The comparison group consisted of Medicare FFS beneficiaries we assigned to 57 comparison practices in each of the baseline and intervention quarters. We identified the comparison practices in data we obtained on 239 potential comparison practices in Alaska from SK&A, a health care data vendor. We limited comparison practices to those in geographically isolated parts of southeastern and southern parts of Alaska, because the PeaceHealth practices are also geographically isolated. We excluded FQHCs from the comparison group because neither of the treatment clinics is an FQHC. These restrictions left us with 57 remaining practices that we used as the comparison group.

We assigned Medicare FFS beneficiaries to the comparison practices in each baseline and intervention quarter using the same rules we used for the intervention group. Further, we defined the high-risk subgroup of comparison members with diabetes, CHF, or hypertension in each quarter using the same rules as for the treatment group.

Although we attempted to use propensity-score matching among the 57 potential comparison practices to form a smaller comparison group that would be very similar to each of the two treatment clinics, we were unable to do so. No comparison practices met our minimum criteria of being similar enough to each of the treatment clinics along all of the variables we considered important, including practice size and service use among assigned beneficiaries. After discussions with CMMI, we concluded that the best approach was to have the comparison group include beneficiaries at all 57 practices in the comparison pool, rather than a poorly matched subset. Section V.C compares the treatment and comparison groups on key variables.

4. Construction of outcomes and covariates

We used Medicare claims from January 1, 2009, to June 30, 2015, for beneficiaries assigned to the treatment and comparison practices to develop two types of variables: (1) outcomes, defined for each person in each baseline or intervention quarter; and (2) covariates, which describe a beneficiary's characteristics at the start of the baseline and intervention periods and are used in the regression models for estimating impacts to adjust for beneficiaries' characteristics before the period began. We used covariates defined at the start of each period, without updating them each quarter, to avoid controlling in each intervention quarter for previous quarters' program effects, as this would bias the effect estimates away from detecting true impacts. Appendix 1 provides details on the methods we used to construct these variables.

Outcomes. For each beneficiary, we calculated six outcomes that we grouped into four domains:

1. Domain: Quality-of-care processes
    a. Diabetes quality-of-care composite (binary variable for each beneficiary); calculated as whether a beneficiary with diabetes had had all four recommended tests—lipid profile, hemoglobin A1c test, dilated eye exam, and nephropathy screening—during the previous 12 months
    b. Ambulatory-care follow-up visit within 14 days of a hospital discharge (binary variable for each beneficiary); calculated as whether all of an individual's discharges in a quarter were followed by an ambulatory visit with a primary care or specialist physician within 14 days of the discharge
2. Domain: Quality-of-care outcomes
    a. Number of inpatient admissions followed by an unplanned readmission within 30 days (number /quarter)
3. Domain: Service use
    a. All-cause inpatient admissions (number/quarter)
    b. Outpatient ED visit rate (number/quarter); outpatient ED visits are defined as ED visits or observational stays that do not end in a hospital admission
4. Domain: Spending
    a. Total Medicare Part A and B spending (dollars/month)

Four of these outcomes—all but the two quality-of-care process measures—are outcomes that CMMI has specified as core for the evaluations of all HCIA programs. Our definition of the readmission measure, however, differs from CMMI's standard definition. CMMI typically defines readmissions as the proportion of inpatient admissions that end in an unplanned readmission. Instead, we analyzed impacts on the *number* of these unplanned readmissions across all beneficiaries per quarter, because this enables us to look at the total impact on readmissions across the treatment group, rather than readmissions contingent on an inpatient admission. We made this decision, in consultation with CMMI, because the intervention might also affect the number and type of admissions.

All outcomes are quarter-specific—meaning that we calculated them for each baseline and intervention quarter separately—except for the quality-of-care process measure for diabetes. Because this measure assesses whether a beneficiary received recommended preventive care services over a year-long period, we calculated this measure over full years rather than quarters: for example, over the baseline year (that is, the period corresponding to the four baseline quarters), over the first year of the intervention period (corresponding to the first four intervention quarters), and so on. We avoided calculating these measures for overlapping periods, meaning that no measurement year included services provided in another measurement year.

Finally, we defined all outcomes for all treatment and comparison group members, except for the two measures of quality-of-care processes. We calculated the measure of 14-day follow-up after discharge among only those patients with at least one hospital discharge in the relevant quarter. We calculated the diabetes composite measure among beneficiaries ages 18 to 75 with diabetes at the beginning of the period (baseline or intervention period).

Covariates. The covariates included (1) whether a beneficiary had each of 10 chronic conditions (including Alzheimer's and related dementia, cancer, CHF, chronic kidney disease, chronic obstructive pulmonary disease, diabetes, depression, hypertension, ischemic heart disease, and stroke). As noted earlier, we identified beneficiaries with these conditions by applying Chronic Condition Warehouse algorithms to claims in the 12 to 36 months (depending on the condition) before the start of the baseline or intervention periods; (2) dual Medicare and Medicaid enrollment; (3) Hierarchical Condition Category (HCC) score, which is a continuous score that CMS developed to predict a beneficiary's future Medicare spending; (4) demographics (age, gender, and race identified as Native American or Alaska Native versus all other races); and (5) original reason for Medicare entitlement (old age, disability, or end-stage renal disease).

5. Regression model

We used a regression model to implement the difference-in-differences design for estimating impacts. For each outcome, the model estimates the relationship between the outcome and a series of predictor variables, assuming that each of the predictor variables has a linear (additive) relationship with the outcome. The predictor variables include the patient-level covariates (defined in Section V.A.4); whether the patient is assigned to a treatment or a comparison practice; an indicator for each practice (which accounts for differences between practices in their patients' outcomes at baseline); indicators for each post-intervention quarter (or, for the diabetes

measure, for the final post-intervention quarter of the year-long measurement period); and an interaction of a beneficiary's treatment status with each post-intervention quarter (or, for the diabetes measure, the final post-intervention quarter of the year-long measurement period).

The estimated relationship between the interaction term and the outcome in a given quarter is the impact estimate for that quarter (or, for the diabetes measure, for the year ending with that quarter). It measures the average difference between outcomes for beneficiaries assigned to the treatment and comparison practices during that period, subtracting out any differences between these groups during the four baseline quarters. By providing separate impact estimates for each intervention quarter (or year, for the diabetes measure), the model enables the program's impacts to change over time. We can also test impacts over discrete sets of quarters or years, which is needed to implement the primary tests discussed in the next section. Finally, the model quantifies the uncertainty in the impact estimates, allowing for statistical tests that determine whether observed differences in outcomes between the treatment and comparison groups are likely due to chance. The model uses robust standard errors to account for clustering of outcomes across quarters for the same beneficiary and a dummy variable for each practice (fixed effects) to account for clustering of outcomes for beneficiaries assigned to the same practice. Appendix 2 provides details on the regression methods.

6. Primary tests

Table V.1 shows the primary tests for PeaceHealth, by domain. Each test specifies a population, outcome, time period, expected direction of effect, and threshold that we count as substantively important. The purpose of these primary tests is to focus the impact evaluation on hypotheses that will provide the most robust evidence about program effectiveness (see Appendix 3 for details and a description of how we selected each test). Both the awardee and CMMI had an opportunity to comment on the primary tests.

Our rationale for selecting these primary tests is as follows:

- **Outcomes.** PeaceHealth's central goal was to reduce hospitalizations, ED visits, and Medicare Part A and B spending, so our primary tests address these three outcomes. In addition, the primary tests address one quality-of-care outcome the intervention is expected to affect: 30-day unplanned hospital readmissions. Finally, we include two quality-of-care process measures that, based on PeaceHealth's theory of action (Section III.A.3), we think the program could improve: (1) a composite measure for whether a beneficiary with diabetes received all four recommended processes of care during the year (lipid profile, hemoglobin A1c test, dilated eye exam, and nephropathy screening); and (2) receipt of a follow-up ambulatory care visit with a primary care or specialist provider within 14 days of hospital discharge.
- **Time period.** PeaceHealth expected program impacts to grow over the first two years before stabilizing in the third year. However, given that the projected impacts were based on the assumption that the intervention would begin soon after the award did, it might be more realistic to expect that the program effects would be delayed by about a year, given the implementation delays. As a result, we conducted the primary tests on outcomes in the

second and third intervention years (January 2014 to June 2015, corresponding to quarters I5 through I10), excluding the first intervention year.

- **Population.** PeaceHealth's impacts should have concentrated among its high-risk population —specifically those with diabetes, CHF, and/or hypertension—but this population was small compared with the full population served by the HCIA-funded program. Because there are trade-offs between analyzing the high-risk subpopulation (for which expected effects are large but the sample size is moderate) and analyzing the entire Medicare FFS population (which is more representative of the program population served but with smaller anticipated effects), we assess both in our primary tests. For the diabetes quality-of-care process measures, we limit the population to beneficiaries ages 18 to 75 with diabetes. For the 14-day follow-up measure, we limit the sample in each quarter to those who had at least one qualifying hospitalization during the quarter for which we could observe whether the person had a 14-day follow-up visit.
- **Direction (sign) of the impact estimate.** For the quality-of-care process measures, we expect the impact estimate to be positive, signaling an increase in the percentage of people receiving recommended care. For all other outcomes, we expect the impact estimates to be negative, indicating a reduction in service use or overall expenditures.
- **Substantive thresholds.** Some impact estimates could be large enough to be policy-relevant (to CMMI and other stakeholders) even if they are not statistically significant; for this reason, we have prespecified thresholds for what we call substantive importance. We express the threshold as a percentage change from the counterfactual—that is, the outcomes that beneficiaries in the treatment group would have had if they had not received the HCIA-funded intervention. The thresholds we use—15 percent for high-risk beneficiaries and 5 percent for all beneficiaries are extrapolated from the literature (Peikes et al. 2011; Rosenthal et al. 2016). We use thresholds from the literature rather than PeaceHealth's target impacts (Table II.1) because PeaceHealth's target impacts were defined for outcomes or populations that did not perfectly align with how we define outcomes and populations in the primary tests.

Table V.1. Specification of the primary tests for PeaceHealth Ketchikan Medical Center

Domain (number of tests in the domain)[a]	Outcome (units)	Time period for impacts (controlling for baseline differences)[b]	Population	Substantive threshold (expected direction of effect)[c]
Quality-of-care processes (3)	Received all four recommended diabetes processes of care in the year (binary [yes or no]/beneficiary/year)	Final intervention year (corresponding to intervention quarters 7 through 10)[d]	Medicare FFS beneficiaries with diabetes and ages 18 to 75 assigned to treatment practices	15.0% (+)
	All inpatient admissions within a quarter were followed by an ambulatory care visit with a primary care or specialist provider within 14 days (binary [yes or no]/beneficiary/year)	Average over intervention quarters 5 through 10[d]	Medicare FFS beneficiaries with CHF, diabetes, and/or hypertension assigned to treatment clinics and who had at least one hospital stay in the quarter	15.0% (+)
	All inpatient admissions within a quarter were followed by an ambulatory care visit with a primary care or specialist provider within 14 days (binary [yes or no]/beneficiary/year)	Average over intervention quarters 5 through 10[d]	All Medicare FFS beneficiaries assigned to treatment clinics and who had at least one hospital stay in the quarter	15.0% (+)
Quality-of-care outcomes (2)	30-day unplanned hospital readmissions (#/beneficiary/quarter)	Average over intervention quarters 5 through 10[d]	Medicare FFS beneficiaries with CHF, diabetes, and/or hypertension assigned to treatment clinics	15.0% (-)
	30-day unplanned hospital readmissions (#/beneficiary/quarter)	Average over intervention quarters 5 through 10[d]	All Medicare FFS beneficiaries assigned to treatment clinics	-5.0% (-)
Service use (4)	All-cause inpatient admissions (#/beneficiary/quarter)	Average over intervention quarters 5 through 10[d]	Medicare FFS beneficiaries with CHF, diabetes, and/or hypertension assigned to treatment clinics	-15.0% (-)
	All-cause inpatient admissions (#/beneficiary/quarter)	Average over intervention quarters 5 through 10[d]	All Medicare FFS beneficiaries assigned to treatment clinics	-5.0% (-)
	Outpatient ED visit rate (#/beneficiary/quarter)	Average over intervention quarters 5 through 10[d]	Medicare FFS beneficiaries with CHF, diabetes, and/or hypertension assigned to treatment clinics	-15.0% (-)

Table V.1 *(continued)*

Domain (number of tests in the domain)[a]	Outcome (units)	Time period for impacts (controlling for baseline differences)[b]	Population	Substantive threshold (expected direction of effect)[c]
	Outpatient ED visit rate (#/beneficiary/quarter)	Average over intervention quarters 5 through 10[d]	All Medicare FFS beneficiaries assigned to treatment clinics	-5.0% (-)
Spending (2)	Medicare Part A and B spending ($/beneficiary/month)	Average over intervention quarters 5 through 10[d]	Medicare FFS beneficiaries with CHF, diabetes, and/or hypertension assigned to treatment clinics	-15.0% (-)
	Medicare Part A and B spending ($/beneficiary/month)	Average over intervention quarters 5 through 10[d]	All Medicare FFS beneficiaries assigned to treatment clinics	-5.0% (-)

[a] We will adjust the *p*-values from the primary test results for the multiple comparisons made within each domain, but not across domains.

[b] The regressions models will control for differences in outcomes between the treatment and comparison groups in the baseline period.

[c] The substantive threshold is the impact as a percentage of the counterfactual. The counterfactual is the outcome the treatment group would have had in the absence of the HCIA-funded intervention.

[d] To implement the primary tests for all outcomes except the quality-of-care process measure related to diabetes, we take the average of the regression-adjusted estimates for intervention quarters 5 through 10. Because the diabetes measure is defined over a year rather than every quarter, we assess impacts on that outcome only in the final year of the intervention.

CHF = congestive heart failure; ED = emergency department; FFS = fee-for-service; HCIA = Health Care Innovation Award.

7. Secondary tests (robustness checks)

We also conducted secondary quantitative tests to help corroborate the findings from the primary tests. This is important because some of the differences observed between the treatment and comparison groups for the primary tests could result from the non-experimental design or random fluctuations in the data. We will have greater confidence in the primary results if they are generally consistent with the expected broader pattern of results.

We conducted three sets of secondary tests for PeaceHealth.

1. First, we repeated the tests designed for the primary tests, but for outcomes during the first four intervention quarters, the period before PeaceHealth fully implemented its program. Because we expect program impacts to grow over time, with few or no impacts in the first year of a practice's participation in the program, the following pattern would be highly consistent with an effective program—little to no measured effects in the first four quarters and larger effects in quarters 5 through 10. In contrast, if we found very large differences in outcomes (favorable or unfavorable) in the first 12 intervention months, this could suggest a limitation in the comparison group, not true program impacts.
2. Second, we reestimated impacts on admissions and spending only among beneficiaries assigned to the treatment and comparison groups by the start of the period, either baseline or intervention. This restriction prevents addition to the intervention sample over time. It is possible that differences in sample addition between the treatment and comparison groups could bias the impact results to some degree if the sample members added over time differ from earlier sample members (for example, they are younger and healthier); this could create differences in mean outcomes between the treatment and comparison groups that are unrelated to the HCIA program. We have explored this possibility because, as we will describe in Section V.D.1, the rate of net sample growth during the baseline period was slightly higher for the treatment group than for the comparison group.

8. Synthesizing evidence to draw conclusions

Within each domain, we planned to draw one of five conclusions about program effectiveness based on the primary test results, the results of secondary tests, and the plausibility of those findings given the implementation evidence:

1. Statistically significant favorable effect (the highest level of evidence)
2. Substantively important (but not statistically significant) favorable effect
3. Substantively important (but not statistically significant) unfavorable effect
4. No substantively large effect
5. Indeterminate effect

We could not conclude that a program had a statistically significant unfavorable effect because, in consultation with CMMI, we decided to use one-sided statistical tests (which do not test for

evidence of unfavorable effects). We used one-sided tests to increase the probability that, if a program truly did have impacts, we would be able to detect them.

Appendix 3 describes our decision rules for each of the five possible conclusions. In short, we concluded that a program had a statistically significant favorable effect in a domain if (1) at least one primary test result in the domain was favorable and statistically significant, after adjusting the statistical tests to account for multiple tests (if applicable) within a domain; or (2) the average impact estimate across all primary tests in the domain was favorable and statistically significant. In both cases, we also had to determine that the primary test results were plausible given the results of the secondary tests and implementation evidence. We concluded that a program had a substantively important favorable effect if the average impact estimate in the domain was substantively important but not statistically significant, and if the result was plausible given the secondary tests and implementation evidence. In contrast, if the average impact estimate was unfavorable (opposite the hypothesized direction), larger than the substantive threshold, and unfavorable effects were plausible given the other evidence, we concluded the program had a substantively important unfavorable effect. If the tests in a domain did not meet any of these criteria, we instead used the following rules. First, if the tests for at least one outcome in the domain (or all outcomes in the domain together) had sufficient statistical power to detect an impact of the size of the substantive threshold with at least 75 percent probability, we concluded there was not a substantively large effect because we are reasonably confident that we would have detected such an effect had there been one. Second, if the power was not sufficient (less than 75 percent) to detect this type of impact, we concluded the impact in the domain was indeterminate. Indeterminate means either that the program truly did not have effects that were substantively large, or that it did, but our statistical tests were unable to detect them. Finally, if the results for the primary tests in a domain were not plausible given the implementation evidence or the secondary, corroborating tests, we did not draw any conclusions about program impacts in that domain.

B. Characteristics of the treatment group at baseline

This section describes the characteristics of the treatment group at the start of the intervention (January 1, 2013). The second column of Table V.2. also shows this information. (Table V.2 serves a second purpose—to show the extent of similarity between the treatment and comparison practices at the start of the intervention—which we describe in Section V.C.) For benchmarking purposes, the last column shows the values of relevant variables for the national Medicare population, when available.

The means for the two treatment clinics are weighted, such that each clinic receives a weight equal to the number of beneficiaries assigned to that clinic in the baseline period. We did this to correspond to how much weight each practice receives in the impact regressions. The observations in those regressions are beneficiary-quarters and each beneficiary receives the same weight. Therefore, the smaller treatment practice (in Prince of Wales, N = 78 at the start of the intervention period) implicitly receives much less weight than the larger treatment practice (in Ketchikan, N = 767 at the start of the intervention period).

Characteristics of the clinics overall. At the start of the intervention, the 2 treatment clinics had a weighted average of 9.7 PCPs (one clinic had 2 whereas the other had 10). Both clinics are owned by the awardee, PeaceHealth Ketchikan Medical Center. Although both clinics are in remote towns in Alaska, the clinic in Ketchikan is in a zip code that the U.S. Census Bureau classifies as urban and, as a result, 97 percent of the treatment beneficiaries received care in a clinic classified to be in an urban area. Neither of the clinics is in a county that has a shortage of health professionals, as designated by the federal Health Resources and Services Administration.

Characteristics of the clinics' Medicare FFS beneficiaries. The characteristics of all Medicare FFS beneficiaries assigned to the treatment clinics during the baseline period (January 1, 2012, through December 31, 2012) were similar to the nationwide FFS averages by some measures but very different by others. The HCC risk score for the treatment group of 1.06 was very close to the national average (1.0). Hospital admission rates (71/1,000 beneficiaries/quarter) and Medicare Part A and B spending ($861/beneficiary/month) were close to the national average, but the outpatient ED visit rate (205/1,000 beneficiaries/quarter) was almost twice as high as the national average.

The high-risk beneficiaries in the treatment group had somewhat higher health care utilization and spending during the baseline period than the full treatment group. They had 25 percent more all-cause inpatient admissions, 12 percent more outpatient ED visits, and 25 percent higher Medicare spending compared with the full treatment group; differences would be even larger if we compared the high-risk group with its complement (that is, members of the treatment group who were not a part of the high-risk group).

The percentage of patients receiving recommended processes of care was low. Only 18 percent of patients with diabetes received all four recommended process of care for diabetes. Further, only 33 percent of patients received an ambulatory care visit within 14 days of hospital discharge. Although we do not have national benchmarks constructed the same way as our process-of-care variables, we can compare the levels to those seen for TransforMED's HCIA population (which spanned 90 practices in 15 states) and to those for CMMI's CPC initiative, which spanned 502 practices in seven regions and used the same specifications for defining the process variables). For the diabetes process-of-care measure, the estimates are 18, 39, and 31 percent for PeaceHealth, TransforMED, and CPC (Taylor et al. 2015), respectively. For the 14-day follow-up measure, the estimates are 33, 58, and 64 percent for PeaceHealth, TransforMED, and CPC (Taylor et al. 2015), respectively. Therefore, the rates seen for PeaceHealth for both the diabetes and 14-day follow-up measures were well below the averages for CPC and TransforMED.

Table V.2. Characteristics of treatment and comparison practices before the intervention start date (January 1, 2013)

Characteristic of practice	Treatment clinics (N = 2)	Comparison practices (N = 57)	Absolute difference[a]	Standardized difference[b]	Medicare FFS national average
Characteristics of the practices overall					
Practice owned by hospital or health system (%)	100	33	67	0.86	n.a.
Number of PCPs	9.7	4.3	5.44	0.76	n.a.
Characteristics of practices' locations					
Located in an urban zip code (%)	96.7	53.5	43.2	0.65	n.a.
Located in a health professionals shortage area (primary care) (2011) (%)	0.0	40.5	-40.5	-0.69	n.a.
Characteristics of all beneficiaries attributed to practices during the baseline year (January 1, 2012, to December 31, 2012)					
Number of beneficiaries	745	354	392	0.43	n.a.
HCC risk score	1.06	1.00	0.06	0.06	1.0
Receipt of an ambulatory care visit within 14 days of all hospital discharges in the quarter, among those with at least one discharge in the quarter (%)	0.33	0.49	-0.16	-0.34	NA
Unplanned readmissions (#/beneficiary/quarter)	4.33	2.41	1.93	0.36	NA
All-cause inpatient admissions (#/1,000 beneficiairies/quarter)	71	69	2.06	0.03	74[c]
Outpatient ED visit rate (#/1,000 beneficiairies /quarter)	205	142	63	0.36	105[d]
Medicare Part A and B spending ($/beneficiary/month)	861	901	-40	-0.04	860[e]
Disability as original reason for Medicare entitlement (%)	0.24	0.22	0.02	0.09	16.7[f]
Dually eligible for Medicare and Medicaid (%)	0.26	0.23	0.04	0.17	21.7[g]
Age (years)	70.41	70.75	-0.34	-0.01	71[h]
Female (%)	0.52	0.53	0.00	-0.01	54.7[f]
Characteristics of high-risk beneficiaries attributed to practices during the baseline year (January 1, 2012, to December 31, 2012)					
Number of high-risk beneficiaries	464	217	247	0.43	n.a.
Receipt of an ambulatory care visit within 14 days of all hospital discharges in the quarter, among those with at least one discharge in the quarter (%)	0.33	0.49	0.16	0.32	NA
Unplanned readmissions (#/beneficiary/quarter)	4.10	2.00	2.10	0.49	
All-cause inpatient admissions (#/1,000 beneficiairies/quarter)	89	85	4	0.05	NA
Outpatient ED visit rate (#/1,000 beneficiairies/quarter)	230	168	62	0.29	NA
Medicare Part A and B spending ($/beneficiary/month)	1,072	1,079	-22	-0.01	NA

Table V.2 *(continued)*

Characteristic of practice	Treatment clinics (N = 2)	Comparison practices (N = 57)	Absolute difference[a]	Standardized difference[b]	Medicare FFS national average
Receipt of all four recommended diabetes process of care measures, among those with diabetes ages 18 to 75 (%)	0.18	0.20	-0.02	-0.12	NA[i]

Sources: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at CMS. Zip code data merged from the Five-Year American Community Survey ZIP Code Characteristics (2012) and county data merged from the Area Health Resources File (2011).

Notes: Each practice gets a weight equal to the number of beneneficiaries assigned to the practice.
Absolute differences might not be exact due to rounding.

[a] The absolute difference is the difference in means between the treatment and comparison groups.

[b] The standardized difference is the difference in means between the treatment and comparison groups divided by the standard deviation of the variable, which is pooled across the treatment and selected comparison groups.

[c] Health Indicators Warehouse (2014b).

[d] Gerhardt et al. (2014).

[e] Boards of Trustees (2013).

[f] Chronic Conditions Data Warehouse (2014, Table A.1).

[g] Health Indicators Warehouse (2014c).

[h] Health Indicators Warehouse (2014a).

[i] Although a national benchmark defining the diabetes process-of-care measure in the same way is not available, 31percent of beneficiaries in the Comprehensive Primary Care Initiatative practices (spread across seven regions) and 39 percent of beneficiaries in the TransforMed practices (spread across 15) received all four recommended process of care measures.

CMS = Centers for Medicare & Medicaid Services; ED = emergency department; FFS = fee-for-service; HCC = Hierarchical Condition Category; PCP = primary care provider.

NA = not available.

n.a. = not applicable.

C. Similarities between treatment and comparison groups at baseline

Assessing the similarities and differences between the treatment and comparison groups at the start of the intervention is critical for assessing the quasi-experimental evaluation design and interpreting its results. Similarities increase the credibility of a key assumption underlying difference-in-differences models—that the change over time in outcomes for the comparison group is the same change that would have happened for the treatment group, had the treatment clinics not received the intervention. As discussed in Section V.A.3, we were unable to create a matched comparison group of practices that resembled the treatment clinics on all important, measurable characteristics. As a result, the comparison group comprised all 57 potential comparison practices we identified in geographically isolated areas similar to PeaceHealth in the Southeast and Southern parts of Alaska.

The third column of Table V.2 shows differences in weighted mean characteristics at the start of the intervention of the 57 comparison practices. Each practice is weighted by its number of assigned beneficiaries in the baseline period. The comparison practices are smaller on average than the treatment clinics, both by the mean number of PCPs (4 versus 10) and mean number of beneficiaries (354 versus 745). Although a hospital (PeaceHealth) owned both treatment clinics,

a hospital or health system owned only 33 percent of the comparison practices. The comparison practices were less likely to be in zip codes classified as urban by the U.S. Census Bureau and more likely to be located in counties identified as health shortage areas. These differences in practice characteristics are all outside of our target of 0.25 standardized differences (the 0.25 target is an industry standard; for example, see Institute of Education Sciences 2014). Our impact analysis, described in Section V.D.3, controlled for any time-invariant influences these differences had on outcomes across practices through practice-level fixed effects.

Despite differences in practices' characteristics, beneficiaries attributed to the comparison practices were similar to those attributed to treatment practices along a number of dimensions. They had similar demographic, health, and eligibility characteristics, such as age, gender, HCC risk scores, percentage with dual Medicare and Medicaid coverage, and percentage with disability as reason for original Medicare entitlement. The groups also had very similar mean rates of admissions, Medicare Part A and B spending, and receipt of recommended diabetes care. However, the comparison group had substantially lower unplanned readmissions rates and outpatient ED visit rates and substantially higher rates of post-discharge ambulatory care visits.

D. Beneficiaries' outcomes and intervention impacts

In this section, we first present sample sizes and mean outcomes, by quarter, for the treatment and comparison groups. These mean outcomes provide context for understanding the difference-in-differences estimates that follow; however, the differences in mean outcomes are not regression-adjusted and not impact estimates by themselves. Next, we present the results of the primary tests, by domain. Then, we present the results of the secondary tests (robustness checks) and assess whether the primary test results are plausible given the secondary test results and the implementation evidence. We end with conclusions in one domain and a discussion about why we cannot draw conclusions in other domains.

1. Sample sizes

The sample sizes for impact estimation differed depending on the outcome. We present sample sizes by domain.

Quality-of-care processes (Table V.3)

- The **14-day follow-up measure** was defined among Medicare FFS beneficiaries who have at least one hospital stay in the quarter. For the treatment group, the sample size ranged from 35 to 63 beneficiaries across the baseline and intervention quarters. For the comparison group, the sample ranged from 541 to 651 across the baseline and intervention quarters. Among high-risk beneficiaries, the sample size ranged from 26 to 39 beneficiaries in the treatment group and 394 to 461 beneficiaries in the comparison group.
- The **diabetes preventive care composite measure** was defined among Medicare FFS beneficiaries ages 18 to 75 with diabetes. The sample size ranged from 111 to 129 for the treatment group and from 1,212 to 1,380 for the comparison group across the baseline year and each of the two intervention years. This population accounted for 10 to 15 percent of the total Medicare FFS sample in the treatment and comparison groups, depending on the year.

Table V.3. Unadjusted mean outcomes (quality-of-care processes) observed among select Medicare FFS beneficiaries, by treatment status and quarter

Period	Quarter	Number of Medicare FFS beneficiaries (panels)		Mean outcomes		
		T	C (not weighted)	T	C	Difference (%)
Among beneficiaries with at least one inpatient admission in the quarter, the percentage of beneficiaries whose inpatient admissions in the quarter were followed by an ambulatory care visit with a primary care or specialist provider within 14 days of discharge (%/quarter)						
Baseline	B1	48 (1)	585 (55)	50.0	50.8	-0.8 (-1.5%)
	B2	47 (1)	549 (56)	48.9	50.3	-1.3 (-2.7%)
	B2	44 (2)	539 (50)	13.6	48.2	-34.6 (-71.7%)
	B4	45 (2)	574 (55)	20.0	44.6	-24.6 (-55.2%)
Intervention	I1	63 (2)	541 (52)	34.9	50.1	-15.2 (-30.3%)
	I2	43 (2)	556 (53)	16.3	52.7	-36.4 (-69.1%)
	I3	35 (2)	568 (51)	17.1	53.7	-36.6 (-68.1%)
	I4	48 (2)	592 (55)	27.1	48.5	-21.4 (-44.1%)
	I5	43 (2)	611 (53)	25.6	51.9	-26.3 (-50.7%)
	I6	61 (2)	583 (54)	24.6	52.7	-28.1 (-53.3%)
	I7	49 (2)	573 (53)	22.4	56.5	-34.1 (-60.3%)
	I8	53 (2)	642 (54)	35.8	53.3	-17.4 (-32.7%)
	I9	50 (2)	651 (53)	18.0	58.2	-40.2 (-69.1%)
	I10	60 (2)	645 (51)	30.0	56.0	-26.0 (-46.4%)

Table V.3 *(continued)*

Period	Quarter	Number of Medicare FFS beneficiaries (panels) T	Number of Medicare FFS beneficiaries (panels) C (not weighted)	Mean outcomes T	Mean outcomes C	Mean outcomes Difference (%)
Among those with CHF, diabetes, or hypertension and at least one inpatient admission in the quarter, the percentage of beneficiaries whose inpatient admissions in the quarter were followed by an ambulatory care visit with a primary care or specialist provider within 14 days of discharge (%/quarter)						
Baseline	B1	39 (1)	457 (53)	43.6	52.3	-8.7 (-16.7%)
	B2	37 (1)	414 (55)	48.6	49.5	-0.9 (-1.8%)
	B2	35 (1)	394 (47)	14.3	47.7	-33.4 (-70.1%)
	B4	33 (2)	414 (51)	18.2	45.4	-27.2 (-60.0%)
Intervention	I1	52 (2)	452 (51)	32.7	51.1	-18.4 (-36.0%)
	I2	33 (2)	420 (50)	18.2	55.5	-37.3 (-67.2%)
	I3	26 (2)	418 (51)	23.1	56.5	-33.4 (-59.1%)
	I4	28 (2)	432 (53)	21.4	50.2	-28.8 (-57.3%)
	I5	30 (2)	423 (52)	23.3	53.4	-30.1 (-56.3%)
	I6	37 (2)	402 (52)	21.6	55.2	-33.6 (-60.8%)
	I7	32 (2)	400 (51)	28.1	58.3	-30.1 (-51.7%)
	I8	39 (2)	449 (48)	38.5	56.3	-17.9 (-31.7%)
	I9	29 (1)	461 (50)	13.8	59.2	-45.4 (-76.7%)
	I10	37 (2)	429 (50)	29.7	57.8	-28.1 (-48.6%)

Table V.3 *(continued)*

Period	Quarter	Number of Medicare FFS beneficiaries (panels)		Mean outcomes		
		T	C (not weighted)	T	C	Difference (%)
Among those with diabetes and ages 18 to 75, the percentage who received all four recommended diabetes processes of care in the year (%/year)						
Baseline	B1–B4[a]	111 (2)	1,374 (55)	18.0	21.5	-3.5 (-16.1%)
Intervention	I1–I4[a]	129 (2)	1,380 (54)	22.5	22.0	0.5 (2.1%)
	I7–I10[a]	114 (2)	1,212 (55)	31.6	22.8	8.8 (38.7%)

Sources: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at the Centers for Medicare & Medicaid Services.

Notes: The baseline quarters are measured relative to the start of the baseline period on January 1, 2012. For example, the first baseline quarter (B1) ran from January 1, 2012, to March 31, 2012. The intervention quarters are measured relative to the start of the intervention period on January 1, 2013. For example, the first intervention quarter (I1) ran from January 1, 2013, to March 31, 2013. In each period (baseline or intervention), the treatment group each quarter included beneficiaries assigned to a treatment practice by the start of the quarter and who met other sample criteria—that is, they were enrolled in FFS Medicare; lived in Alaska; and met any restrictions of the measure with respect to age, chronic conditions, or recent hospital admissions. In addition, for the measures of diabetes, we required beneficiaries to be observable for the full 12 months covered by the measure. In each period (baseline or intervention), the comparison group included all beneficiaries assigned to a comparison practice by the start of the quarter and who met the other sample criteria. See text for details.

The difference between the treatment and comparison groups in a quarter is calculated by subtracting the mean outcome for the comparison group from the mean outcome for the treatment group. The percentage difference equals that difference divided by the mean outcome for the comparison group.

[a] The diabetes quality-of-care process measure was calculated over year-long periods, corresponding to the baseline and intervention quarters shown in the table.

B = baseline C = control; CHF = congestive heart failure; FFS = fee-for-service; I = intervention; T = treatment.

Quality-of-care outcomes, service use, and spending. The sample sizes for all outcomes in these three domains were the same. In the first baseline quarter (B1), the treatment group included 718 beneficiaries assigned to the two treatment clinics and the comparison group included 9,247 beneficiaries assigned to the 57 comparison practices (Table V.4). The sample sizes increased during the four baseline quarters (by 24 and 17 percent from B1 to B4 for the treatment and comparison groups, respectively). This net increase indicates that sample addition (due to beneficiaries being newly attributed to the treatment or comparison practices) exceeded sample attrition (due to beneficiaries dying, switching from FFS Medicare to managed care, or moving out of the state). In the first intervention quarter (I1), the treatment group included 845 beneficiaries and the comparison group included 9,658 beneficiaries. The sample sizes increased during the intervention period (by 30 and 22 percent from I1 to I10 for the treatment and comparison groups, respectively). The high-risk subgroup was roughly half the sample size of the full treatment and comparison populations (Table V.5).

2. Mean outcomes for the treatment and comparison groups, by domain and quarter

Quality-of-care processes. For the treatment group, the share of beneficiaries with a hospital stay who had an ambulatory care visit within 14 days of discharge was highest—at roughly 50 percent—during the first two baseline quarters. In the second two baseline quarters, this share fell to 13.6 to 20 percent and during the intervention period this share ranged from 16.3 to 35.8 percent. For the comparison group, this share was 44.6 to 50.8 percent during the baseline period and modestly higher during the intervention period, ranging from 48.5 to 58.2 percent. These patterns were similar among the high-risk group of treatment and comparison beneficiaries.

During the baseline year, 18.0 percent of treatment and 21.5 percent of comparison beneficiaries ages 18 to 75 with diabetes received all four recommended processes of care. This percentage was 31.6 for the treatment group and 22.8 for the comparison group in the second program year.

Quality-of-care outcomes. Among the treatment group, 30-day unplanned readmissions rates fluctuated during both the baseline and intervention periods from 2.8 to 10.7 readmissions per 1,000 beneficiaries per quarter, without any trend. Among the high-risk subgroup, the rate fluctuated from 2.1 to 17.6, also without any trend. The rate for the comparison group also fluctuated but within a smaller range, from to 6.1 to 10.1 readmissions for the full sample and 8.6 to 14.3 readmissions per 1,000 beneficiaries per quarter for the high-risk subgroup.

Service use. During the baseline period, both the treatment and comparison groups had all-cause inpatient admissions rates of about 77.0 per 1,000 beneficiaries in the first quarter. Rates were slightly lower in subsequent baseline quarters, ranging from 68.1 to 71.1. During the intervention period, the treatment group rate fluctuated from quarter to quarter, ranging from 54.5 to 81.7 without any consistent trend of increasing or decreasing rates. The comparison group rate also varied without any consistent trend, but over a narrower range from 60.7 to 71.9. For the high-risk subgroups of both treatment and comparison groups, the rate was higher, ranging from 76.4 to 102.0 admissions per 1,000 beneficiaries.

Table V.4. Unadjusted mean outcomes (quality-of-care outcomes, service use, and spending) measured for all Medicare FFS beneficiaries, by treatment status and quarter

	Number of Medicare FFS beneficiaries (practices)		30-day unplanned hospital readmissions (#/1,000/ quarter)			All-cause inpatient admissions (#/1,000/quarter)			Outpatient ED visit rate (#/1,000/quarter)			Medicare Part A and B spending ($/month)		
Q	T	C (not wgt)	T	C	Diff (%)	T	C	Diff (%)	T	C	Diff (%)	T	C	Diff (%)
Baseline period (January 1, 2012, to December 31, 2012)														
B1	718 (2)	9,247 (57)	2.8	7.8	-5.0 (-64.2%)	76.6	77.3	-0.7 (-0.9%)	235.4	140.5	94.8 (67.5%)	$865	$943	$-78 (-8.3%)
B2	778 (2)	9,853 (57)	3.9	7.6	-3.8 (-49.3%)	68.1	67.7	0.4 (0.6%)	194.1	137.1	57.0 (41.6%)	$796	$884	$-87 (-9.9%)
B3	844 (2)	10,423 (57)	10.7	7.5	3.2 (42.5%)	71.1	62.7	8.3 (13.3%)	228.7	147.8	80.9 (54.7%)	$813	$861	$-47 (-5.5%)
B4	887 (2)	10,862 (57)	5.6	7.9	-2.3 (-28.8%)	69.9	67.1	2.8 (4.1%)	164.6	144.0	20.6 (14.3%)	$986	$912	$74 (8.1%)
Intervention period (January 1, 2013, to June 30, 2015)														
I1	845 (2)	9,658 (57)	7.1	9.4	-2.3 (-24.6%)	81.7	71.9	9.8 (13.6%)	185.8	145.4	40.4 (27.8%)	$1,021	$943	$78 (8.2%)
I2	893 (2)	10,196 (57)	9.0	9.2	-0.3 (-2.8%)	63.8	67.9	-4.0 (-6.0%)	208.3	143.8	64.5 (44.8%)	$748	$917	$-169 (-18.4%)
I3	936 (2)	10,566 (57)	10.7	7.6	3.1 (41.1%)	54.5	64.8	-10.3 (-16.0%)	205.1	139.2	65.9 (47.3%)	$819	$967	$-148 (-15.3%)
I4	964 (2)	10,854 (57)	7.3	10.1	-2.9 (-28.3%)	59.1	70.2	-11.1 (-15.8%)	198.1	135.6	62.6 (46.1%)	$777	$956	$-178 (-18.7%)
I5	996 (2)	11,028 (57)	5.0	8.9	-3.9 (-43.5%)	57.2	69.9	-12.7 (-18.1%)	211.8	130.5	81.4 (62.4%)	$893	$960	$-66 (-6.9%)
I6	1,002 (2)	11,211 (57)	9.0	7.6	1.4 (18.5%)	71.9	64.8	7.1 (11.0%)	241.5	138.0	103.5 (75.0%)	$1,165	$995	$170 (17.1%)
I7	1,030 (2)	11,416 (57)	6.8	6.1	0.7 (10.8%)	70.9	60.7	10.2 (16.8%)	201.9	139.3	62.6 (45.0%)	$1,035	$994	$41 (4.1%)
I8	1,063 (2)	11,612 (57)	3.8	7.9	-4.2 (-52.5%)	54.6	68.5	-14.0 (-20.4%)	163.7	135.2	28.4 (21.0%)	$845	$951	$-106 (-11.1%)
I9	1,080 (2)	11,701 (57)	5.6	7.4	-1.9 (-25.3%)	60.2	67.9	-7.8 (-11.4%)	210.2	139.2	71.0 (51.0%)	$890	$974	$-84 (-8.6%)
I10	1,101 (2)	11,807 (57)	10.0	8.6	1.4 (15.6%)	68.1	68.1	0.0 (0.0%)	214.4	147.7	66.7 (45.2%)	$1,013	$969	$44 (4.5%)

Table V.4 *(continued)*

Sources: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at the Centers for Medicare & Medicaid Services.

Notes: The baseline quarters are measured relative to the start of the baseline period on January 1, 2012. For example, the first baseline quarter (B1) ran from January 1, 2012, to March 31, 2012. The intervention quarters are measured relative to the start of the intervention period on January 1, 2013. For example, the first intervention quarter (I1) ran from January 1, 2013, to March 31, 2013. In each period (baseline or intervention), the treatment group each quarter included all beneficiaries assigned to a treatment practice by the start of the quarter and who met other sample criteria—that is, they were enrolled in FFS Medicare, lived in Alaska and were observable. In each period, the comparison group included all beneficiaries who were assigned to a comparison practice by the start of the quarter and who met the other sample criteria. See text for details.

The difference between the treatment and comparison groups in a quarter is calculated by subtracting the mean outcome for the comparison group from the mean outcome for the treatment group. The percentage difference equals that difference divided by the mean outcome for the comparison group.

B = baseline; C = comparison; Diff = difference; ED = emergency department; FFS = fee-for-service; I = intervention; Q = quarter; T = treatment.

Table V.5. Sample sizes and unadjusted mean outcomes (quality of care outcomes, service use, and spending) for Medicare FFS beneficiaries with CHF, diabetes, or hypertension, in the treatment and comparison groups for PeaceHealth, by quarter

	Number of Medicare FFS beneficiaries (practices)		30-day unplanned hospital readmissions (#/1,000/quarter)			All-cause inpatient admissions (#/1,000/quarter)			Outpatient ED visit rate (#/1,000/quarter)			Medicare Part A and B spending ($/month)		
Q	T	C (not wgt)	T	C	Diff (%)	T	C	Diff (%)	T	C	Diff (%)	T	C	Diff (%)
Baseline period (January 1, 2012, to December 31, 2012)														
B1	470 (1)	5,971 (57)	2.1	10.6	-8.4 (-79.8%)	91.5	94.5	-3.0 (-3.1%)	263.8	157.3	106.6 (67.8%)	$925	$1,111	$-186 (-16.7%)
B2	492 (2)	6,180 (57)	6.1	9.5	-3.4 (-36.1%)	89.4	82.5	6.9 (8.4%)	203.3	154.5	48.7 (31.5%)	$1,025	$1,053	$-28 (-2.7%)
B3	516 (2)	6,333 (57)	17.4	10.1	7.3 (72.6%)	85.3	76.4	8.8 (11.6%)	269.4	167.9	101.5 (60.5%)	$1,023	$1,004	$19 (1.9%)
B4	522 (2)	6,459 (57)	9.6	11.8	-2.2 (-18.6%)	92.0	84.7	7.3 (8.6%)	182.0	166.1	15.9 (9.6%)	$1,319	$1,103	$216 (19.6%)
Intervention period (January 1, 2013, to June 30, 2015)														
I1	549 (2)	6,320 (57)	7.3	11.9	-4.6 (-38.6%)	102.0	91.8	10.2 (11.1%)	213.1	165.0	48.1 (29.1%)	$1,306	$1,167	$139 (11.9%)
I2	556 (2)	6,474 (57)	10.8	11.7	-0.9 (-8.1%)	79.1	82.0	-2.9 (-3.5%)	214.0	160.3	53.7 (33.5%)	$902	$1,088	$-185 (-17.1%)
I3	567 (2)	6,510 (57)	17.6	11.2	6.4 (57.3%)	65.3	81.0	-15.7 (-19.4%)	227.5	156.8	70.7 (45.1%)	$948	$1,158	$-210 (-18.1%)
I4	566 (2)	6,522 (57)	5.3	14.3	-9.0 (-62.8%)	56.5	87.4	-30.9 (-35.3%)	199.6	155.6	44.0 (28.3%)	$846	$1,149	$-302 (-26.3%)
I5	569 (2)	6,482 (57)	8.8	9.6	-0.8 (-8.1%)	70.3	82.7	-12.4 (-15.0%)	237.3	152.9	84.4 (55.2%)	$1,082	$1,160	$-78 (-6.7%)
I6	560 (2)	6,440 (57)	5.4	11.2	-5.8 (-52.1%)	78.6	80.1	-1.6 (-1.9%)	283.9	159.2	124.7 (78.3%)	$1,465	$1,189	$276 (23.2%)
I7	560 (2)	6,384 (57)	7.1	8.6	-1.5 (-17.1%)	82.1	76.6	5.5 (7.2%)	212.5	166.4	46.1 (27.7%)	$1,251	$1,202	$48 (4.0%)
I8	557 (2)	6,351 (57)	3.6	9.8	-6.2 (-63.2%)	73.6	87.4	-13.8 (-15.8%)	208.3	158.3	49.9 (31.5%)	$1,075	$1,144	$-69 (-6.1%)
I9	549 (2)	6,267 (57)	7.3	11.0	-3.7 (-33.8%)	69.2	89.2	-20.0 (-22.4%)	235.0	168.2	66.8 (39.7%)	$965	$1,182	$-216 (-18.3%)
I10	542 (2)	6,208 (57)	9.2	11.9	-2.7 (-22.6%)	83.0	86.5	-3.5 (-4.0%)	241.7	175.2	66.5 (38.0%)	$1,177	$1,117	$61 (5.4%)

Table V.5 *(continued)*

Sources: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at the Centers for Medicare & Medicaid Services.

Notes: The baseline quarters are measured relative to the start of the baseline period on January 1, 2012. For example, the first baseline quarter (B1) ran from January 1, 2012, to March 31, 2012. The intervention quarters are measured relative to the start of the intervention period on January 1, 2013. For example, the first intervention quarter (I1) rans from January 1, 2013, to March 31, 2013. In each period (baseline or intervention), the treatment group each quarter included all beneficiaries assigned to a treatment practice by the start of the quarter and who met other sample criteria—that is, they were enrolled in FFS Medicare, lived in Alaska and were observable. In each period, the comparison group included all beneficiaries who were assigned to a comparison practice by the start of the quarter and who met the other sample criteria. See text for details.

The difference between the treatment and comparison groups in a quarter is calculated by subtracting the mean outcome for the comparison group from the mean outcome for the treatment group. The percentage difference equals that difference divided by the mean outcome for the comparison group.

B = baseline; C = comparison; CHF = congestive heart failure; Diff = difference; ED = emergency department; FFS = fee-for-service; I = intervention; Q = quarter; T = treatment.

The outpatient ED visit rates were higher for the treatment group than the comparison group in every baseline and intervention quarter, both for the full sample and the high-risk subgroup. During the baseline period, the rate for the full sample declined for the treatment group, from 235.4 visits per 1,000 beneficiaries in B1 to 165 in B4, and fluctuated from 137.1 to 147.8 for the comparison group. During the intervention quarters, the rates fluctuated from quarter to quarter with no clear pattern for both groups, ranging from 163.7 to 241.5 for the treatment group and 130.5 to 147.7 for the comparison group. Among the high-risk subgroups, rates were higher, and consistently higher among the treatment group than the comparison group.

Spending. Mean monthly Medicare Part A and B spending fluctuated from $748 to $1,165 over the baseline and intervention quarters, without any trend among the full treatment group. Mean spending among the comparison group hovered close to $900 during the period. Treatment group spending was lower than comparison group spending in 3 of 4 baseline quarters and 6 of 10 intervention quarters, with no clear trend in the differences. Among the high-risk subgroups, mean spending was slightly higher, ranging from $846 to $1,465 among the high-risk treatment group and from $1,004 to $1,202 among the high-risk comparison group.

3. Results for primary tests, by domain

Overview. The primary tests conducted for this report cover the full primary test period (I5 through I10). For the quality-of-care process domain, we found substantively large favorable impacts driven by estimates of program impacts on the diabetes process of care measure. The primary tests also found substantively large but not statistically significant impacts on quality-of-care outcomes. We found no impacts that were statistically significant or larger than the substantive thresholds in either the service use or spending domains (Table V.6).

Quality-of-care processes. The likelihood of receiving recommended processes of care for diabetes was 57 percent higher for the treatment group than the estimated counterfactual. (Our estimated counterfactual—the outcome the treatment group members would have had in the absence of the HCIA program—is the treatment group mean minus the difference-in-differences estimate.) This statistically significant favorable estimate is substantively large because it is larger than the substantive threshold of 15 percent. The likelihood of receiving an ambulatory care visit within 14 days of hospital discharge was 36 percent lower in the treatment group than its estimated counterfactual (and 37 percent lower in the high-risk subset of the treatment group). This unfavorable estimate is substantively large but we cannot conclude that it is statistically significant because our one-sided statistical tests assess only improvements in outcomes. The combined estimate across the three tests in the quality-of-care processes domain was -5.4 percent, an unfavorable point estimate that was not substantively large. The statistical power to detect substantively large effects was moderate to poor for the individual and combined tests.

Table V.6. Results of primary tests for PeaceHealth

Primary test definition					Statistical power to detect an effect that is[a]		Results			
Domain (number of tests in the domain)	Outcome (units)	Time period for impacts (controlling for baseline differences)[b]	Population	Substantive threshold (impact as a percentage relative to the counterfactual)[c]	Size of the substantive threshold	Twice the substantive threshold[d]	Treatment group mean	Regression-adjusted difference between the treatment and estimated counterfactual[c] (standard error)	Percentage difference[e]	*p*-value[f]
Quality-of-care processes (3)	Received all four recommended diabetes processes of care in the year (binary [yes or no]/beneficiary/ year)	Final intervention year (corresponding to intervention quarters 7 through 10)	Medicare FFS beneficiaries with diabetes and ages 18 to 75 assigned to treatment practices	15.0%(+)	23.2%	42.7%	31.6	11.5** (5.5)	57.2%	0.04
	All inpatient admissions within a quarter were followed by an ambulatory care visit with a primary care or specialist provider within 14 days (binary [yes or no]/beneficiary/ year)	Average over intervention quarters 5 through 10	Medicare FFS beneficiaries with CHF, diabetes, and/or hypertension assigned to treatment clinics and who had at least one hospital stay in the quarter	15.0%(+)	46.4%	86.5%	25.8	-15.4 (5.2)	-37.4%	>0.99
	All inpatient admissions within a quarter were followed by an ambulatory care visit with a primary care or specialist provider within 14 days (binary [yes or no]/beneficiary/ year)	Average over intervention quarters 5 through 10	All Medicare FFS beneficiaries assigned to treatment clinics and who had at least one hospital stay in the quarter	15.0%(+)	53.7%	92.9%	26.1	-14.7 (4.4)	-36.0%	>0.99
	Combined	Varies by test	Varies by test	15.0%	45.4%	85.3%	n.a.	n.a.	-5.4%	0.66

Table V.6 *(continued)*

Primary test definition					Statistical power to detect an effect that is[a]		Results			
Domain (number of tests in the domain)	Outcome (units)	Time period for impacts (controlling for baseline differences)[b]	Population	Substantive threshold (impact as a percentage relative to the counterfactual)[c]	Size of the substantive threshold	Twice the substantive threshold[d]	Treatment group mean	Regression-adjusted difference between the treatment and estimated counterfactual[c] (standard error)	Percentage difference[e]	*p*-value[f]
Quality-of-care outcomes (2)	30-day unplanned hospital readmissions (#/1,000 beneficiaries/quarter)	Average over intervention quarters 5 through 10	Medicare FFS beneficiaries with CHF, diabetes, and/or hypertension assigned to treatment clinics	15.0%(-)	20.0%	34.5%	6.9	-4.2 (3.8)	-37.8%	0.17
	30-day unplanned hospital readmissions (#/1,000 beneficiaries/ quarter)	Average over intervention quarters 5 through 10	All Medicare FFS beneficiaries assigned to treatment clinics	5.0%(-)	12.7%	15.9%	6.7	-0.2 (2.4)	-3.3%	0.49
	Combined	Average over intervention quarters 5 through 10	Varies by test	10.0%(-)	17.6%	28.2%	n.a.	n.a.	-20.6%	0.23
Service use (4)	All-cause inpatient admissions (#/1,000 beneficiaries/ quarter)	Average over intervention quarters 5 through 10	Medicare FFS beneficiaries with CHF, diabetes, and/or hypertension assigned to treatment clinics	15.0%(-)	49.2%	89.2%	76.1	-19.0 (11.3)	-19.9%	0.12
	All-cause inpatient admissions (#/1,000 beneficiaries/ quarter)	Average over intervention quarters 5 through 10	All Medicare FFS beneficiaries assigned to treatment clinics	5.0%(-)	20.8%	36.5%	63.8	-8.2 (7.7)	-11.3%	0.30

Table V.6 *(continued)*

Primary test definition					Statistical power to detect an effect that is[a]		Results			
Domain (number of tests in the domain)	Outcome (units)	Time period for impacts (controlling for baseline differences)[b]	Population	Substantive threshold (impact as a percentage relative to the counterfactual)[c]	Size of the substantive threshold	Twice the substantive threshold[d]	Treatment group mean	Regression-adjusted difference between the treatment and estimated counterfactual[c] (standard error)	Percentage difference[e]	*p*-value[f]
	Outpatient ED visit rate (#/1,000 beneficiares/ quarter)	Average over intervention quarters 5 through 10	Medicare FFS beneficiaries with CHF, diabetes, and/or hypertension assigned to treatment clinics	15.0%(-)	57.8%	95.3%	236.4	14.0 (22.6)	6.3%	0.56
	Outpatient ED visit rate (#/1,000 beneficiaries/ quarter)	Average over intervention quarters 5 through 10	All Medicare FFS beneficiaries assigned to treatment clinics	5.0%(-)	25.2%	47.9%	207.3	19.7 (15.3)	10.5%	0.78
	Combined	Average over intervention quarters 5 through 10	Varies by test	10.0%(-)	47.2%	87.3%	n.a.	n.a.	-3.6%	0.33
Spending (2)	Medicare Part A and B spending ($/beneficiary/ month)	Average over intervention quarters 5 through 10	Medicare FFS beneficiaries with CHF, diabetes, and/or hypertension assigned to treatment clinics	15.0%(-)	52.2%	91.8%	1,169	-74 (139)	-5.9%	0.36
	Medicare Part A and B spending ($/beneficiary/ month)	Average over intervention quarters 5 through 10	All Medicare FFS beneficiaries assigned to treatment clinics	5.0%(-)	23.1%	42.4%	974	5 (89)	0.5%	0.50
	Combined	Average over intervention quarters 5 through 10	Varies by test	10.0%(-)	39.9%	78.0%	n.a.	n.a.	-2.7%	0.39

Table V.6 *(continued)*

Sources: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at the Centers for Medicare & Medicaid Services.

Note: The results for each outcome are based on a difference-in-differences regression model, as described in the text. Estimates are calculated for Medicare beneficiaries who are observable in the relevant time period: that is, beneficiaries who are enrolled in Medicare FFS (Part A and B), are alive, and have Medicare as their primary payer. Additional sample restrictions apply to the quality-of-care process measures; see text for details.

[a] The power calculation is based on actual standard errors from the analysis. For example, in the second to last row, a 5 percent effect on Medicare Part A and B spending (from the counterfactual of $974 + $5 = $979) would be a change of $49. Given the standard error of $89 from the regression model, we would be able to detect a statistically significant result 23.1 percent of the time if the impact was truly -$49, assuming a one-sided statistical test at the $p = 0.10$ significance level.

[b] We estimated impacts as the average across intervention quarters 5 through 10 for all outcomes but one: the quality-of-care process measure for diabetes. For this measure, we calculated outcomes instead over the year-long period covering intervention quarters 7 through 10.

[c] The substantive threshold is the impact as a percentage of the counterfactual. The counterfactual is the outcome the treatment group would have had in the absence of the HCIA-funded intervention. Our estimate of the counterfactual is the treatment group mean minus the regression-adjusted difference-in-differences estimate.

[d] We show statistical power to detect a very large effect (twice the size of the substantive threshold) because this provides additional information about the likelihood that we will find effects if the program is indeed effective. If power to detect effects is less than 75 percent even for a very large effect, then the evaluation is extremely poorly powered for that outcome.

[e] Percentage difference is calculated as the regression-adjusted difference between the treatment and comparison group, divided by the adjusted comparison group mean.

[f] *p*-values test the null hypothesis that the regression-adjusted difference-in-differences estimate is less than or equal to zero for outcomes in the quality-of-care processes domain, or greater than or equal to zero in all other domains (a one-sided test). Because it is a one-sided test, as the difference-in-differences estimate approaches infinity in an unfavorable direction (negative for process-of-care measures and positive for all other measures), the *p*-value approaches 1, whereas it would approach 0 in a two-sided test. We adjusted the *p*-values for the multiple (three) comparisons made within the quality-of-care processes domain, and (separately) for the two comparisons made within the quality-of-care outcomes domain, for the four comparisons made within the service use domain, and the two comparisons made in the spending domain.

*/**/*** Significantly different from zero at the .10/.05/.01 levels, one-tailed test, respectively. No difference-in-differences estimates were significantly different from zero at the .05 level.

CHF = congestive heart failure; ED = emergency department; FFS = fee-for-service; HCIA = Health Care Innovation Award.

n.a. = not applicable.

Quality-of-care outcomes. The rate of 30-day unplanned readmissions for the full treatment group and the high-risk treatment subgroup were 3.3 and 37.8 percent lower, respectively, than our estimates of the counterfactuals. These lower rates for the treatment group were in the favorable direction but not statistically significant. The estimate for the high-risk subgroup was larger than the substantive threshold but the rate for the full population was not. After combining results across the two populations, the combined effect was 20.6 percent, larger than the substantive threshold of 10 percent and in the favorable direction.

The statistical power to detect effects the size of the substantive threshold was poor for both the individual tests and the combined effect in the domain.

Service use. The admission rate for the full treatment group and high-risk subgroup were 11.3 percent and 19.9 percent lower, respectively, than each of their estimated counterfactuals. Although neither of these differences was statistically significant, both were substantively large. The ED visit rate was 10.5 percent higher—higher than the substantive threshold of 5 percent—than the counterfactual for the full treatment group. The high-risk treatment group had an ED visit rate that was 6.3 percent higher than its counterfactual. After combining results across the four tests in two outcomes in this domain, the combined estimate was a 3.6 percent reduction in service use. This favorable estimate was neither substantively large nor statistically significant. Power to detect effects that were the size of the substantive thresholds was marginal for the test for ED visits among the high-risk group and poor for all the other tests and for the combined measure.

Spending. The treatment group averaged $974 per beneficiary per month in Part A and B spending during the 5th through 10th intervention quarters, a value 0.5 percent (or $5) higher than the estimated counterfactual. This difference was much smaller than the substantive threshold of 5.0 percent. Among the high-risk subgroup, the treatment group had average spending that was 5.9 percent lower than the counterfactual, but this estimate was also neither statistically nor substantively large. Statistical power to detect an effect the size of the substantive threshold was marginal for the high-risk group and poor for the test on the full treatment group and the combined estimate.

4. Results for secondary tests

Estimates during the first intervention year (January 1, 2013, to December 31, 2013). As shown in Table V.7, there are a number of substantively large and/or statistically significant differences between the treatment group and estimated counterfactual in the secondary test period. The treatment group had a 15.0 percent higher rate for the diabetes quality-of-care measure and an 18.8 percent lower rate of unplanned readmissions than the estimated counterfactual. These are large, favorable estimates, but they are much smaller than the estimates in the primary test period, which is consistent with the program ramping up over this period. Secondary tests for the 14-day follow-up measure, admissions, and spending generated estimated differences between the treatment group and the estimated counterfactual that were larger than the differences estimated over the primary test period and in several cases statistically significant.

Table V.7. Results of secondary tests for PeaceHealth

Domain (number of tests in the domain)	Outcome (units)	Time period for impacts (controlling for baseline differences)	Population	Treatment group mean	Regression-adjusted difference between treatment group mean and counterfactual (standard error)	Percentage difference[a]	*p*-value[b]
Quality-of-care processes (3)	Received all four recommended diabetes processes of care in the year (binary [yes or no]/beneficiary/year)	First intervention year (corresponding to intervention quarters 1 through 4)	Medicare FFS beneficiaries with diabetes and ages 18 to 75 assigned to treatment practices	22.5	2.9 (4.7)	15.1%	0.26
	All inpatient admissions within a quarter were followed by an ambulatory care visit with a primary care or specialist provider within 14 days (binary [yes or no]/beneficiary/year)	Average over intervention quarters 1 through 4	Medicare FFS beneficiaries with CHF, diabetes, and/or hypertension assigned to treatment clinics and who had at least one hospital stay in the quarter	23.8	-13.0 (5.5)	-35.3%	>0.99
	All inpatient admissions within a quarter were followed by an ambulatory care visit with a primary care or specialist provider within 14 days (binary [yes or no]/beneficiary/year)	Average over intervention quarters 1 through 4	Medicare FFS beneficiaries assigned to treatment clinics and who had at least one hospital stay in the quarter	23.9	-13.2 (4.8)	-35.6%	0.99
Quality-of-care outcomes (2)	30-day unplanned hospital readmissions (#/1,000 beneficiaries/quarter)	Average over intervention quarters 1 through 4	Medicare FFS beneficiaries with CHF, diabetes, and/or hypertension assigned to treatment clinics	10.3	-2.4 (4.7)	-18.8%	0.31
	30-day unplanned hospital readmissions (#/1,000 beneficiaries/quarter)	Average over intervention quarters 1 through 4	Medicare FFS beneficiaries assigned to treatment clinics	8.5	0.3 (3.1)	3.3%	0.53

Table V.7 *(continued)*

Domain (number of tests in the domain)	Outcome (units)	Time period for impacts (controlling for baseline differences)	Population	Treatment group mean	Regression-adjusted difference between treatment group mean and counterfactual (standard error)	Percentage difference[a]	*p*-value[b]
Service use (4)	All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	Average over intervention quarters 1 through 4	Medicare FFS beneficiaries with CHF, diabetes, and/or hypertension assigned to treatment clinics	75.7	-20.1** (11.9)	-21.0%	0.04
	All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	Average over intervention quarters 1 through 4	Medicare FFS beneficiaries assigned to treatment clinics	64.8	-9.5 (8.4)	-12.8%	0.13
	Outpatient ED visit rate (#/1,000 beneficiaries/quarter)	Average over intervention quarters 1 through 4	Medicare FFS beneficiaries with CHF, diabetes, and/or hypertension assigned to treatment clinics	213.6	-5.5 (21.9)	-2.5%	0.40
	Outpatient ED visit rate (#/1,000 beneficiaries/quarter)	Average over intervention quarters 1 through 4	Medicare FFS beneficiaries assigned to treatment clinics	199.3	-5.2 (17.0)	-2.7%	0.62
Spending (2)	Medicare Part A and B spending ($/beneficiary/month)	Average over intervention quarters 1 through 4	Medicare FFS beneficiaries with CHF, diabetes, and/or hypertension assigned to treatment clinics	1,001	-208** (120)	-17.2%	0.04
	Medicare Part A and B spending ($/beneficiary/month)	Average over intervention quarters 1 through 4	Medicare FFS beneficiaries assigned to treatment clinics	841	-105 (83)	-11.1%	0.10

Sources: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at the Centers for Medicare & Medicaid Services.

Note: The results for each outcome are based on a difference-in-differences regression model, as described in the text. Estimates are calculated for Medicare beneficiaries who are observable in the relevant time period: that is, beneficiaries who are enrolled in Medicare FFS (Part A and B), are alive, and have Medicare as their primary payer. Additional sample restrictions apply to the quality-of-care process measures; see text for details.

[a] Percentage difference is calculated as the regression-adjusted difference between the treatment and comparison group, divided by the adjusted comparison group mean.

[b] The *p*-values from the secondary test results were not adjusted for multiple comparisons within or across domains.

CHF = congestive heart failure; ED = emergency department; FFS = fee-for-service.

Large differences (favorable and unfavorable) during the first intervention year, a period during which we and PeaceHealth did not expect to see large program effects, suggest there might be some unobservable differences between treatment and comparison groups that partially drove the results. This finding is not surprising, given that we were unable to construct a comparison group that was well balanced on all important baseline characteristics; as a result, there are some differences between the two groups on important variables, and these might be correlated with other important and unobserved variables that affected outcomes. These secondary test results remind us to interpret the results of our primary tests cautiously, given that these potential unobservable differences and large observed differences could bias the results.

Estimates limiting the sample to prevent sample addition. The secondary test results (not included) limited to those beneficiaries attributed at the start of the baseline or intervention period are consistent with the primary test results. They show no evidence that differential sample addition between the treatment and comparison practices drove the results seen in the primary test.

5. Consistency of impact estimates with implementation findings

The impact estimates in the primary tests are plausible given the implementation findings. The primary tests showed substantively large favorable and statistically significant impacts on diabetes quality-of-care processes. This finding is consistent with PeaceHealth's emphasis on high-risk patients and its care management components focusing exclusively on high-risk patients, including patients with diabetes. Care coordinators identified and contacted patients with uncontrolled chronic conditions, and scheduled appointments for patients who needed a routine screening or test, including those in the diabetes composite quality-of-care process measure. They also provided them with free diabetic testing supplies.

The substantively large decline in 14-day follow-up visits was surprising given the program's transitional care component, but not implausible. We would have expected to see an increase in 14-day follow-up visits, but the 24-hour follow-up call with care coordinators could have served as a substitute for an in-person visit, especially given the remoteness of the area and transportation difficulties. The substantively large decline in 30-day readmissions for high-risk beneficiaries is also plausible, given the program's transitional care component, with additional support from short- and longer-term case management.

Finally, the lack of significant or substantive findings on service use or spending in the primary test period is plausible even though the implementation evidence shows the program was active over this period. For example, PeaceHealth served a total of 3,881 program participants, which was 111 percent of its target for the three-year award (Section III.B.2). However, even with a well-implemented intervention, it is possible that the program was unable to change beneficiaries' or providers' behaviors in ways that would affect impact outcomes during the primary test period covered in this report.

6. Conclusions about program impacts, by domain

Based on all evidence available, we could draw a conclusion in only one of the four domains—quality-of-care process outcomes. Table V.8 summarizes this conclusion, the tests that support it, and why we cannot draw conclusions in the other domains.

- **We could not draw conclusions on program impacts on quality-of-care outcomes, service use, or spending.** Although some of the tests in these domains produced substantively large estimates, we cannot draw favorable or unfavorable conclusions on program impacts in these domains for several reasons. First, as discussed in Section V.C, the comparison group did not meet industry standards for baseline equivalence to the treatment group along a number of important variables. Difference-in-differences estimates should control for any time-invariant differences between the treatment and comparison groups, but our findings from the secondary tests suggest that this might not fully control for all of the differences. The secondary tests showed large differences between the treatment group and estimated counterfactual for admissions and spending during a time when the program should have had smaller, if any impacts. This, combined with the lack of baseline balance, suggests that time-varying differences between the treatment and comparison groups other than the HCIA program affected service use and spending. Therefore, we do not think it is valid to draw conclusions for service use or spending—or for quality-of-care outcomes, given that the one outcome (30-day readmissions) was highly related to another outcome, all-cause admissions, that showed implausibly large difference in the secondary test period. Our inability to draw conclusions is likely due in part to small sample sizes in the treatment group. The substantial quarter-to-quarter variation in treatment group outcomes, likely from noise due to small sample sizes and outliers, makes it difficult to identify any impacts even if the program had impacts.
- **The program had a substantively large and statistically significant impact on quality-of-care processes.** This conclusion is based on the large, favorable impact estimate for diabetes care. However, the point estimates for the two other process-of-care measures in the domain—ambulatory care follow-up within 14 days of the index stay for the full Medicare population and for the high-risk subset—were substantively large and *unfavorable.* This could be because the nurse telephone calls after the hospital admissions substituted for ambulatory care visits (causing the 14-day follow-up rate to decline), rather than prompting such visits. Despite the concerns that limit us from drawing conclusions in the three other domains, we are comfortable drawing conclusions in the quality-of-care process domain for several reasons. First, as discussed in Section V.D.5, the results were consistent with implementation evidence on diabetes care. Second, the diabetes measure had a clear trend over time, with modest improvement among the treatment group in the first intervention year and larger improvement in the second year, as can be seen in unadjusted means and primary and secondary test results. Finally, we do not think the implausible secondary results for service use and spending invalidate the comparison group for process-of-care measures because (1) the service use and spending measures were calculated over a different (broader) population than the process of care measures, (2) processes of care are very different from service use and spending, and (3) the secondary tests for the diabetes measures themselves corroborated that the impact estimate during the first intervention year is consistent with our and PeaceHealth's expectation to not see large program effects.

Table V.8. Final conclusions about the impacts of PeaceHealth's HCIA program on patient outcomes, by domain

Domain	Final conclusion	Evidence supporting conclusion		
		Primary test result(s) that supported conclusion	Primary test result(s) plausible given secondary tests?	Primary test result(s) plausible given implementation evidence?
Quality-of-care process	Statistically significant favorable effect	Statistically significant favorable effect on composite measure for diabetes care	Yes	Yes
Quality-of-care outcomes	No conclusion	Differences between treatment and comparison groups in 30-day unplanned readmissions were substantively large and favorable in the test of the combined effect across two populations (all beneficiaries and high-risk beneficiaries)	No	Yes
Service use	No conclusion	No statistically significant or substantively important effect; power was poor to marginal to detect an effect on either of the two outcomes in the domain, for tests on all beneficiaries and high-risk beneficiaries	No	Yes
Spending	No conclusion	No statistically significant or substantively important effect; power was poor to marginal to detect an effect on the single outcome in the domain, for tests on all beneficiaries and high-risk beneficiaries	No	Yes

Sources: Tables V.6 and V.7

HCIA = Health Care Innovation Award.

VI. DISCUSSION AND CONCLUSIONS

PeaceHealth used its $3 million HCIA to implement the coordinated care program. The program involved four interrelated components: (1) general transitional care services for all patients discharged from the PeaceHealth Ketchikan Medical Center and intensive transitional care services for patients with CHF; (2) short-term care management for patients with a temporary medical or social hurdle; (3) longer-term case management for patients requiring assistance to effectively manage their chronic conditions; and (4) population health management, including redefining the scrub-and-huddle process and providing outreach to paneled patients to improve preventive care. Through these four intervention components, PeaceHealth aimed to improve quality of care for Medicare FFS beneficiaries; reduce the need for expensive hospitalizations and ED visits, particularly among high-risk beneficiaries; and lower total Medicare spending.

Despite some delays in the first year of the program, PeaceHealth implemented the program consistent with its core design. Several measures capture the generally successful implementation:

- PeaceHealth hired 10.0 FTE staff members with HCIA funding, almost meeting its original staffing goal of 11.0 FTE staff. Of these newly hired staff members, 4.5 FTE staff were care coordinators who delivered services that spanned all four program components.
- The two treatment clinics provided intervention services to 3,881 patients, exceeding the program's original target by 331 patients.
- Service metrics for transitional care indicate that, during the third year of the program, care coordinators consistently made follow-up calls to more than 70 percent of all patients discharged from the Ketchikan Medical Center hospital or ED, depending on the quarter.

The results from our implementation and impact evaluations enable us to draw conclusions on program impacts in only one of four domains—quality-of-care processes. These results suggest that the program led to an improvement in this domain, based solely on evidence that shows an increase in the percentage of Medicare FFS beneficiaries at PeaceHealth with diabetes who received all four recommended diabetes processes of care. The results do not provide evidence that the program increased a second quality-of-care measure on post-discharge ambulatory care visits.

The improvements in diabetes measure suggest that PeaceHealth's significant investment in population health management had its intended effects on processes of care. In particular, care coordinators' efforts in contacting patients with diabetes, and use of the scrub-and-huddle process to scheduled appointments and routine tests, resulted in an increase in the receipt of recommended tests among patients with diabetes. Another reason the program might have had such large effects on this process-of-care measure is that only 18 percent of beneficiaries with diabetes received all recommended care during the baseline period, a very low rate with considerable room for improvement.

We are unable to draw conclusions in the quality-of-care outcomes, service use, and spending domains. Although the implementation evidence was generally consistent with impact estimates in these domains, baseline comparisons between the treatment and comparison groups and the results from the secondary tests suggest that the two groups differed in important ways that could bias the estimates of program impacts. Despite the fact that we cannot draw conclusions in these domains, we have included the results in this report for transparency and so stakeholders can review the evidence and draw their own conclusions.

Our inability to draw conclusions in three domains does not appear to be due to major problems implementing the intervention as planned. Rather, our inability to draw conclusions points to two limitations in the evaluation. First, PeaceHealth's HCIA-funded intervention affected all patients at two primary care practices in Alaska. We were unable to identify a similar set of practices in similar parts of Alaska to serve as a comparison group. For this reason, we designed our evaluation to estimate program impacts using 57 practices in remote parts of Alaska

as a comparison group. The comparison group was unmatched, and some differences existed between the treatment and comparison groups during a 12-month period before the intervention began. Second, because the PeaceHealth patient population was small—in both the number of practices and the number of beneficiaries—and because our evaluation covered only Medicare FFS beneficiaries, the treatment group was very small, limiting statistical power and resulting in imprecise estimates.

CMMI and other stakeholders could consider a number of changes to the design of similar programs in the future to increase the potential to draw conclusions about program impacts on patients' outcomes. One possible solution to the lack of similar comparison practices might be to randomize patients within the program population so that some receive the new program services (such as meetings with care coordinators) and others do not. This would allow valid estimates of the impact of program services, even without an external comparison group. In addition, the problem of small treatment population size might be alleviated with more timely, high quality Medicaid data. Nondual Medicaid beneficiaries comprised 15 percent of PeaceHealth's program participants. Adding this population to the evaluation would improve statistical power and the relevance of the impact estimates to the intervention overall.

REFERENCES

Boards of Trustees, Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds. "2013 Annual Report of the Boards of Trustees of the Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds." Table V.D1. Washington, DC: Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds, 2013. Available at http://downloads.cms.gov/files/TR2013.pdf. Accessed August 13, 2014.

Chronic Conditions Data Warehouse. "Table A.1 Medicare Beneficiary Counts for 2003–2012." Baltimore, MD: CMS, 2014. Available at https://www.ccwdata.org/cs/groups/public/documents/document/ccw_website_table_a1.pdf Accessed November 19, 2014.

Gerhardt, Geoffrey, Alshadye Yemane, Keri Apostle, Allison Oelschlaeger, Eric Rollins, and Niall Brennan. "Evaluating Whether Changes in Utilization of Hospital Outpatient Services Contributed to Lower Medicare Readmission Rate." *Medicare & Medicaid Research Review,* vol. 4, no. 1, 2014, pp. E1–E13.

Gilman, Boyd, Purvi Sevak, Victoria Peebles, Greg Peterson, Catherine DesRoches, Sandi Nelson, Laura Blue, Keith Kranker, Kate Stewart, Frank Yoon, and Lorenzo Moreno. "Findings: PeaceHealth BlueCross BlueShield." In Moreno, Lorenzo, Boyd Gilman, Greg Peterson, Catherine DesRoches, Sheila Hoag, Linda Barterian, Laura Blue, Katherine Bradley, Emily Ehrlich, Kristin Geonnotti, Lauren Hula, Keith Kranker, Rumin Sarwar, KeriAnn Wells, Joseph Zickafoose, Sandi Nelson, Kate Stewart, and Frank Yoon. "Evaluation of Health Care Innovation Awards (HCIA): Primary Care Redesign Programs." Second annual report to CMS. Volume II: Individual Program Summaries. Princeton, NJ: Mathematica Policy Research, December 11, 2015.

Health Indicators Warehouse. "Average Age of Medicare Beneficiaries (mean)." Hyattsville, MD: Health Indicators Warehouse, National Center for Health Statistics, 2014a. Available at http://www.healthindicators.gov/Indicators/Average-age-of-Medicare-beneficiaries-mean_308/Profile/ClassicData. Accessed November 19, 2014.

Health Indicators Warehouse. "Hospital Inpatient Medicare Admissions (per 1,000 beneficiaries)." Hyattsville, MD: Health Indicators Warehouse, National Center for Health Statistics, 2014b. Available at http://www.healthindicators.gov/Indicators/Hospital-inpatient-Medicare-admissions-per-1000-beneficiaries_2001/Profile/ClassicData. Accessed August 13, 2014.

Health Indicators Warehouse. "Medicare Beneficiaries Who Are Also Eligible for Medicaid (percent)." Hyattsville, MD: National Center for Health Statistics, HIW, 2014c. Available at http://www.healthindicators.gov/Indicators/Medicare-beneficiaries-eligible-for-Medicaid-percent_317/Profile/ClassicData. Accessed August 4, 2015.

Institute of Education Sciences. "What Works Clearinghouse: Procedures and Standards Handbook, Version 3.0." Washington, DC: U.S. Department of Education, IES, 2014. Available at http://ies.ed.gov/ncee/wwc/DocumentSum.aspx?sid=19. Accessed September 15, 2014.

Peikes, Deborah, Stacy Dale, Eric Lundquist, Janice Genevro, and David Myers. "Building the Evidence Base for the Medical Home: What Sample and Sample Size Do Studies Need? White Paper." AHRQ Publication No.11-0100-EF. Rockville, MD: Agency for Healthcare Research and Quality, October 2011.

Rosenthal, M.B., S. Alidina, M. Friedberg, S. Singer, D. Eastman, Z. Li, and E. Schneider. "A Difference-in-Differences Analysis of Changes in Quality, Utilization, and Cost Following the Colorado Multi-Payer Patient-Centered Medical Home Pilot." *Journal of General Internal Medicine,* vol. 31, no. 3, 2016, pp. 289–296.

Taylor, Erin Fries, Stacy Dale, Deborah Peikes, Randall Brown, Arka Ghosh, Jesse Crosson, Grace Anglin, Rosalind Keith, Rachel Shapiro, and contributing authors. "Evaluation of the Comprehensive Primary Care Initiative: First Annual Report." Prepared for the U.S. Department of Health and Human Services, Centers for Medicare & Medicaid Services. Princeton, NJ: Mathematica Policy Research, January 2015.

CHAPTER 7

RUTGERS CENTER FOR STATE HEALTH POLICY

Purvi Sevak, Cara Stepanczuk, Katharine Bradley, Greg Peterson, Boyd Gilman, Laura Blue, Keith Kranker, Kate Stewart, Sheila Hoag, and Lorenzo Moreno

This page has been left blank for double-sided copying.

RUTGERS CENTER FOR STATE HEALTH POLICY

CHAPTER SUMMARY

Introduction. Rutgers Center for State Health Policy (CSHP) used its $14.3 million Health Care Innovation Award (HCIA) to implement a community-based care management/care-coordination program, also known as a "hotspotting" program, at four provider organizations. Based on an existing care management/care coordination hotspotting model designed by the Camden Coalition of Healthcare Providers (Camden Coalition), the CSHP program used multidisciplinary, community-based care teams to connect participants who were frequent users of hospital services ("high utilizers") to appropriate clinical and social services, help them manage their conditions, and overcome socioeconomic obstacles to care. The four sites that implemented the program served 1,068 participants from January 2013 to June 2015 (when HCIA-funded operations concluded) across four diverse institutional and geographic settings. CSHP aimed to reduce average annual costs of care by 14.8 percent by the end of the award by reducing patients' use of inappropriate acute care—such as inpatient admissions and emergency department (ED) visits—and by increasing use of appropriate primary and specialty care.

Objectives. This report (1) describes the design and implementation of CSHP's HCIA-funded intervention, (2) estimates the impacts of the intervention on patients' outcomes and Medicare spending during the award, and (3) uses both implementation and impact findings to identify possible explanations for the observed impacts.

Methods. We reviewed CSHP's program documents and self-monitoring metrics, conducted interviews with CSHP leadership and site-specific program staff, and surveyed program staff about their experiences. To estimate impacts, we compared outcomes for Medicare fee-for-service (FFS) patients served by the four implementation sites with outcomes for Medicare patients in similar locations and with similar characteristics, adjusting for any differences in outcomes between the two groups during a one-year baseline period. We did not include Medicaid beneficiaries or uninsured patients in the impact evaluation due to limitations in available data.

Program design and implementation. The intervention was a community-based care management/care coordination program that encompassed three different activities: (1) enrolling patients soon after hospital discharge; (2) providing care coordination and care management services (for example, medication reconciliation, arranging for transportation, and assistance applying for social services) through mobile care teams; and (3) providing training and coaching to improve patients' capacity to manage their own medical and social conditions. The available implementation evidence indicates that the intervention was implemented largely as planned, although all four sites experienced implementation challenges. For example, participants had very complex needs and faced a variety of barriers to appropriate care, such that they required longer than expected program participation periods and ongoing program support to succeed in changing their use of health system resources. Second, significant limitations in the social service and health care systems—such as lack of affordable housing and specialist availability—limited care teams' ability to stabilize participants and encourage behavioral change. Only 62

percent of participants graduated from the program and about one-third of participants dropped out of the program before meeting their individual goals.

Impacts on patients' outcomes. The impact estimates indicate that, during the three years of the award, the intervention had a statistically significant favorable effect on quality-of-care outcomes, driven mostly by a decrease in the number of 30-day unplanned hospital readmissions. However, the estimates show that the intervention had an indeterminate effect on patients' outcomes in the other three evaluation domains: quality-of-care processes, service use, and Medicare spending. (Outcomes include the proportion of patients discharged from a hospital who received a primary care or specialist visit within 14 days, all-cause inpatient admissions, outpatient ED visit rates, and Medicare inpatient and total Part A and B spending). There was no evidence of statistically significant or substantively large favorable effects in these three domains, but the statistical power to detect effects for these domains was marginal.

Conclusion. Although the CSHP program appeared to have improved quality-of-care outcomes among Medicare FFS beneficiaries, there was no evidence that CSHP achieved its goal of reducing health care spending among this population. The lack of observed effects on Medicare spending—and on outcomes in the quality-of-care processes and service use domains—appeared not be due to a failure to implement the program as planned. The lack of effects could be due to a combination of two factors: (1) challenges in sustaining long-term behavioral change in a population with complex medical and social needs and (2) limitations in the local health and social service systems that the program was designed to leverage. It is also possible that the program had effects on outcomes other than quality-of-care outcomes, but our evaluation failed to detect them due to insufficient statistical power or because effects were concentrated among Medicaid and uninsured populations, which our estimates did not include.

Summary of intervention and impact results for Rutgers Center for State Health Policy

Intervention description			
Awardee description		Research group at Rutgers University that guided and funded implementation at four program sites	
Award amount ($ millions)		$14.3 million	
Award extended beyond June 2015?		No	
Locations		High-poverty areas in four cities (Allentown, Pennsylvania; Aurora, Colorado; Kansas City, Missouri; and San Diego, California)	
Target population		Frequent users of hospital services (inpatient or outpatient ED), typically 2 or more hospitalizations in prior 6 months	
Interventions		Care management to address medical, behavioral, and social needs, • Delivered by multidisciplinary care teams[a] • Teams scheduled medical appointments and provided transportation • Patients coached on physician visits and self-management • Patients linked to social and behavioral health services (for example, SSDI benefits and substance abuse treatment centers)	
Metrics of intervention delivered		• Enrolled 1,068 people (all insurance types) • For those enrolled - 10 contacts/month on average for 4.2 months - 66% met care goals and so graduated[b]	
Impact evaluation methods			
Core design		Contemporaneous differences model with matched comparison group	
Treatment group	Definition	Medicare FFS beneficiaries who enrolled in the program	
	# of beneficiaries during primary test period[c]	113 to 149	
Comparison group definition		Matched Medicare FFS beneficiaries living in same or similar geographic areas as treatment beneficiaries	
Impact results: Quality-of-care processes domain			
Ambulatory care visit within 14 days of discharge (% of beneficiaries/quarter)		Comparison mean[d]	37.4%
		Impact estimate (% difference)	+3.6 pp (+9.7%)
Impact conclusion[e]		Indeterminate effect	
Impact results: Quality-of-care outcomes domain			
30-day unplanned hospital readmissions (#/1,000 beneficiaries/quarter)		Comparison mean[d]	365
		Impact estimate (% difference)	-126 (-34.4%)*
Inpatient admissions for ACSCs (#/1,000 beneficiaries/quarter)		Comparison mean[d]	215
		Impact estimate (% difference)	-27 (-12.4%)
Combined impact estimate[f]		-23.4%**	
Impact conclusion[e]		**Statistically significant favorable effect**	
Impact results: Service use domain			
All-cause inpatient admissions (#/1,000 beneficiaries /quarter)		Comparison mean[d]	784
		Impact estimate (% difference)	-116 (-14.8%)
Outpatient ED visits (#/1,000 beneficiaries/quarter)		Comparison mean[d]	1196
		Impact estimate (% difference)	+57 (+4.8%)
Combined impact estimate[f]		-5.0%	
Impact conclusion[e]		Indeterminate effect	
Impact results: Spending domain			
Medicare Part A and B spending ($/beneficiary/month)		Comparison mean[d]	$5,332
		Impact estimate (% difference)	-$468 (-8.8%)
Medicare inpatient spending ($/beneficiary/month)		Comparison mean[d]	$3,048
		Impact estimate (% difference)	-$40 (-1.3%)
Combined impact estimate[f]		-5.0%[g]	
Impact conclusion[e]		Indeterminate effect	

Note: See the CSHP chapter for details on the intervention, impact methods, and impact results.

[a] The composition of the multidisciplinary teams varied across the four implementation sites. Teams included combinations of nurses, nurse practitioners, social workers, community health workers, peer health coaches, medical assistants, behavioral health providers, and community volunteers.

[b] The other 33 percent of beneficiaries exited the intervention without graduating because they moved from the catchment area, became unreachable by care team staff, declined to participate further, or died.

[c] Number of beneficiaries in the full treatment group across the quarters in the primary test period.

Summary of intervention and impact results for Rutgers Center for State Health Policy *(continued)*

[d] The comparison mean is the estimate of the outcome the treatment group beneficiaries would have had if they had not received the intervention. It is equal to the mean for the treatment group over the intervention quarters (in the primary test period) minus the impact estimate.

[e] We drew conclusions at the domain level based on the results of prespecified primary tests and consistency with implementation evidence. For each domain, we could draw one of five conclusions: (1) Statistically significant favorable effect (the highest level of evidence), (2) Substantively important (but not statistically significant) favorable effect, (3) Substantively important (but not statistically significant) unfavorable effect, (4), No substantively large effect, and (5) Indeterminate effect. Section IV.A.7 of this report describes the decision rules we used to reach each of these possible conclusions.

[f] The combined estimate is the average across all the individual estimates in the domain, in which the impact estimate for each individual outcome is expressed as percentage change relative to the comparison group.

[g] The combined estimate included the impact estimate for Medicare Part A and B spending (reported in this table) and the estimate for inpatient spending only (reported in the full chapter only).

*Significantly different from zero at the .10 level, one-tailed test.

**Significantly different from zero at the .05 level, one-tailed test.

***Significantly different from zero at the .01 level, one-tailed test.

ACSC = ambulatory care-sensitive condition; CSHP = Center for State Health Policy; ED = emergency department; FFS = fee-for-service; pp = percentage point; SSDI = Social Security Disability Insurance.

I. INTRODUCTION

This report presents findings from the evaluation of Rutgers Center for State Health Policy's (CSHP) Health Care Innovation Award (HCIA), with a focus on program impacts on patients' outcomes. Section II provides an overview of CSHP's HCIA-funded intervention and the design of the impact evaluation. Section III describes the design and implementation of the intervention, including how the program could be expected to affect study outcomes through changes in patients' behavior. Section IV describes our methods for, and results and conclusions from, estimating program impacts on patients' outcomes in four domains: quality-of-care processes, quality-of-care outcomes, service use, and spending. Section V draws conclusions by synthesizing the impact and implementation findings. The impact estimates and conclusions in this report are final because they cover CSHP's three-year award period of the HCIA, which ended in June 2015.

II. OVERVIEW OF CSHP'S HCIA-FUNDED INTERVENTION AND THE IMPACT EVALUATION

A. CSHP's HCIA-funded intervention

CSHP received a $14.3 million award to implement a hotspotting program—that is, a community-based care management/care coordination program—at four provider organizations. Based on an existing hotspotting model designed by the Camden Coalition of Healthcare Providers (Camden Coalition), the CSHP program used multidisciplinary, community-based care teams to (1) connect participants who were frequent users of hospital services (high utilizers) to appropriate clinical and social services, (2) help them manage their conditions, and (3) overcome socioeconomic obstacles to care. CSHP is a research group at Rutgers University; its primary roles as an HCIA awardee were to guide the four provider organizations that served as implementation sites, coordinate technical assistance to the sites from partner organizations, and administer the HCIA funding. The implementation sites served 1,068 participants from January 2013 to June 2015 (when HCIA-funded operations concluded) across four diverse institutional and geographic settings: (1) an independent physician association in San Diego, California (MultiCultural Primary Care Medical Group); (2) a nonprofit community health center in Aurora, Colorado (Metro Community Provider Network); (3) a nonprofit health system with two hospitals in Kansas City, Missouri (Truman Medical Center); and (4) a nonprofit operator of two federally qualified health centers (FQHCs) in Allentown, Pennsylvania (Neighborhood Health Centers of the Lehigh Valley).

The CSHP program aimed to reduce the average annual cost of care by 14.8 percent by the end of the award (Table II.1). CSHP expected to achieve this outcome through three key activities: (1) enrolling patients soon after hospital discharge, (2) providing care management/care coordination services through mobile care teams, and (3) providing training and coaching to improve patients' capacity to manage their own medical and social conditions. Care team composition varied by site, and care teams included different combinations of nurses, nurse practitioners, social workers, community health workers, peer health coaches, medical assistants, and behavioral health providers. CSHP expected that its program activities would decrease unnecessary hospital admissions, decrease unnecessary emergency department (ED)

visits, and increase use of appropriate primary and specialty care. These shifts in use of health care services were expected, in turn, to reduce total spending. (Section III.A.3 describes the awardee's theory of action in detail.)

Table II.1. Summary of CSHP's HCIA program and our evaluation for estimating its impacts on patient outcomes

Program description	
Award amount	$14,347,808
Award start date	July 1, 2012
Implementation date	January 2, 2013
Award end date	June 30, 2015
Awardee description	CSHP is a research group at Rutgers University; its primary role as an HCIA recipient was to guide the four implementation sites, coordinate the technical assistance provided by partner organizations, and administer the HCIA funding.
Intervention overview	CSHP supported four institutions in adapting the care coordination hotspotting model pioneered by the Camden Coalition. This model features multidisciplinary, community-based care teams that connect frequent users of hospital services to appropriate clinical and social services, help them manage their conditions, and address socioeconomic obstacles to care.
Intervention component	**Care management/care coordination for high-risk patients.** Multidisciplinary, mobile care teams worked with patients to address their medical, behavioral, and social needs. Care team structure varied by site, and teams included different combinations of nurses, nurse practitioners, social workers, community health workers, peer health coaches, medical assistants, and behavioral health providers. Care teams helped patients secure primary care and specialist appointments and provided additional services as needed, including coaching patients through physician visits, providing transportation to physician visits, and helping link patients to social services such as shelter or Social Security disability benefits.
Target population	The initial target was 2,425 frequent users of inpatient and ED services, although this was revised to 1,691 after the first year of implementation. Typically, individuals were eligible if they had two or more hospital admissions within the prior six months, although the specific rules varied from site to site. Within these parameters, the program enrolled individuals whom staff believed were most likely to benefit from care coordination and whose needs did not exceed program resources.
Target impacts on patients' outcomes	• Reduce total annual cost of care by 14.8 percent • Reduce hospital admissions and ED visits (amount not specified) • Increase primary care use and efficient use of health care resources (amount not specified)
Workforce development	Sites hired 35.3 full-time-equivalent staff; data are not available to confirm how many were fully or partially supported through HCIA funding. In addition, one site used 35 community volunteers who worked an estimated 600 hours to provide support services.
Location	Allentown, Pennsylvania; Aurora, Colorado; Kansas City, Missouri; San Diego, California
Impact evaluation	
Core design	Contemporaneous differences with matched comparison group, adjusted for differences in baseline characteristics
Treatment group	Medicare FFS beneficiaries enrolled at any of the four CSHP implementation sites from January 1, 2013, to March 31, 2015, and who were continuously enrolled in Medicare FFS for 12 baseline months before program enrollment
Comparison group	Medicare FFS beneficiaries, continuously enrolled in Medicare FFS for 12 baseline months, with at least one chronic condition and one hospital admission or ED visit, whose hospital discharge status was within the set of discharge statuses we observed among treatment beneficiaries, and whose zip codes in the Medicare enrollment database indicated residence in geographic areas that were the same or similar to geographic areas in which the treatment group resided

Table II.1 *(continued)*

Intervention component(s) included in impact evaluation	Care management/care coordination for high-risk patients. Although implemented independently and with some variation, all four sites focused on care management to affect outcomes.
Extent to which the treatment group reflects the awardee's target population (for the component(s) evaluated)	Low. Only 26 percent of program enrollees (through March 31, 2015) had Medicare FFS as their primary payer, either alone or dually with Medicaid. To match treatment group members to comparison group members, we further restricted the Medicare FFS population to Medicare beneficiaries who were continuously enrolled in Medicare FFS for all 12 baseline months before program enrollment.
Study outcomes, by domain	**1.** Quality-of-care processes. 14-day follow-up to hospitalization **2.** Quality-of-care outcomes. 30-day unplanned readmissions and inpatient admissions for ambulatory care-sensitive conditions 3. Service use. All-cause inpatient admissions and outpatient ED visits 4. Spending. Medicare Part A and B spending, Medicare inpatient spending

Source: Review of CSHP reports, including its original application, operational plan, and 15 quarterly narrative reports to the Centers for Medicare & Medicaid Services.

CSHP = Center for State Health Policy; ED = emergency department; FFS = fee-for-service; HCIA = Health Care Innovation Award.

B. Overview of impact evaluation

To estimate program impacts on patients' outcomes, we compared outcomes for Medicare fee-for-service (FFS) beneficiaries participating in the HCIA intervention (the treatment group) to outcomes for beneficiaries in a matched comparison group, adjusting for individual differences in characteristics between these two groups before the intervention began. These characteristics include baseline levels of utilization and spending of the treatment and comparison beneficiaries. The bottom panel of Table II.1 summarizes our impact evaluation design.

We selected beneficiaries for the potential comparison group if they met two claims-based criteria for CSHP program eligibility: (1) at least one of 25 chronic conditions and (2) at least one outpatient ED visit or hospital discharge during the three-year HCIA funding period. In addition, potential comparison beneficiaries had to reside in the same or similar locations as the treatment group beneficiaries. Although CSHP used subjective criteria as well as utilization information to select program participants (for example, selecting people considered likely to benefit from the intervention), these subjective criteria are not replicable in claims data. This means we could not account for them when selecting the comparison group.

We estimated impacts on outcomes with Medicare FFS claims data and grouped them into four domains: (1) quality-of-care processes, (2) quality-of-care outcomes, (3) service use, and (4) spending. Across the HCIA awardees in primary care redesign (PCR), we designed our impact evaluations to identify promising interventions or intervention components—consistent with evaluation goals from the Center for Medicare & Medicaid Innovation (CMMI) to find programs that could be scaled or retested as part of a future model test. Before conducting the analysis, we specified a series of primary tests, describing the evidence we would need to conclude that the program was effective, and the awardee and CMMI reviewed these tests. Each test specified a population, outcome, period, expected direction of effect, and threshold that we count as substantively important. The purpose of these primary tests was to focus the impact evaluation

on hypotheses that would provide the most robust evidence about program effectiveness. We used the results from the primary tests to draw conclusions about program impacts in each of the four evaluation domains. Because we sought to identify promising programs, rather than only those with unequivocally demonstrated success, we conducted one-sided statistical tests (that is, testing only for program benefits) and used a threshold for statistical significance of 0.1, which is less strict than the conventional standard of 0.05.

Our impact evaluation reflects the effects of the intervention on the 26 percent of CSHP's patient population that was in Medicare FFS (including those who were dually eligible for Medicaid). Because the evaluation's treatment group was limited to Medicare beneficiaries, the evaluation design partially aligns with CSHP's intervention; however, the evaluation's treatment group omits uninsured patients, Medicare Advantage beneficiaries, Medicaid beneficiaries who are not dually eligible for Medicare, and the small number of commercially insured patients enrolled in the program. Because the sample size was too small for site-specific impact estimates, the impacts are aggregated across all four program sites.

III. PROGRAM IMPLEMENTATION

This section provides a detailed description of CSHP's HCIA-funded intervention, highlighting how it evolved over time and its theory of action. Then it assesses the evidence on the extent to which the intervention was implemented as planned, based on measures of implementation timeliness, program enrollment, service delivery, staffing, and stakeholder relationships. Third, the section summarizes the facilitators and barriers associated with implementation effectiveness.

We based our evaluation of CSHP's program implementation on telephone discussions and email correspondence with program administrators, as well as information collected during site visit interviews with frontline staff conducted in June 2014 and April 2015. Although we also use information from the awardee's quarterly reports to CMMI and self-monitoring program metrics, we did not verify the quality of the performance data reported by the awardee.

A. Program design and adaptation

1. Target population and patient identification, recruitment, and enrollment

CSHP's HCIA-funded intervention targeted frequent users of inpatient and ED services who resided in the areas surrounding the four implementation sites. The program accepted participants regardless of payer type. In this section, we describe how CSHP identified and enrolled program participants.

Target population of patients. The four implementation sites differed in how they defined the target population. Initially, all sites followed a recommendation from the Camden Coalition to limit participants to those with two or more hospital admissions within the prior six months. Two of the sites changed the eligibility criteria early in the award period to expand the pool of potential participants. Site-specific eligibility criteria were as follows:

- Two sites used the original criterion of two or more hospital admissions in the prior 6 months.
- One site recruited individuals with two or more hospital admissions in the prior 6 months or three or more admissions in the prior 12 months.
- One site recruited individuals with three or more hospital events (admissions or ED visits) in the prior 6 months.

Within these eligibility requirements, the program enrolled individuals whom staff believed were most likely to benefit from care coordination and whose needs did not exceed program resources. For example, implementation sites typically excluded individuals with cancer, substance abuse, or serious behavioral problems because they were less likely than other potential participants to benefit from the services available. However, sites differed in how they applied these exclusions. At least two sites adjusted eligibility criteria according to caseload and program capacity.

Identification, recruitment, and enrollment of patients for care management and care coordination. CSHP's sites also varied in their methods of patient identification, recruitment, and enrollment. Two sites identified eligible patients through hospital electronic health records, and the other two used referrals from hospitals, other health care providers, and community organizations. After identifying potential participants, care team members approached them at bedside (before discharge) or in the community (after discharge) and explained the program. According to a CSHP administrator, approaching medically and socially complex patients in the hospital led to patients' greater acceptance of program services and engagement in care plans. In contrast, care teams experienced more difficulty engaging patients if the first contact occurred after discharge. This was a challenge for the two sites dependent on referrals, as they typically did not receive referrals until after potential participants had been discharged. After patients consented to participate, care team staff scheduled a home visit to complete program enrollment and begin care management and care coordination services.

2. Intervention components

CSHP's intervention had one component that encompassed a number of key activities for beneficiaries enrolled in the program. Enrolled participants were assigned to multidisciplinary mobile care teams, which worked to address their medical, behavioral, and social needs. Care teams worked with participants at their homes or in other locations in the community, such as libraries, to provide care management and care coordination. Care teams provided the following services based on the needs of each participant:

- Developing and monitoring progress on goal-oriented care plans
- Scheduling, preparing for, and accompanying participants to appointments with primary care providers and specialists
- Reconciling partcipants' medications
- Conducting home visits and telephone calls to support disease management

- Arranging for transportation to and from appointments and pharmacies
- Assisting with enrollment in social service (such as housing or Social Security disability benefits) and behavioral health service programs (such as substance abuse treatment centers)
- One site also provided direct medical and behavioral health care services

The duration of program participation varied, with each site tailoring the time frame to its patient population and institutional preference. The average length of participation across the four sites, among those with a specified intervention exit date, was 4.2 months, with site-specific averages ranging from 2.4 to 6.3 months.

3. Theory of action

Based on a review of CSHP's program activities and goals, we developed a theory of action to describe the mechanisms through which program administrators expected the program to improve the outcomes we selected for the impact evaluation. (Table II.1 provides a list of these outcomes.) CSHP expected the HCIA-funded program to improve patients' outcomes in the following ways:

1. **Mobile care teams provided care management and care coordination services to improve patients' use of health system and capacity for self-care.** The CSHP program employed care teams to help program participants manage their medical and social conditions.
2. **Improvements in patients' capacity to manage their medical and social conditions reduced the need for potentially avoidable and costly acute care services.** Several of the care management and care coordination services provided by care teams aimed to help patients connect to and improve their relationships with primary and specialty care physicians. Care teams also provided education about appropriate use of primary and specialty care as alternatives to emergency and hospital care. These actions should have resulted in increased follow-up primary and specialty care after hospitalization. Care teams also helped participants learn to manage their conditions and reduce their use of emergency and hospital services for non-emergent issues. This should have resulted in reduced rates of 30-day unplanned readmissions, inpatient admissions for ambulatory care-sensitive conditions, all-cause inpatient admissions, and outpatient ED visits.
3. **More efficient utilization led to reduced total cost of care.** As a result of increased use of follow-up preventive, primary, and specialty care and reduced use of ED and inpatient services, we would expect to see a reduction in Medicare Part A and B spending and Medicare inpatient spending.

Text box III.1. Example from CSHP illustrating the program's theory of action

"Patient A in Kansas City has multiple chronic conditions and poly-substance abuse, a history of homelessness, frequent ED visits, and no PCP [primary care provider]. At the initial contact with the care team, the patient stated that he would "never want to conform to the rules." The care team's strategy is to first establish firm trust. They accomplished this by identifying opportunities to provide basic help, such as involving family members in explaining the impact on diet of modifying cooking practices, supplying a scale and log to support the modification, organizing and explaining the purpose of medications, arranging for transportation and enabling the patient to do so, scheduling and accompanying patients to medical and social service appointments. Within weeks, the patient has started scheduling transportation and keeping his appointments independent of the care team, and now states that he cares about his health. His sister reflects, "He used to use the ER [emergency room] for everything. Now he asks when his appointment is."

Source: CSHP's third quarterly report to CMS.

4. Intervention staff and workforce development

Table III.1 provides key details about staff hired for the HCIA-funded intervention. As mentioned earlier, each site hired, managed, and organized its own personnel. Each site developed a different staff structure and workflow. In general, care teams included clinical staff who focused on meeting participants' medical needs and nonclinical staff who focused on their social needs. HCIA funding supported care teams, as did financial and in-kind support from program sites and other implementation partners (such as local advocacy groups). Sites also employed staff in administrative and data specialist roles that supported the care teams. One site relied on community volunteers to support the care team as well, who provided peer coaching, community outreach, administrative and documentation support, and patient assessment and follow-up.

Table III.1. Key details about intervention staff

Role	Staff members	Staff/team responsibilities	Adaptations?
Care team staff	Varied by site; included various combinations of nurses, nurse practitioners, social workers, community health workers, peer health coaches, medical assistants, and behavioral health providers	Care teams worked with patients at their homes or other locations in the community, such as libraries, to provide care management and care coordination. See Section III.A.2 for a detailed list of responsibilities.	Yes. Over time, sites shifted care team staff and workflow to optimize patients' enrollment, care management, and program graduation rates, as well as to balance resource constraints.
Support staff	Varied by site; included various combinations of supervisors, data specialists, stakeholder engagement/advocacy staff, and community volunteers	- Supervisors managed staff and improved program operations over time. - Data specialists tracked program operations and participants. - Stakeholder engagement/advocacy staff sought funding for ongoing program operations. - Volunteers provided participant follow-up and post-graduation services.	Yes. Sites made changes to improve program processes and balance resource constraints, such as increasing the use of volunteers and adjusting the focus of advocacy staff over time.

Source: Interviews and document review.

B. Implementation effectiveness

This section examines the evidence on implementation effectiveness—that is, it analyzes measures of the intervention delivered and, when possible, compares those measures to the services the awardee intended to deliver. We assess the evidence on implementation effectiveness in five areas: (1) implementation timeliness, (2) program enrollment, (3) service delivery, (4) staffing, and (5) stakeholder relationships. To conduct this analysis, we used data from interviews with program administrators and selected frontline staff, self-reported metrics included in CSHP's self-monitoring and measurement reports to CMMI, participants' data from CSHP, and responses to a survey of program staff.

1. Program timeline

All four sites began to enroll and provide program services to participants soon after the targeted program launch date (December 2012). As the CSHP final program narrative to CMMI noted, "As soon as the operational plan was approved by CMS [the Centers for Medicare & Medicaid Services], in late 2012, the clinical sites began hiring and constituting care teams." However, sites reported that it took 6 to 12 months to become fully operational. The four sites reached full operational capacity at different speeds due to variation in resources and administrators' ability to develop efficient workflows and protocols.

2. Program enrollment

The four sites enrolled 63 percent (1,068) of the overall (revised) enrollment target of 1,691 (Table III.2). Enrollment success varied by site; two sites met their site-specific enrollment targets and two did not. CSHP's final report to CMMI confirmed that the four sites faced challenges in meeting enrollment goals due to (1) difficulties obtaining real-time data alerts to identify potential participants who were in the hospital or had recently been hospitalized; and (2) longer-than-expected program participation periods, which reduced the ability of care teams to enroll new participants. Administrators did not set goals for length of program participation, but initially expected that sites would follow the Camden Coalition's model of providing 60 to 90 days of intensive care management/care coordination. Instead, participants stayed in the program for an average of 4.2 months across sites.

One reason for long participation periods was that it was difficult to determine participants' readiness for graduation from the program. The Camden Coalition did not provide protocols for assessing readiness for graduation, so each site developed its own guidelines and methods for this process. These readiness assessments usually centered on participants' progress, or lack of progress, toward meeting their goals. Care teams graduated 666 participants by the end of the award, about two-thirds of those enrolled. One-third of participants exited the intervention without graduating because they moved out of the catchment area, became unreachable by care team staff, declined to participate further, or died. As of June 2015, 4 percent of participants were still active, according to CSHP's final report to CMMI. Many of those who graduated reportedly remained involved in the program in some way. For example, at least two of the sites provided services to participants after graduation because they believed that graduates would destabilize without continued support.

Table III.2. Enrollment measures

Measure	Target	Actual	Met target?
Program enrollment	1,691[a]	1,068 (through June 2015)	No
Program graduation	Not specified	666 (through June 2015)	n.a.

Source: Analysis of CSHP's HCIA quarterly reports, December 2012 through August 2015.

[a] CSHP's initial target was about 2,425 frequent users of inpatient and ED services, although this was revised to 1,691 after the first year of implementation.

CSHP = Center for State Health Policy; ED = emergency department.

n.a. = not applicable.

3. Service-related measures

CSHP did not define specific goals for service-related metrics, although the four sites tracked care team contacts with or on behalf of participants per month (these might include, for example, telephone calls to physicians or other service providers); time spent with or on behalf of participants; and initial home visits within a week of discharge (Table III.3). Figure III.1 shows the quarterly number of care team contacts with or on behalf of participants per participant-month. The elevated level of activity in the final quarter raised the overall average number of care team contacts per participant per month to 10.3 (for an average of six hours per participant per month). The awardee's quarterly reports to CMMI did not include other process measurements or goals. However, data from a survey on training that Mathematica administered to program staff in 2015 indicate that program staff spent most of their time conducting activities consistent with the program's design, including calling, coaching, and assisting participants in an effort to connect them to resources and improve their capacity to manage their medical and social conditions. (Because no formal or informal training was provided as part of CSHP's program, we do not present any other findings from the HCIA Primary Care Redesign Trainee Survey.)

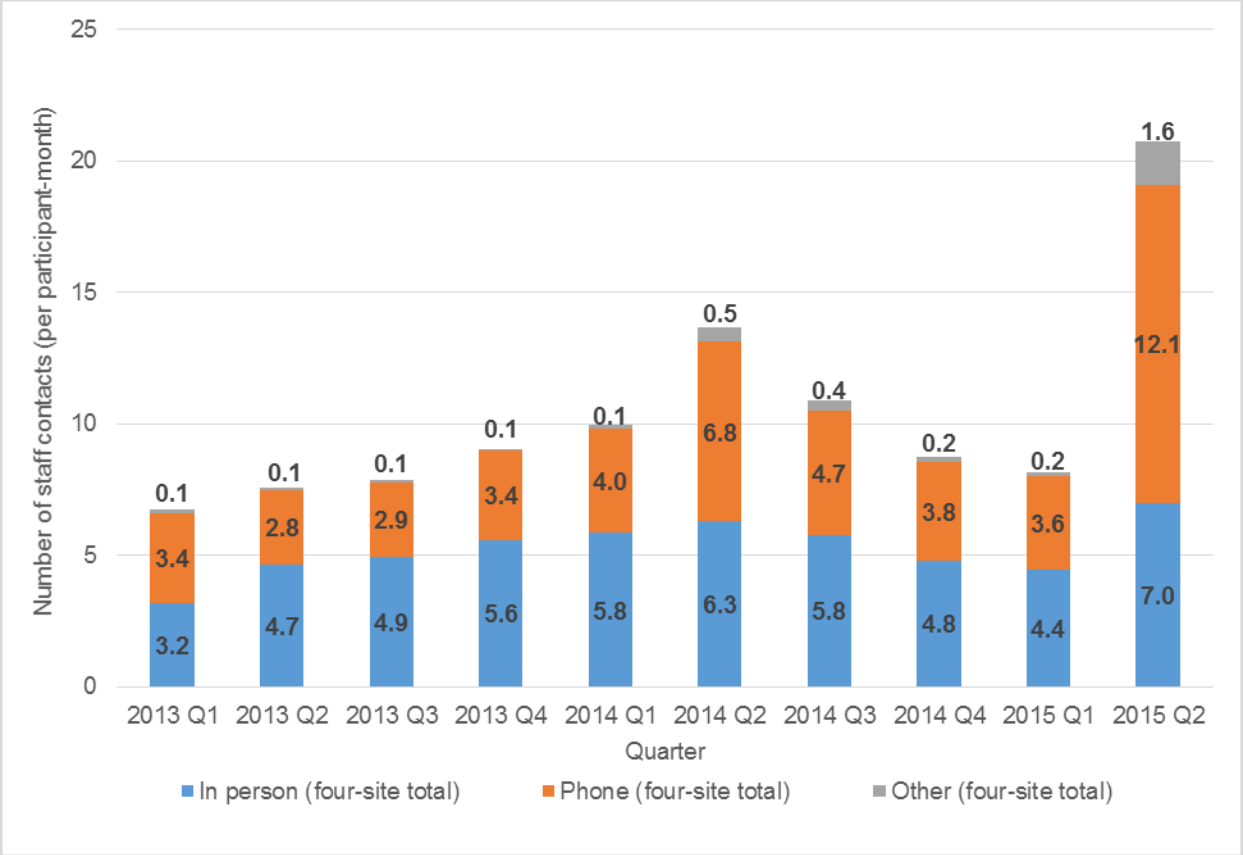
Table III.3. Service metrics and targets

Service metrics	Awardee's target	Actual
Staff contacts with or on behalf of participants	Not specified	Mean of 10.3 per participant per month (through June 2015)
Time spent with or on behalf of participants	Not specified	Mean of nearly 6 hours per participant per month
Initial home visits within 7 days of index discharge	Not specified	59 percent

Source: Analysis of CSHP's HCIA quarterly reports, December 2012 through August 2015, and interviews from second site visit, April 2015.

CSHP = Center for State Health Policy; HCIA = Health Care Innovation Award.

Figure III.1. Care team contacts with or on behalf of participants per participant-month



Source: Analysis of CSHP's HCIA quarterly reports, December 2012 through August 2015.
CSHP = Center for State Health Policy; HCIA = Health Care Innovation Award; Q = quarter.

4. Staffing measures

The sites established care teams in all four locations by January 2013, although, as mentioned earlier, it took 6 to 12 months for sites to reach full operational capacity. Six months after program launch, there were 10.5 total full-time-equivalent (FTE) staff and 12 months after program launch, there were 39.2 total FTEs, including existing and new staff. The four sites hired a total of 35.3 FTE new care team and support staff (73 percent of their target) by the end of the three-year award period.

All four sites experienced difficulty in hiring and maintaining staff for three reasons: (1) delays in CMS approval of the program's operational plan, (2) challenges selecting care team staff with the skills needed to successfully interact with patients with complex medical and social needs, and (3) staff perceptions of employment insecurity due to the time-limited nature of HCIA funding. These staffing issues reduced sites' capacity for contacting participants and enrolling new patients. Some sites reported modifying their workflows in response to difficulty filling vacant positions. According to one of CSHP's reports to CMMI, uncertainty about

program financing after the three-year award period led CSHP to minimize new hires in the final quarter of implementation.

5. Stakeholder relationships

Stakeholder relationships were an essential part of the CSHP program: stakeholders were a source of referrals, helped coordinate medical and behavioral health services, shared data to help monitor participants, and served as connections to public benefits such as housing or transportation, among others. Some groups even provided some funding to partially support program staff. However, implementation sites had mixed success building relationships with different types of stakeholders. Although sites reported success in their attempts to partner with local primary care providers, health plans, and community organizations, they struggled to secure data-sharing agreements and funding commitments from hospitals to support the continuation of program services after HCIA funding ended.

C. Summary of facilitators of and barriers to implementation

Several factors facilitated implementation of CSHP's HCIA-funded intervention and others hindered implementation. We described those factors in detail in the second annual report (Bradley et al. 2015). Here we summarize key facilitators and barriers from that report and provide limited new information that supports earlier findings (Table III.4).

Three factors served as both facilitators of and barriers to program implementation. First, the adaptability of the program enabled sites to optimize implementation for their organizational contexts and patient populations. Specifically, sites tailored eligibility criteria, length of program participation, and workflows based on staff and participants' needs. As one program manager noted, having the freedom to adapt the intervention was the reason for the site's success. However, adaptability was also a barrier to effective implementation, as explained by a CSHP administrator: "Rapid innovation requires an enormous amount of energy, and poses significant challenges to a small health center as well as a large hospital system for different reasons: a small organization may be nimble but tends to be resource-poor, and a large organization may have greater financial resources but can frequently be a 'big ship to turn.'" Second, although hospital partnerships remained weak, partnerships with local primary care providers, health plans, community organizations, and local political leaders improved sites' ability to help participants. Third, team communication and cohesion was important for care teams' success; those with a solid supervisory structure and frequent collaboration across all levels of staff reported better performance and staff satisfaction.

Finally, two other important barriers to implementation were (1) the medical and social complexity of the target population and (2) limitations in the social service and health care systems. Care teams often struggled to help participants overcome multiple, interconnected barriers to care and stabilize their chronic conditions. Limitations in the social and health safety net exacerbated these challenges. For example, lack of affordable housing made it difficult to stabilize participants simultaneously contending with homelessness and uncontrolled medical issues.

Table III.4. Summary of key facilitators of and barriers to the implementation of CSHP's HCIA-funded intervention

Facilitators (+) and barriers (-)	Description based on findings in second annual report	Additional supporting data not available in the second annual report, if applicable
Program characteristics		
(+) Adaptability of the program to meet the needs of participants and staff **(-) Rapid adaptation** and frequency of changes to care teams' roles	(+) Adaptation enabled the sites to conform to the organizational, cultural, and financial characteristics of the sites' host institutions and to accommodate the views of important local stakeholders, such as hospitals. Adaptation also enabled sites to make improvements in response to self-monitoring data and to bolster staff engagement by incorporating staff suggestions. (-) However, frequent changes to care team roles, made in response to staff turnover and caseload fluctuations, could make implementation feel chaotic.	None
Implementation process		
(+) Engaging stakeholders such as local primary care providers, health plans, community organizations, and local political leaders **(-) Engaging hospital stakeholders**	(+) Program administrators and staff reported that their efforts to engage external stakeholders helped to support program implementation and sustain the progress they made. For example, positive working relationships with primary care providers helped to improve participant–clinician communication, increase participants' access to care, and strengthen collaboration to meet the complex needs of participants. (-) However, staff experienced difficulty working with hospitals; in particular, they struggled to obtain referrals for new patients, utilization data for current enrollees, and financial support from local hospitals.	CSHP continued to find building stakeholder relationships to be a facilitator of and a barrier to implementation during the last year of program operation: (+) In addition to financial support from sites' host organizations, three of the implementation sites obtained funding commitments from other sources to continue program operations beyond the HCIA funding period. (-) Although some hospitals were supportive of the program in spirit, they were not willing to commit funding for program operations beyond the HCIA funding period.
Internal factors		
(+) Team communication and cohesion **(-) Team communication and cohesion**	Team communication was important for program implementation, although sites had different levels of success in this area over time. Care teams with a solid supervisory structure and frequent collaboration across all levels of staff experienced greater implementation success and staff satisfaction.	All four sites reported that a cohesive, passionate care team composed of strong personalities with varied perspectives was both an incredible asset and a source of continuous challenge.

Table III.4 *(continued)*

Facilitators (+) and barriers (-)	Description based on findings in second annual report	Additional supporting data not available in the second annual report, if applicable
External factors		
(-) Patients with complex needs and resource constraints	Many participants faced a variety of barriers to appropriate care, including lack of stable income, health insurance, legal residency, English language proficiency, knowledge of the health system and chronic disease management, stable housing, social support, and transportation. Many also had issues with cultural barriers, mental illness and substance abuse (despite informal program eligibility criteria that excluded some patients with these conditions), and traumatic experiences that made stabilizing their chronic conditions more difficult. Participants' issues often took longer to resolve than the intervention's time line typically allowed.	The HCIA Primary Care Redesign Trainee Survey asked program staff who had received HCIA-funded training about the extent to which patients' resistance to the program was a barrier to staff performing their duties effectively. Some respondents (16 percent) thought it was a major barrier and half thought it was a minor barrier.
(-) Limitations in the social service and health care systems	A general lack of affordable housing, insufficient transportation services, and poor access to specialty care were the most significant environmental barriers to stabilizing participants' social and medical conditions.	The Trainee Survey asked about the extent to which inadequate community support was a barrier to staff performing their duties effectively. Some respondents (20 percent) thought it was a major barrier and 40 percent thought it was a minor barrier. The Trainee Survey also asked about the extent to which clinicians' resistance to the program was a barrier to staff performing their duties effectively. Nearly a third of respondents thought it was either a minor or major barrier (22 and 8 percent, respectively).

Sources: Interviews from second site visit, April 2015; document review through September 2015; HCIA Primary Care Redesign Trainee Survey.

Note: We reviewed four domains associated with implementation experience: (1) program characteristics, (2) implementation process, (3) internal factors, and (4) external environment. Implementation research suggests that barriers and facilitators within these domains are important determinants of implementation effectiveness.

CSHP = Center for State Health Policy; HCIA = Health Care Innovation Award.

D. Conclusions about the extent to which the program, as implemented, reflected the core design

The evidence indicates that CSHP implemented its HCIA-funded intervention largely as planned. First, all four sites began providing intervention services on schedule, although they did not reach full operational capacity for another 6 to 12 months. Second, although all four sites faced challenges to enrollment, two met their site-specific enrollment targets by the end of the award. Third, frontline staff reported spending their time on care management/care coordination activities in a way that was consistent with the design of the intervention. However, the lack of established protocols and clear process goals limits our ability to assess service delivery. Fourth, although all four sites experienced challenges in hiring or maintaining staff, they compensated somewhat by shifting workflows. Finally, although some stakeholders were hesitant to commit funding to support the program, sites partnered successfully with health care, payer, and community-based organizations to improve the care of their participants.

Sites implemented the intervention largely as planned, but one key aspect of program implementation could have limited its success. As noted previously, only 62 percent of participants graduated from the program and 33 percent of enrollees exited the intervention without graduating (4 percent of participants were still active as of June 2015). In addition, at least two sites provided services to participants after graduation because they believed that graduates would destabilize without continued support. These metrics are consistent with two important implementation barriers. First, CSHP sites reported that participants had very complex medical and social needs and faced a variety of barriers to appropriate care, such that they required longer-than-expected participation periods and ongoing support to succeed in changing their use of health system resources. In many cases it was not possible to resolve participants' issues or fully address their needs. Second, significant limitations in the social service and health care systems limited care teams' ability to stabilize participants and encourage behavioral change.

IV. PROGRAM IMPACTS ON PATIENTS' OUTCOMES

In this section of the report, we draw conclusions, based on available evidence, about the impacts of CSHP's HCIA program on patients' outcomes in four domains: quality-of-care processes, quality-of-care outcomes, service use, and spending. We first describe the methods for estimating impacts (Section IV.A.) and then the characteristics of the treatment and comparison groups at the start of the intervention (Section IV.B). We next demonstrate that the treatment beneficiaries were similar at the start of the intervention to the beneficiaries we selected for the comparison group, which is important for limiting potential bias in impact estimates (Section IV.C). Finally, in Section IV.D, we describe the quantitative impact estimates, their plausibility given implementation findings, and our conclusions about program impacts in each domain. Our conclusions in this report are final because the analyses include CSHP's full HCIA award period.

The findings in this report update the impact results from the second annual report for CSHP (Bradley et al. 2015). Secifically, we (1) included additional treatment group beneficiaries by extending the enrollment period, (2) rematched treatment group beneficiaries to potential comparison beneficiaries so that all treatment beneficiaries had one or more matched comparison

beneficiaries, (3) extended the period that outcomes were measured in claims data by 6 months (from December 31, 2014 to August 31, 2015), and (4) added one outcome measure (ambulatory-care follow-up visit within 14 days of a hospital discharge).

A. Methods

1. Overview

We estimated program impacts on patients' outcomes as the difference in outcomes between Medicare FFS beneficiaries in the treatment group and those in a matched comparison group, adjusting for observed differences in outcomes between these groups during the year before the intervention began. We prespecified primary tests, describing the evidence we would need to conclude that the program was effective, and the awardee and CMMI reviewed these. Each test specified a population, outcome, time period, expected direction of effect, and threshold that we count as substantively important. The purpose of these primary tests was to focus the impact evaluation on hypotheses that would provide the most robust evidence about program effectiveness. We used the results from the primary tests, in conjunction with the implementation evidence, to draw conclusions about program impacts in each of the four evaluation domains. The remaining subsections describe each component of the impact evaluation in more detail.

2. Treatment group definition

The treatment group included Medicare FFS beneficiaries who enrolled in the CSHP program from January 1, 2013, through March 31, 2015. (One of the four CSHP program sites—Truman Medical Center in Kansas City, Missouri—enrolled two Medicare FFS beneficiaries in November and December 2012, before the program start date elsewhere; the treatment group includes those two beneficiaries for completeness.) The program enrolled a handful of beneficiaries after March 31, 2015, but we did not include them in the treatment group because when the program ended on June 30, 2015, they had been exposed to the program for only a short time and might not yet have experienced program impacts. According to one of CSHP's reports to CMMI, about 26 percent of program enrollees (through March 31, 2015) were enrolled in Medicare FFS, either alone or dually with Medicaid.

We limited the analysis sample to those continuously enrolled in FFS Medicare and observable in Medicare data during the four quarters before their program enrollment (the baseline period). We did this to make it easier to match treatment beneficiaries to potential comparison beneficiaries. Continuous enrollment ensured that we had a complete record of beneficiaries' service use in the year before program enrollment.

The treatment group included participants at all four CSHP program sites: Metro Community Provider Network in Aurora, Colorado; MultiCultural Primary Care Medical Group in San Diego, California; Neighborhood Health Centers of the Lehigh Valley in Allentown, Pennsylvania; and Truman Medical Center in Kansas City, Missouri.

Additional sample restrictions in each quarter. To be included in the analytic sample in any given quarter, each treatment group member had to meet two additional criteria. First, because we defined our evaluation outcomes quarterly (described in Section IV.A.4), and with

the intervention quarters specified relative to each beneficiary's enrollment date, we required that the last intervention quarter ended no later than August 31, 2015. We included July and August in the evaluation period, even though HCIA funding ended in June 2015, because the program expected to affect evaluation outcomes beyond the period of services receieved. Ending at August 31, 2015, enabled us to use all our available claims data for this report through September 30, 2015, still allowing one month of data beyond August 31 to observe 30-day unplanned readmissions. Second, the beneficiary's outcomes had to be observable in Medicare claims for at least one day during the quarter. Outcomes were observable for beneficiaries who were enrolled in Medicare FFS (Part A and B), were alive, and had Medicare as their primary payer (including beneficiaries who were dually eligible for Medicaid).

3. Comparison group definition

We constructed a comparison group of Medicare FFS beneficiaries who were similar to the treatment group beneficiaries on observable characteristics during the baseline period (that is, the four quarters before enrollment). This section describes how we constructed the matched comparison group and Section IV.C shows the balance we achieved between the two groups on selected matching variables. The multistage matching technique used for this report adds several additional steps to the approach we used in the second annual report (Bradley et al. 2015). The main difference is that we used an expanded set of matching variables to better predict treatment status in our propensity-score models. As a result, treatment beneficiaries previously included in our second annual report might have new matched comparisons in this report. Because we also added treatment beneficiaries who had enrolled from July 1, 2014, to March 31, 2015, we also expanded the comparison pool to beneficiaries who could have enrolled about the same time as these new treatment beneficiaries.

We constructed the comparison group through three steps:

First, we identified a pool of *potential* comparison members among Medicare FFS beneficiaries who met the minimum claims-based criteria for CSHP program eligibility that all participants met. These criteria were that they had (1) at least one of 25 chronic conditions we could observe in claims data; and (2) at least one outpatient ED visit or hospital discharge at some point from November 1, 2012, to March 31, 2015. We further limited this pool to those whose hospital discharge status was within the set of discharge statuses we observed among treatment beneficiaries and whose zip codes in the Medicare Enrollment Database indicated residence in geographic areas that either included the treatment group or were similar in size and composition to geographic areas in which the treatment group resided. However we did not require that comparison beneficiaries be matched only to treatment beneficiaries who lived in the same zip code or state. The comparison pool zip codes covered the following cities or counties in each state:

- In Pennsylvania: Allentown, Bath, Bethlehem, Emmaus, Lancaster, Macungie, Nazareth, Northampton, Reading, Red Hill, and Scranton

- In Colorado: Adams, Arapahoe, Bent, Boulder, Denver, Douglas, El Paso, Elbert, Fremont, Gilpin, Jefferson, Larimer, Lincoln, Logan, Morgan, Otero, Pueblo, Teller, and Weld counties
- In California: San Diego and Los Angeles
- In Missouri: Kansas City and St. Louis

Second, for each potential comparison beneficiary, we created a *pseudo-enrollment date* to approximate the date the beneficiary would have enrolled in the intervention if he or she had been in the treatment group. The pseudo-enrollment date was drawn to correspond with CSHP enrollment dates. Specifically, for each potential comparison beneficiary, we randomly added a number of days to the ED visit or discharge date that triggered the beneficiary's eligibility for the treatment or comparison group to get the pseudo-enrollment date. We drew the number of days from a frequency distribution of days from treatment beneficiaries' last ED visit or discharge before program enrollment to program enrollment. If a potential comparison beneficiary was discharged multiple times from November 1, 2012, to March 31, 2015, we considered the beneficiary as a potential comparison beneficiary at each pseudo-enrollment date. However, we ensured that a beneficiary could not be in the comparison group more than once. We then limited the comparison pool to those continuously enrolled in FFS Medicare and observable in claims data during the four (baseline) quarters before their pseudo-enrollment date, consistent with the treatment group.

Third, we used the Enrollment Database and claims in the 12 to 36 months before program enrollment (treatment group) or pseudo-enrollment (potential comparison group) to develop baseline characteristics to use to then develop a set of matched comparison beneficiaries similar to treatment beneficiaries on observed baseline characteristics. Matching aims to reduce selection bias in observational studies by selecting comparison beneficiaries from the pool who are roughly equivalent to the treatment group across key, observable baseline characteristics. The goal of matching is to achieve baseline equivalence between the treatment and matched comparison groups on the variables in the matching process (Stuart 2010). However, in addition to claims-based eligibility criteria, the CSHP program selected treatment group members based on their perceived willingness to change and social support—factors that are not observable in claims data. We could not incorporate these factors into matching. Thus it is possible the treatment and comparison beneficiaries might have differed on unobservable characteristics, even if they were well matched on observable characteristics.

Exact matching. For CSHP, we used exact matching to stratify the sample by three characteristics. First, we exact matched on whether the original reason for Medicare entitlement was old age or something else (that is, disability and end-stage renal disease [ESRD]) because 85 percent of the treatment beneficiaries were originally entitled to Medicare due to disability or ESRD. These beneficiaries were younger and had different patterns of service use on average than beneficiaries entitled due to old age. Second, we exact matched on whether the beneficiary was discharged to hospice. This is important because fewer than 1 percent of treatment beneficiaries were discharged to hospice and we did not want the comparison group to include disproportionately more beneficiaries discharged to hospice, whose health care utilization,

spending, and mortality trajectories likely differed from nonhospice enrollees. Third, we exact matched on (pseudo-) enrollment month to minimize differential attrition over intervention quarters between treatment and matched comparison beneficiaries due to different lengths of time observable in the Medicare data. We did not exact match on state of residence due to difficulty achieving good matches on this variable; however, as we describe next, we matched on state in the propensity-score model.

Propensity-score matching. Within each of the groups created by the three exact match variables, we applied a two-stage matching approach. First, given the large size of the comparison pool (roughly 1.65 million beneficiaries), we used the nearest-neighbor matching approach to first narrow the comparison pool based on service use (ED visits, inpatient admissions, and unplanned readmissions) and Medicare spending during the baseline quarters. Second, we used propensity-score matching to match treatment to comparison beneficiaries on these same measures of service use and spending and other variables, including demographic characteristics, state of residence, zip code-level poverty rate, Medicare-Medicaid dual enrollment status, health status, and chronic conditions.

Within the family of propensity-score matching methods, we implemented a technique called full matching to form matched sets that contained one treatment and one or more comparison beneficiaries. Pair matching, in contrast, would have matched one treatment to one comparison beneficiary. The important benefit of full matching is that it achieves maximum bias reduction on observed matching characteristics and, subject to this constraint, maximizes the size of the comparison sample (Rosenbaum 1991; Hansen 2004). We matched each treatment beneficiary to up to 10 beneficiaries from the potential comparison group to create a more stable comparison group against which to compare the treatment group's experiences.

Additional sample restrictions in each quarter. To be included in the analytic sample, a matched comparison beneficiary had to meet the same additional criteria as the treatment group members—that is, the end of the last intervention quarter had to be no later than August 31, 2015, and the beneficiary had to be observable in Medicare claims for at least one day of the quarter.

4. Construction of outcomes and covariates

We used Medicare claims from November 1, 2009, to September 30, 2015, for beneficiaries assigned to the treatment and comparison groups to develop two types of variables: (1) outcomes, defined for each person in each baseline or intervention quarter; and (2) covariates, which describe a beneficiary's demographic, Medicare enrollment-related, and health-related characteristics during four baseline quarters for use as control variables in the regression models. As noted earlier, we defined the quarters relative to the beneficiary's enrollment or pseudo-enrollment date. Control variables were measured during the baseline period to avoid the potential bias that could occur if the intervention affected both control variables and outcomes. For example, the intervention might result in greater contact with the health system and earlier diagnoses of diseases and conditions, which could affect both health-related characteristics and outcomes. If we adjusted for changes in health-related status during the intervention period, we

might adjust away part of the impact of the intervention. Appendix 1 provides details on the methods we used to construct these variables.

Outcomes. For each person, we calculated seven outcomes grouped into four domains:

1. Domain: Quality-of-care processes
   a. Ambulatory-care follow-up visit within 14 days of a hospital discharge (binary variable for each beneficiary); calculated as whether all of an individual's discharges in a quarter were followed by an ambulatory visit with a primary care or specialist physician within 14 days of the discharge
2. Domain: Quality-of-care outcomes
   a. Inpatient admissions for ambulatory care-sensitive conditions (ACSCs, number/quarter)
   b. Number of inpatient admissions followed by an unplanned readmission within 30 days (number/quarter)
3. Domain: Service use
   a. All-cause inpatient admissions (number/quarter)
   b. Outpatient ED visit rate (number/quarter); outpatient ED visits are defined as ED visits or observational stays that do not end in a hospital admission
4. Domain: Spending
   a. Total Medicare Part A and B spending ($/month)
   b. Medicare inpatient spending ($/month)

Four of these outcomes—30-day unplanned readmissions, all-cause inpatient admissions, outpatient ED visits, and total Medicare spending—are outcomes that CMMI has specified as core for the evaluations of all HCIA programs. Our definition of the readmission measure, however, differs from CMMI's standard definition. CMMI typically defines readmissions as the proportion of inpatient admissions that end in an unplanned readmission. Instead, we analyzed impacts on the *number* of these unplanned readmissions across all beneficiaries per quarter because this enabled us to examine the total impact on readmissions across the treatment group, rather than readmissions contingent on an inpatient admission. We made this decision, in consulation with CMMI, because the intervention might also affect the number of and type of admissions.

All outcomes are quarter-specific—meaning that we calculated them for each baseline and intervention quarter separately. Also, we defined all outcomes for all treatment and comparison group members, except for the quality-of-care processes measure. We calculated the measure of 14-day follow-up post-discharge among only those beneficiaries with at least one hospital discharge in the relevant quarter.

Covariates. The covariates, defined at the enrollment (treatment group) or pseudo-enrollment date (comparison group) include (1) measures of chronic conditions created by

applying Chronic Condition Warehouse algorithms to claims in the 12 to 36 months (depending on the condition) before the beneficiary's enrollment or pseudo-enrollment date, including the number of major chronic conditions (among 25 mostly physical health conditions) and 6 specific chronic conditions (Alzheimer's disease, cancer, chronic kidney disease, chronic obstructive pulmonary disease, congestive heart failure, and diabetes); (2) the number of mental health conditions (out of 6); (3) Hierarchical Condition Category (HCC) score, which is a continuous score that CMS developed to predict a beneficiary's future Medicare spending; (4) ED visits, inpatient admissions, ACSCs, total spending, and inpatient spending in the baseline period; (5) number of unplanned readmissions and the percentage of discharges followed by a primary care office visit within 14 days; (6), whether the hospital visit before enrollment was an outpatient ED visit or inpatient stay; (7) discharge status; (8) an indicator for dual Medicare and Medicaid enrollment; (9) demographics (age, gender, and race and ethnicity); and (10) the 2012 zip code-level poverty rate in the beneficiary's home zip code.

5. Regression model

We used a regression model to implement a contemporaneous differences analysis. For each quarter-specific outcome, the model estimated the relationship between the outcome and the covariates described earlier and a series of quarter-specific intervention indicator variables for whether the beneficiary was in the treatment group. The estimated relationship between the quarter-specific treatment indicator and outcomes measured the average difference in outcomes for beneficiaries in the treatment and comparison groups in that quarter, while controlling for any differences in outcomes associated with differences in the covariates.

We designed the model to measure differences in treatment and comparison group outcomes separately for each quarter, because it is possible that the program's impacts had changed since the beneficiary first received program services. We could also examine differences over discrete sets of quarters, which was needed to implement the primary tests discussed in the next section. Finally, the model quantified the uncertainty in the estimates, allowing for statistical tests that determine whether observed differences in outcomes between the treatment and comparison groups are likely due to chance. Appendix 2 provides details on the regression methods, including descriptions of the weights used in the model and how the regressions account for correlation in outcomes across quarters for a given individual.

6. Primary tests

Table IV.1 shows the primary tests for CSHP, by domain. Each test specified a population, outcome, time period, expected direction of effect, and threshold that we counted as substantively important. The purpose of these primary tests was to focus the impact evaluation on hypotheses that could provide the most robust evidence about program effectiveness (see Appendix 3 for detail and a description of how we selected each test). We provided both the awardee and CMMI an opportunity to comment on the primary tests.

Our rationale for selecting these primary tests is as follows:

- **Outcomes.** Based on CSHP's theory of action, we specified primary tests in four domains in which we expected the program to have an effect. In the quality-of-care processes domain, we included one measure for receipt of a follow-up ambulatory care visit with a primary care or specialist provider within 14 days of hospital discharge. In the quality-of-care outcomes domain, we expected the program to reduce admissions for ACSCs and 30-day unplanned readmissions. In the service use domain, we expected the program to reduce all-cause admissions and ED visits. Finally, given the expected reduction in service use, CSHP expected to reduce spending by 14.8 percent. Given that, in the spending domain we expected the program to reduce both inpatient hospital spending and total Medicare Part A and B spending.

Table IV.1. Specification of the primary tests for CSHP

Domain (number of tests in the domain)[a]	Outcome (units)	Time period for impacts (controlling for baseline differences)[b]	Population	Substantive threshold[c] (expected direction of effect)
Quality-of-care processes (1)	All inpatient admissions within a quarter were followed by an ambulatory care visit with a primary care or specialist provider within 14 days (binary [yes or no]/beneficiary/year)	Average over intervention quarters 1 through 4[d]	Medicare FFS beneficiaries in the treatment group with at least one hospital stay in the quarter	15.0 (+)
Quality-of-care outcomes (2)	Inpatient admissions for ambulatory care-sensitive conditions (#/beneficiary/quarter)	Average over intervention quarters 1 through 4[d]	Medicare FFS beneficiaries in the treatment group	15.0 (-)
	30-day unplanned hospital readmissions (#/beneficiary/quarter)	Average over intervention quarters 1 through 4[d]	Medicare FFS beneficiaries in the treatment group	15.0 (-)
Service use (2)	All-cause inpatient admissions (#/beneficiary/quarter)	Average over intervention quarters 1 through 4[d]	Medicare FFS beneficiaries in the treatment group	15.0 (-)
	Outpatient ED visit rate (#/beneficiary/quarter)	Average over intervention quarters 1 through 4[d]	Medicare FFS beneficiaries in the treatment group	15.0 (-)
Spending (2)	Medicare Part A and B spending ($/beneficiary/month)	Average over intervention quarters 1 through 4[d]	Medicare FFS beneficiaries in the treatment group	11.0 (-)
	Medicare inpatient spending ($/beneficiary/month)	Average over intervention quarters 1 through 4[d]	Medicare FFS beneficiaries in the treatment group	15.0 (-)

[a] We adjust the *p*-values from the primary test results for the multiple comparisons made within each domain, but not across domains.

[b] The regression models control for differences in characteristics and outcomes between the treatment and comparison groups during the baseline year when estimating program impacts.

[c] The substantive threshold is the impact as a percentage of the counterfactual. The counterfactual is the outcome the treatment group would have had in the absence of the HCIA-funded intervention.

[d] To implement the primary tests, we take the average of the regression adjusted estimates for intervention quarters 1 through 4.

CSHP = Center for State Health Policy; ED = emergency department; FFS = fee-for-service; HCIA = Health Care Innovation Award.

- **Time period.** We expected reductions in outcomes across all domains to be largest during program participation and that they could become harder to identify over time as the health of the treatment and comparison group members evolved. Because the length of the intervention varied to accommodate patients' needs (the intervention lasted 4.2 months post-discharge, on average, but in some cases was much longer, up to 22 months), we chose to specify our primary tests based on outcomes in the 12 months following a participant's enrollment date (that is, intervention quarters 1 through 4 [I1 to I4]). To implement each primary test, we took the average of the regression-adjusted estimates across the four quarters (I1 to I4) for that outcome.
- **Population.** CSHP's program sought to influence outcomes across all domains for all program enrollees. There was no program subgroup CSHP identified as expected to have different program impacts from other enrollees. Therefore, our primary tests included all (observable) Medicare FFS beneficiaries. Although CSHP did enroll patients with non-FFS Medicare coverage, such as patients with Medicaid and patients without any insurance coverage, we have no data on them.
- **Direction (sign) of the impact estimate.** For the quality-of-care process measure, we expected the impact estimate to be positive, signaling an increase in the percentage of people receiving follow-up care after a hospital discharge. For all other outcomes, we expected the impact estimates to be negative, indicating a reduction in service use or overall expenditures.
- **Substantive thresholds.** Some impact estimates could be large enough to be policy relevant (to CMMI and other stakeholders) even if they are not statistically significant; for this reason, we pre-specified thresholds for what we call substantive importance. We expressed the substantive threshold as a percentage change from the counterfactual—that is, the outcomes that beneficiaries in the treatment group would have had if they had not received the treatment. The decline of 11 percent that we chose for substantive importance for total Medicare spending is 75 percent of CSHP's anticipated impact on spending. (We used 75 percent recognizing that CSHP could still be considered successful if it came close to, but did not achieve, its fully anticipated effects.) We extrapolated the 15 percent threshold for all other outcomes from the literature (Peikes et al. 2011; Rosenthal et al. 2016) because CSHP did not specify by how much it expected to improve these outcomes.

7. Synthesizing evidence to draw conclusions

Within each domain, we drew one of five conclusions about program effectiveness, based on the primary test results and the plausibility of those findings given the implementation evidence:

1. Statistically significant favorable effect (the highest level of evidence)
2. Substantively important (but not statistically significant) favorable effect
3. Substantively important (but not statistically significant) unfavorable effect
4. No substantively large effect
5. Indeterminate effect

We cannot conclude that a program had a statistically significant unfavorable effect because, in consultation with CMMI, we decided to use one-sided statistical tests (which do not test for evidence of unfavorable effects). We used one-sided tests to increase the probability that, if a program truly did have impacts, we would be able to detect them.

Appendix 3 describes our decision rules for each of the five possible conclusions. In short, we concluded that a program had a statistically significant favorable effect in a domain if (1) at least one primary test result in the domain was favorable and statistically significant, after adjusting the statistical tests to account for multiple tests (if applicable) within a domain; or (2) the average impact estimate across all primary tests in the domain was favorable and statistically significant. In both cases, we also had to determine that the primary test results were plausible given the implementation evidence. We concluded that a program had a substantively important favorable effect if the average impact estimate in the domain was substantively important but not statistically significant, and if the result was plausible given the implementation evidence. In contrast, if the average impact estimate was unfavorable (opposite the hypothesized direction), larger than the substantive threshold, and unfavorable effects were plausible given the other evidence, we concluded the program had a substantively important unfavorable effect. If the tests in a domain met none of these criteria, we drew one of two conclusions. First, if the tests for at least one outcome in the domain (or all outcomes in the domain together) had sufficient statistical power to detect an impact of the size of the substantive threshold with at least 75 percent probability, we concluded that there was not a substantively large effect because we were reasonably confident that we would have detected a substantively large effect had there been one. Alternatively, if the power was not sufficient to detect this type of impact, we concluded the impact in the domain was indeterminate. Indeterminate means either that the program truly did not have effects that were substantively large, or that it did, but our statistical tests were not able to detect them.

B. Characteristics of the treatment group at baseline

This section describes the characteristics of the treatment group at the start of the intervention (January 2013). We also show this information in the second column of Table IV.2. (Table IV.2 serves a second purpose—to show the equivalence of the treatment and comparison groups at the start of the intervention—which we describe in Section IV.C)

The characteristics of the treatment group were consistent with CSHP's target population—frequent users of hospital care, typically with two or more inpatient hospital or ED visits in the prior six months. The mean HCC risk score of 3.9 was nearly four times the national average (1.0), indicating that the treatment group could be expected to have Medicare spending 3.9 times the national average over the subsequent year. The treatment group members typically had multiple chronic conditions, with an average of 7.7 chronic conditions and 1.4 mental health conditions. A high percentage had diabetes (70.5), chronic kidney disease (64.4), chronic obstructive pulmonary disease (57.1), or congestive heart failure (56.4). These condition-specific rates were each two to five times the national average. In the 12 months before program enrollment, Medicare spending averaged $69,960, almost seven times the national average of $10,320. The mean number of hospitalizations and ED visits in the 12 months before program

enrollment was 4.1 and 5.5, respectively; these were also more than 10 times the national averages.

The treatment group was also distinct from the average Medicare population along demographic characteristics and reason for Medicare eligibility. The mean age of the treatment group was 58 years, whereas the Medicare FFS average age was 71 years. It is not surprising then that the original reason for Medicare entitlement was disability or ESRD for 85 percent of the treatment group. Slightly more than half of the treatment group was black, compared with 10 percent of the Medicare FFS population. The average poverty rate in the zip codes listed in treatment group members' enrollment data was 26.2 percent, which is almost twice the national poverty rate in 2012 of 15 percent (DeNavas-Walt et al. 2013).

C. Equivalence of treatment and comparison groups at baseline

Demonstrating that the treatment and comparison groups were similar at the start of the intervention is important for the evaluation design. This similarity increases the credibility of a key assumption underlying contemporaneous differences models—that the outcomes observed for the comparison group during the intervention period are the same, on average, as the outcomes that would have been observed for the treatment group, had the treatment group not received the intervention.

Table IV.2 shows that the treatment and matched comparison groups were remarkably similar at the start of the intervention on most matching variables. (The second column of the table shows the unmatched comparison pool, which was generated from the nearest-neighbor matching we used to narrow the pool to to those similar to the treatment group, and the third column shows the matched comparion pool.) By construction, there were no differences between the two groups on the exact matching variables. There were some slight differences between treatment group beneficiaries and matched comparison group beneficiaries on the variables we matched through propensity scores, but the standardized differences across the propensity-score matching variables were all within our target of 0.25 standardized differences. All but three variables were within 0.05 standardized differences, and all were within 0.10 standardized differences. (The 0.25 target is an industry standard; for example, see Institute of Education Sciences [2014]).

Our full matching process substantially improved the balance for most variables compared with the full, unmatched comparison pool of 1,649,380 beneficiaries (results not shown). This improvement was very important given how the treatment population differed from the national Medicare FFS population, as discussed previously. Although we placed a number of restrictions to limit the comparison pool to those more likely to be eligible for the CSHP program, Table IV.2 shows that the unmatched (but restricted) comparison pool was still quite different from the treatment group.

Table IV.2. Characteristics of treatment and comparison groups at baseline for CSHP

Characteristic	Treatment group (n = 149)	Unmatched comparison group (n = 2,926)	Matched comparison group (n = 1,130)	Absolute difference[a]	Standardized difference[b]	Medicare FFS average
Exact match variables[c]						
Original reason for entitlement is disability or ESRD (%)	85.2	55.9	85.2	0	0	NA
Discharged to hospice (%)	0.7	0.003	0.7	0	0	NA
Propensity matched variables[d]						
Demographic characteristics						
Age (years)	58.5	65.2	59.0	-0.5	-0.036	71[e]
Female (%)	45.6	42.2	43.4	2.2	0.045	54.7[f]
Race: Black (%)	51.7	16.8	48.0	3.7	0.074	10.4[f]
Race: Hispanic (%)	8.1	4.7	8.5	-0.4	-0.018	2.8[f]
Zip code poverty rate (%)	26.2	16.0	25.5	0.7	0.065	15.0[g]
Medicare-related characteristics						
Dual status at enrollment (%)	70.5	39.2	69.2	1.3	0.027	21.0[h]
Health status and chronic conditions						
HCC risk score	3.9	2.3	3.8	0.1	0.037	1.0
Chronic conditions (# out of 25)[i]	7.7	5.6	7.6	0.1	0.036	NA
Mental health conditions (# out of 6)[j]	1.4	1.1	1.4	0.0	0.008	NA
Alzheimer's (%)	7.4	12.1	8.7	-1.3	-0.054	4.9[k]
Cancer (%)	5.4	11.3	5.3	0.1	0.002	NA
CHF (%)	56.4	27.0	56.0	0.4	0.008	15.3[k]
CKD (%)	64.4	34.8	63.0	1.4	0.029	16.2[k]
COPD (%)	57.1	27.8	56.2	0.9	0.017	11.8[k]
Diabetes (%)	70.5	37.8	71.2	-0.7	-0.015	28.0[k]
Service use and spending 6 months before enrollment						
All-cause inpatient admissions (#/beneficiary/6 months)	2.7	0.9	2.7	-0.0	-0.021	0.148[l]
Inpatient admissions for ambulatory care-sensitive conditions (#/beneficiary/6 months)	0.8	0.2	0.9	-0.1	-0.005	NA
Outpatient ED visits (#/beneficiary/6 months)	3.2	2.3	3.3	-0.1	-0.011	0.210[m]
Medicare Part A and B spending ($/beneficiary/6 months)	42,940	18,591	43,337	-397	-0.011	5,160[n]
Medicare FFS inpatient spending ($/beneficiary/6 months)	29,120	9,804	28,978	142	0.005	2,610[n]

Table IV.2 *(continued)*

Characteristic	Treatment group (n = 149)	Unmatched comparison group (n = 2,926)	Matched compari-son group (n = 1,130)	Absolute difference[a]	Standard-ized difference[b]	Medicare FFS average
Service use and spending 12 months before enrollment						
30-day unplanned hospital readmissions (#/beneficiary/11 months)	1.5	0.3	1.5	0.0	0.008	NA
Hospitalizations followed by 14-day follow-up (%)	45.4	62.4	45.7	-0.3	-0.008	NA
All-cause inpatient admissions (#/beneficiary/year)	4.1	1.4	4.0	0.1	0.018	0. 296[l]
Inpatient admissions for ambulatory care-sensitive conditions (#/beneficiary/year)	1.2	0.3	1.2	0.0	0.014	NA
Outpatient ED visits (#/beneficiary/year)	5.5	3.6	5.5	-0.0	-0.002	0.420[m]
Medicare Part A and B spending ($/beneficiary/year)	69,960	29,882	69,831	129	0.002	10,320[n]
Medicare FFS inpatient spending ($/beneficiary/year)	45,627	14,934	44,705	922	0.022	5,230[n]
Characteristics of trigger event (%)						
Inpatient admission (%)	65.1	32.4	63.7	1.4	0.03	NA
Days between discharge and (pseudo-) enrollment	18.7	17.7	18.8	-0.1	-0.01	NA
Discharged to home	91.3	92.8	92.3	-1.0	-0.042	NA
Discharged to facility	8.1	7.2	7.0	0.9	0.043	NA
Site (%)						
Neighborhood Health Centers of Lehigh Valley	23.5	17.3	23.8	-0.3	-0.007	NA
Truman Medical Center	23.5	36.4	25.3	-1.8	-0.041	NA
MultiCultural Medical Group	2.7	6.6	2.2	0.5	0.027	NA
Metro Community Provider Network	50.3	39.7	48.7	1.6	0.033	NA

Sources: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at CMS. Zip code poverty rate merged from the 2012 Five-Year American Community Survey ZIP Code Characteristics.

Notes: Characteristics are measured at the time of enrollment (for the treatment group) or pseudo-enrollment (for the potential and matched comparison groups). The matched comparison group means are weighted based on the number of matched comparisons per treatment beneficiary. For example, if four comparison beneficiaries are matched to one treatment beneficiary, the four comparison beneficiaries each have a matching weight of 0.25.

The unmatched comparison group shown is the group that came out of the first stage of matching, which used nearest-neighbor matching to narrow the pool and make it much more similar to the treatment group than the initial pool of potential comparisons.

The chronic condition flags are calculated using one to three years of claims before the enrollment or pseudo-enrollment date (depending on the condition), using the Chronic Conditions Warehouse definitions. The flags for Alzheimer's-related disorders and senile dementia used a look-back period beginning three years before enrollment.

Absolute differences might not be exact due to rounding.

[a] The absolute difference is the difference in means between the treatment and matched comparison groups.

[b] The standardized difference is the difference in means between the treatment and comparison groups divided by the standard deviation of the variable, which is pooled across the treatment and comparison groups.

[c] Variables for which we required treatment and comparison members to match on exactly.

Table IV.2 *(continued)*

[d] Variables that we matched on through a propensity score, which captures the relationship between beneficiaries' characteristics and their likelihood of being in the treatment group.

[e] Health Indicators Warehouse (2014a).

[f] Chronic Conditions Warehouse (2014a, Table A1).

[g] DeNavas-Walt et al. (2013).

[h] We estimated the Medicare FFS average using the percentage of Medicare beneficiaries who were dually eligible in 2011; 2010 MSIS data were used for Florida, Kansas, Maine, Maryland, Montana, New Mexico, New Jersey, Oklahoma, Texas, and Utah, and then adjusted to 2011 CMS-64 spending levels (Kaiser Family Foundation State Health Facts 2016).

[i] We use 25 of the 27 chronic condition categories defined by the Chronic Conditions Warehouse (Chronic Conditions Data Warehouse 2016). We exclude the Alzheimer's disease and the acute myocardial infarction flags because other flags include these conditions.

[j] The six mental health conditions are conduct disorders and hyperkinetic syndrome, anxiety disorder, bipolar disorder, personality disorders, schizophrenia and other psychotic disorders, and depressive disorders, as defined by the Chronic Conditions Warehouse (Centers for Medicare & Medicaid Services 2013).

[k] Chronic Conditions Warehouse (2014b, Table B2).

[l] Health Indicators Warehouse (2014b).

[m] Gerhardt et al. (2014).

[n] Boards of Trustees (2013).

CHF = congestive heart failure; CKD = chronic kidney disease; CMS = Centers for Medicare & Medicaid Services; COPD = chronic obstructive pulmonary disease; ED = emergency department; ESRD = end-stage renal disease; FFS = fee-for-service; HCC = Hierarchical Condition Category; MSIS = Medicaid Statistical Information System.

NA = not available.

Compared with the matched comparison group presented in the second annual report (Bradley et al. 2015), the current matched comparison group was better balanced to the treatment group. For example, the standardized differences between the current treatment and comparison groups were slightly lower and more of the measures were within 0.05 standardized differences than with the original groups in the second annual report.

D. Beneficiaries' outcomes and intervention impacts

This section first presents sample sizes and mean outcomes, by quarter, for the treatment and comparison groups. These mean outcomes provide context for understanding the contemporaneous differences estimates that follow; however, the differences in mean outcomes are not regression-adjusted and not impact estimates by themselves. Next, we present the results of the primary tests, by domain. Then, we assess whether the primary test results are plausible given the implementation evidence. We end with conclusions about program impacts in each domain. These conclusions are final because this report covers the full HCIA funding period for CSHP.

1. Sample sizes

The sample sizes for impact estimation differed by domain. The one measure of quality-of-care processes, the 14-day follow-up measure, was defined among Medicare FFS beneficiaries who had at least one hospital stay in the quarter. For the treatment group, the sample size ranged from 43 to 73 beneficiaries across intervention quarters, accounting for about 36 to 49 percent of all treatment beneficiaries in each quarter (Table IV.3). For the comparison group, the sample ranged from 248 to 482 across intervention quarters, accounting for a similar proportion of the total comparison group. After weighting the comparison group to account for the larger number

of comparison than treatment beneficiaries, the comparison group sample sizes were still similar to those in the treatment group.

Sample sizes for the quality-of-care outcomes, service use, and spending domains were the same. In the first intervention quarter (I1), the treatment group included 149 treatment group beneficiaries and 1,130 comparison group beneficiaries (Table IV.4). The sample decreased by about 10 percent in each subsequent intervention quarter, as expected, because (1) some beneficiaries did not enroll or pseudo-enroll early enough to follow for a second, third, or fourth intervention quarter before the end of our evaluation period in August 2015; and (2) some treatment or comparison group members exited the sample due to death or becoming unobservable. The ratio of comparison group (n = 1,130) to treatment group (n = 149) beneficiaries in I1 (7.6) declined in each intervention quarter; the ratio of comparison group (n = 784) to treatment group (n = 113) beneficiaries in I4 was 6.9.

We found that differential mortality between the treatment and comparison groups caused this decline in the ratio of treatment to comparison beneficiaries. Specifically, the comparison group had slightly higher mortality. By the start of I2, 2 percent of the treatment and 6 percent of the comparison beneficiaries had died; 3 percent of the treatment and 8 percent of the comparison beneficiaries had died by the start of I3; and 4 percent of the treatment group and 11 percent of the comparison group had died by I4 (results not shown).

Higher mortality among the comparison group could present challenges for estimating impacts for two reasons. First, if this higher mortality is the result of favorable program impacts on survival—that is, the program helps to keep the sickest treatment group beneficiaries alive—it would likely cause a downward bias of estimates of program impacts on other outcomes. For example, the sicker treatment group members would likely have more admissions and higher spending on average than the surviving comparison beneficiaries. Second, the differential mortality could signal lack of equivalence at baseline on some unobservable factors correlated with mortality. For example, program staff might have preferentially enrolled beneficiaries who they thought were likely to live to participate fully in the program. The influence of any such unobservable differences on impact estimates is unclear. It could make any apparent impacts appear larger than they actually were, if treatment group members were healthier on average than comparison group members. However, unobservable differences could also bias estimates of program impacts downward, if the sickest patients in the comparison group dropped out of the sample (due to death) while relatively sick members of the treatment group remained in the sample.

2. Unadjusted mean outcomes for the treatment and comparison groups, by domain and quarter

Quality-of-care processes. For both the treatment and comparison groups, 33 to 46 percent of beneficiaries who had any hospital stays in a baseline quarter had all of those stays followed by an ambulatory care visit within 14 days of discharge. The proportion of treatment group beneficiaries with 14-day follow-up visits was higher in I1 and I3 than the proportion among comparison group beneficiaries (by 30 and 14 percent, respectively); the proportions were similar in I2 and I4 (Table IV.3).

Table IV.3. Sample sizes and unadjusted means for quality-of-care process measures for Medicare FFS beneficiaries in the treatment and comparison groups for CSHP, by quarter

	Number of Medicare FFS beneficiaries			Mean outcomes		
Quarter	T	C (not weighted)	C (weighted)	T	C	Difference (%)
Among those with at least one inpatient admission in the quarter, all inpatient admissions in the quarter were followed by an ambulatory care visit with a primary care or specialist provider within 14 days of discharge (binary [yes or no]/beneficiary/year)						
I1	73	482	67	42.5	32.8	9.7 (29.6%)
I2	49	351	51	38.8	38.0	0.8 (2.1%)
I3	44	283	40	45.5	39.8	5.7 (14.3%)
I4	43	248	34	37.2	37.0	0.2 (0.5%)

Source: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at CMS. See Table IV.2 for sources for the Medicare FFS averages.

Note: The quarters are 3-month periods after a beneficiary's enrollment date (treatment group) or pseudo-enrollment date (comparison group). That is, Intervention Quarter 1 is the first 3 months after enrollment or pseudo-enrollment, and Intervention Quarter 2 is month 4 to 6. The means are weighted: each treatment group beneficiary receives a weight of 1; each comparison beneficiary receives a weight equal to the reciprocal of the total number of comparison beneficiaries that match to the same treatment beneficiary. The sample includes beneficiaries whose enrollment or pseudo-enrollment date was between November 1, 2012 and March 31, 2015. The difference between the treatment and comparison groups in a quarter is calculated by subtracting the mean outcome for the comparison group from the mean outcome for the treatment group. The percentage difference equals that difference divided by the mean outcome for the comparison group.

C = comparison group; CMS = Centers for Medicare & Medicaid Services; Diff = difference; ED = emergency department; FFS = fee-for-service; T = treatment group.

NA = not available.

n.a. = not applicable.

Table IV.4. Sample sizes and unadjusted mean outcomes (quality-of-care outcomes, service use, and spending) for Medicare FFS beneficiaries in the treatment and comparison groups for CSHP, by quarter

	Medicare FFS average	Intervention quarter 1			Intervention quarter 2			Intervention quarter 3			Intervention quarter 4		
		T	C	Diff (%)	T	C	Diff (%)	T	C	Diff (%)	T	C	Diff (%)
Number of Medicare FFS beneficiaries (unweighted)	49 million	149	1,130	n.a.	134	987	n.a.	122	871	n.a.	113	784	n.a.
Inpatient admissions for ambulatory care-sensitive conditions (#/1,000 beneficiaries/ quarter)	NA	167.8	198.1	-30.3 (-15.3%)	223.9	195.4	28.5 (14.6%)	139.3	211.0	-71.7 (-34.0%)	221.2	179.1	42.1 (23.5%)
30-day unplanned hospital readmission rate (#/1,000 beneficiaries/ quarter)	NA	302.0	358.1	-56.1 (-15.7%)	231.3	327.8	-96.4 (-29.4%)	245.9	334.8	-88.9 (-26.6%)	177.0	326.6	-149.6 (-45.8%)
All-cause inpatient admissions (#/1,000 beneficiaries/ quarter)	74	825.5	758.2	67.3 (8.9%)	641.8	737.9	-96.1 (-13.0%)	647.5	776.8	-129.2 (-16.6%)	557.5	717.4	-159.9 (-22.3%)
Outpatient ED visit rate (#/1,000 beneficiaries/ quarter)	105	1,315.9	1,345.9	-30.0 (-2.2%)	1,588.2	1,091.6	496.6 (45.5%)	1,186.9	1,030.6	156.3 (15.2%)	924.3	1,159.1	-234.9 (-20.3%)
Medicare Part A and B spending ($/beneficiary/ month)	$860	$5,261	$5,714	$-454 (-7.9%)	$4,615	$5,197	$-582 (-11.2%)	$4,766	$5,009	$-242 (-4.8%)	$4,814	$4,970	$-155 (-3.1%)
Medicare FFS inpatient spending ($/beneficiary/ month)	NA	$3,113	$3,269	$-156 (-4.8%)	$2,717	$2,938	$-221 (-7.5%)	$3,104	$2,859	$245 (8.6%)	$3,098	$2,795	$303 (10.8%)

Source: Analysis of the Medicare Enrollment Database and claims data a.ccessed through the Virtual Research Data Center at CMS. See Table IV.2 for sources for the Medicare FFS averages.

Note: The quarters are three-month periods after a beneficiary's enrollment date (treatment group) or pseudo-enrollment date (comparison group). That is, intervention quarter 1 is the first three months after enrollment or pseudo-enrollment, and intervention quarter 2 is months four to six. The means are weighted: each treatment group beneficiary receives a weight of 1; each comparison beneficiary receives a weight equal to the reciprocal of the total number of comparison beneficiaries who match to the same treatment beneficiary. The sample includes beneficiaries whose enrollment or pseudo-enrollment date was from November 1, 2012, to March 31, 2015. The difference between the treatment and comparison groups in a quarter is calculated by subtracting the mean outcome for the comparison group from the mean outcome for the treatment group. The percentage difference equals that difference divided by the mean outcome for the comparison group.

C = comparison group; CMS = Centers for Medicare & Medicaid Services; CSHP = Center for State Health Policy; Diff = difference; ED = emergency department; FFS = fee-for-service; T = treatment group.
NA = not available.
n.a. = not applicable.

Quality-of-care outcomes. Among the treatment group, the number of ACSCs ranged from 139 to 224 per 1,000 beneficiaries per quarter in the four intervention quarters. This rate was lower than the comparison group rate of ACSCs in I1 and I3, but higher than the comparison group rate in I2 and I4 (Table IV.4).

The number of 30-day unplanned readmissions ranged from 177 to 302 per 1,000 beneficiaries per quarter for the treatment group. This rate was substantially lower than the rate among the comparison group in each quarter, with the difference between the rates ranging from 16 to 46 percent of the comparison group rate in each quarter.

Service use. All-cause inpatient admissions for the treatment group ranged from 558 to 826 per 1,000 beneficiaries per quarter, with the rate generally declining over time. The mean admission rate for the treatment group was 9 percent higher than for the comparison group in I1 but 13 to 22 percent lower than the comparison group in the three subsequent quarters.

Outpatient ED visits ranged from 924 to 1,588 per 1,000 beneficiaries per quarter for the treatment group and generally declined over each quarter. Relative to the comparison group rate, the treatment group outpatient ED visit rate was similar in I1, higher by 15 to 46 percent in I2 and I3, and lower by 20 percent in I4.

Spending. Total Medicare Part A and B spending in the treatment group averaged about $4,900 per beneficiary per month in the intervention quarters. In each quarter, this was 3 to 11 percent lower than total spending in the comparison group.

For both treatment and comparison groups, inpatient spending made up about $3,000 of total spending. Inpatient spending among the treatment group was lower than inpatient spending among the comparison group in I1 and I2 but higher than among the comparison group in I3 and I4.

3. Results for primary tests, by domain

Overview. The primary tests conducted for this report are final, as they cover the full primary test period (up to four intervention quarters for each beneficiary).

For three of the study domains—quality-of-care processes, service use, and spending—the regression-adjusted differences between the treatment and comparison groups were small (Table IV.5). None of these differences were statistically significant or larger than the substantive thresholds in either a favorable or unfavorable direction. However, Table IV.5 also shows that, in general, the tests had poor statistical power to detect effects of the size of the substantive thresholds.

In contrast, in the quality-of-care outcomes domain, we found statistically significant, substantively large, and favorable differences between the treatment and comparison groups. We estimated the largest impacts on the measure of 30-day unplanned hospital readmissions.

Table IV.5. Results of primary tests for CSHP

Primary test definition					Statistical power to detect an effect that is[a]		Results			
Domain (# of tests in domain)	Outcome (units)	Time period for impacts	Population	Substantive threshold (impact as a percentage relative to the counterfactual[b])	Size of the substantive threshold	Twice the substantive threshold[c]	Treatment group mean	Regression-adjusted difference between the treatment and estimated counterfactual[b] (standard error)	Percentage difference[d]	*p*-value[e]
Quality-of-care processes (1)	All inpatient admissions within a quarter were followed by an ambulatory care visit with a primary care or specialist provider within 14 days (binary [yes or no]/beneficiary/year)	Average over intervention quarters 1 through 4	Medicare FFS beneficiaries in the treatment group with at least one hospital stay in the quarter	+15.0%	49.0%	89.1%	41.0	3.6 (4.5)	9.7%	0.21
Quality-of-care outcomes (2)	Inpatient admissions for ambulatory care-sensitive conditions (#/1,000 beneficiaries/quarter)	Average over intervention quarters 1 through 4	Medicare FFS beneficiaries in the treatment group	-15.0%	37.9%	74.7%	188	-27 (33)	-12.4%	0.33
	30-day unplanned readmissions (#/1,000 beneficiaries/quarter)	Average over intervention quarters 1 through 4	Medicare FFS beneficiaries in the treatment group	-15.0%	29.7%	58.6%	239	-126* (73)	-34.4%	0.08
	Combined (%)	Average over intervention quarters 1 through 4	Medicare FFS beneficiaries in the treatment group	-15.0%	45.6%	85.6%	n.a.	n.a.	-23.4%**[f]	0.03
Service use (2)	All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	Average over intervention quarters 1 through 4	Medicare FFS beneficiaries in the treatment group	-15.0%	56.2%	94.4%	668	-116 (82)	-14.8%	0.14
	Outpatient ED visits (#/1,000 beneficiaries/quarter)	Average over intervention quarters 1 through 4	Medicare FFS beneficiaries in the treatment group	-15.0%	46.2%	86.2%	1253	57 (151)	4.8%	0.54
	Combined (%)	Average over intervention quarters 1 through 4	Medicare FFS beneficiaries in the treatment group	-15.0%	67.6%	98.6%	n.a.	n.a.	-5.0%	0.28

Table IV.5 *(continued)*

Primary test definition					Statistical power to detect an effect that is[a]		Results			
Domain (# of tests in domain)	Outcome (units)	Time period for impacts	Population	Substantive threshold (impact as a percentage relative to the counterfactual[b])	Size of the substantive threshold	Twice the substantive threshold[c]	Treatment group mean	Regression-adjusted difference between the treatment and estimated counterfactual[b] (standard error)	Percentage difference[d]	*p*-value[e]
Spending (2)	Medicare Part A and B spending ($/beneficiary/month)	Average over intervention quarters 1 through 4	Medicare FFS beneficiaries in the treatment group	-11.0%	54.2%	93.2%	$4,864	-$468 (423)	-8.8%	0.16
	Medicare inpatient spending ($/beneficiary/month)	Average over intervention quarters 1 through 4	Medicare FFS beneficiaries in the treatment group	-15.0%	50.1%	90.1%	$3,008	-$40 (355)	-1.3%	0.49
	Combined (%)	Average over intervention quarters 1 through 4	Medicare FFS beneficiaries in the treatment group	-13.0%	52.5%	92.0%	n.a.	n.a.	-5.0%	0.30

Sources: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at the Centers for Medicare & Medicaid Services.

Notes: The results for each outcome are based on a contemporaneous differences regression model that included one to four intervention quarter observations per beneficiary, as described in the text. For each quarter, the model calculates the regression-adjusted difference between outcomes for beneficiaries assigned to the treatment and comparison groups that quarter. The impact estimates from the four intervention quarters were averaged to obtain an average impact estimate across the four quarters. The quarters are 91- or 92-day increments after enrollment in the CSHP program for treatment group members, or after the pseudo-enrollment date for comparison beneficiaries. For example, if a treatment beneficiary was enrolled in the CSHP program on July 16, 2013, his or her first intervention quarter is July 16 through October 15, 2013; his or her second intervention quarter is October 16, 2013, through January 15, 2014. The estimates were adjusted for any differences in beneficiary-level covariates (defined in Section IV.A.4) at the beginning of the intervention period.

The treatment and comparison groups are limited to beneficiaries who were continuously enrolled in FFS Medicare for each of the four quarters before the enrollment or pseudo-enrollment date. Furthermore, in each intervention quarter, the sample consists of Medicare FFS beneficiaries who were (1) enrolled early enough to be potentially followed up for all 91 or 92 days in the quarter and (2) whose outcomes were observable for at least one day during the quarter. The sample includes those who were in the sample for at least one of the intervention quarters. Outcomes are observable if the beneficiary is alive, enrolled in Medicare Part A and B, not enrolled in a comprehensive managed care plan, and has Medicare as his or her primary payer of medical bills. Outcomes are constructed through August 31, 2015. In each regression model, comparison group beneficiaries are weighted based on the number of matched comparisons per treatment beneficiary. For example, if four comparison beneficiaries are matched to one treatment beneficiary, each of the four comparison beneficiaries has a weight of 0.25.

Table IV.5 *(continued)*

[a] The power calculation is based on actual standard errors from analysis. For example, in the second row, a 15 percent effect on inpatient admissions for ambulatory care-sensitive conditions (from the counterfactual of 188 + 27 = 215) would be a change of 32 admissions. Given the standard error of 33 from the regression model, we would be able to detect a statistically significant result only 37.9 percent of the time if the impact was truly 32 admissions, assuming a one-sided statistical test at the $p = 0.10$ significance level.

[b] The counterfactual is the presumed treatment group outcome in the absence of the HCIA-funded intervention. Our estimate of the counterfactual is the treatment group mean minus the regression-adjusted contemporaneous differences estimate.

[c] The substantive threshold is the impact as a percentage of the counterfactual. The counterfactual is the outcome the treatment group would have had in the absence of the HCIA-funded intervention. Our estimate of the counterfactual is the treatment group mean minus the regression-adjusted contemporaneous differences estimate.

[d] Percentage difference is calculated as the regression-adjusted contemporaneous differences estimate, divided by the estimate of the counterfactual.

[e] *p*-values test the null hypothesis that the regression-adjusted contemporaneous differences estimate is less than or equal to zero for outcomes in the quality-of-care processes domain, or greater than or equal to zero in all other domains (a one-sided test). Because it is a one-sided test, as the contemporaneous differences estimate approaches infinity in an unfavorable direction (negative for process-of-care measures and positive for all others), the *p*-value approaches 1, whereas it would approach 0 in a two-sided test. In each domain except quality-of-care processes, we adjusted the *p*-values from the primary test results for the multiple (two) comparisons made within the domain. We adjusted the *p*-values separately for the comparison made within the quality-of-care processes domain and (separately) for the two comparisons made within each of the other domains.

[f] The standard error for the combined percentage difference for the outcomes in the quality-of-care outcomes domain was 12.8 percentage points.

*/**/*** Significantly different from zero at the .10/.05/.01 levels, one-tailed test, respectively. No difference-in-differences estimates were significantly different from zero at the .01 level.

CSHP = Center for State Health Policy; ED = emergency department; FFS = fee-for-service; HCIA = Health Care Innovation Award.

n.a. = not applicable.

Quality-of-care processes. The likelihood of receiving an ambulatory care visit within 14 days of hospital discharge was 9.7 percent higher in the treatment group than its estimated counterfactual (the estimate of the counterfactual is the treatment group mean minus the regression-adjusted contemporaneous differences estimate), a favorable difference that was neither substantively large nor statistically significant ($p = 0.21$). The statistical power to detect substantively large effects on this outcome was poor (49 percent) for this measure.

Quality-of-care outcomes. The rate of ACSC admissions for the treatment group was 12.4 percent lower than the counterfactual. However, this difference in admissions was not statistically significant ($p = 0.33$, after adjusting for multiple statistical tests in the domain). The 30-day unplanned readmission rate for the treatment group was 34 percent lower than the counterfactual, and was statistically significant at the 10 percent level ($p = 0.08$, after adjusting for multiple statistical tests in the domain). After combining results across the two outcomes in this domain, the combined effect was a 23 percent favorable effect on quality-of-care outcomes, statistically significant at the 5 percent level ($p = 0.03$).

The statistical power to detect effects the size of the substantive threshold was poor for both ACSC admissions (37.9 percent) and 30-day unplanned readmissions (29.7 percent). Power was better, though still poor (45.6 percent), for the combined effect in the domain.

Service use. The treatment group's all-cause inpatient admission rate was 14.8 percent lower than the estimated counterfactual, which was close to the substantive threshold of 15 percent. The outpatient ED visit rate was 4.8 percent higher than the estimated counterfactual. Neither of these differences was statistically significant or substantively large. After combining results across the two outcomes in this domain, the combined effect was 5 percent in the favorable direction, but not statistically significant or substantively large, relative to the prespecified threshold. Power to detect effects that were the size of the substantive thresholds was marginal or poor for the admissions and outpatient ED visit measures individually (56.2 and 46.2, respectively) and marginal for the two outcomes combined (67.6 percent).

Spending. The treatment group averaged $4,864 per beneficiary per month in Part A and B spending, a value 8.8 percent (or $468) lower than the estimated counterfactual. The treatment group averaged $3,008 per beneficiary per month in inpatient spending, almost identical to ($40 lower than) the estimated counterfactual. Neither of these differences was statistically significant or substantively large.

The statistical power to detect effects the size of the substantive threshold was marginal for both Part A and B (54.2 percent) and inpatient spending (50.1 percent), as well as the combined effect in the domain (52.5 percent).

Aggregate estimates for CMMI's core measures. The estimates presented for the CMMI core outcomes—that is, for 30-day unplanned readmissions, all-cause inpatient admissions, the outpatient ED visit rate, and Medicare Part A and B spending—have so far been expressed per 1,000 beneficiaries per quarter (or, for spending, per beneficiary per month). Table IV.6 translates these estimates into estimates of aggregate impacts among Medicare beneficiaries

enrolled in the CSHP program during the year-long primary test period in this report. We calculated these aggregate impacts by multiplying the per beneficiary point estimates by the average number of Medicare beneficiaries in the treatment group and by the number of quarters or months during the primary test year. The statistically significant estimated reduction in 30-day unplanned readmissions translates to an aggregate reduction of 65 readmissions. The other aggregate estimates in Table IV.6 should be interpreted with caution because the estimates are not statistically significant. (The *p*-values for these aggregate estimates are the same as for the main results shown in Table IV.5).

Table IV.6. Results for primary tests for CMMI's core outcomes expressed as aggregate effects for all Medicare FFS beneficiaries in the treatment group

Outcome (units)	Aggregate impact estimate during the primary test year (I1 through I4)	*p*-value
30-day unplanned readmissions (#)	-65	0.08
All-cause inpatient admissions (#)	-60	0.14
Outpatient ED visits (#)	30	0.54
Medicare Part A and B spending ($)	-$727,580	0.16

Source: Authors' calculation, based on analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at the Centers for Medicare & Medicaid Services.

Note: To estimate the aggregate impact during the primary test year (intervention quarters 1 through 4) we (1) multiplied the per beneficiary per quarter (or month) estimate from Table IV.5 by the average number of Medicare FFS beneficiaries in the treatment group during the four primary test quarters then (2) scaled the estimate to a year by multiplying the resulting product by 4 (or 12). The *p*-values are taken from Table IV.5 and are therefore one-sided (testing that the program improved outcomes) and adjusted for multiple comparisons conducted within each outcome domain.

CMMI = Center for Medicare & Medicaid Innovation; ED = emergency department; FFS = fee-for-service.

4. Consistency of quantitative estimates with implementation findings

The impact estimates in the primary tests were plausible given the implementation findings. The primary test results showed favorable effects that were statistically significant for only one domain, quality-of-care outcomes; the rest of the domains did not have statistically significant or substantively large effects. The implementation evidence shows the program was active during the award period. For example, as described in Section III.B.3, on average care team staff provided nearly six hours of care management/care coordination services per participant per month (and 10.3 contacts per participant per month) to 1,068 frequent users of hospital care. However, even with a well-implemented intervention, it is possible that the program was not able to change beneficiaries' behaviors in ways that would affect service use and spending. The combination of favorable findings on quality-of-care outcomes and lack of substantive effects on service use and spending could also reflect the high clinical needs of the patients, shown in Table IV.2 and the care team's success in getting patients this needed care.

The statistically significant (favorable) effect in the quality-of-care outcomes domain was consistent with CSHP's aim to reduce unnecessary hospital-based care and the care teams'

efforts to help participants overcome obstacles to receiving needed medical and social services and to educate patients about appropriate use of primary and specialty care as alternatives to emergency and hospital care. We would have expected to see parallel improvement in the 14-day follow-up measure (in the quality-of-care processes domain), but the estimated effect on this outcome was not statistically significant or substantively large. It is possible the program had indeterminate effects on this domain because of two factors: (1) the 14-day window might have been too short for the participants to secure appointments with physicians after discharge, particularly given participants' barriers to receiving care; and (2) care management/care coordination activities that helped reduce readmissions could have occurred within 14 days, but were not observed in claims because they were conducted by nonclinical care team members who do not bill Medicare for services (which is how CSHP's program is designed).

5. Conclusions about program impacts, by domain

Based on all evidence currently available, we have drawn the following conclusions about program impacts during the the primary test period. Table IV.7 summarizes these conclusions and their support.

- **The program had an indeterminate effect on quality-of-care processes.** The primary test results were neither substantively large nor statistically significant. However, the statistical power was poor (45.6 percent) to detect effects the size of the substantive threshold. As a result, null findings from the primary test in this domain could be due to (1) the program truly not having a substantively large effect, or (2) the program having a substantively large effect but our tests failing to detect it.
- **The program had a statistically significant (and substantively large) *favorable* effect on quality-of-care outcomes.** The primary test results showed a statistically significant and substantively large favorable estimate for the quality-of-care outcomes domain, driven by a statistically significant and large favorable estimate for 30-day unplanned readmissions, in particular. Although the ACSC admissions outcome was not statistically significant on its own, it was in the favorable direction (lower for the treatment group than the matched comparison group), and the combined effect estimate across both outcomes was statistically significant and substantively important (that is, larger than the substantive threshold in the domain of 15 percent).
- **The program had an indeterminate effect on service use and Medicare spending.** The primary test results were neither substantively large nor statistically significant. However, the statistical power was marginal (67.6 percent for the combined effect in the service use domain and 52.5 percent for the combined effect on spending) to detect effects the size of the substantive threshold. As a result, as with the quality-of-care processes domain, null findings from the primary tests in these domains could be due to (1) the program truly not having substantively large effects or (2) the program having substantively large effects but our tests failing to detect them.

Table IV.7. Conclusions about the impacts of CSHP's HCIA program on patients' outcomes, by domain

		Evidence supporting conclusion	
Domain	Conclusion	Primary test result(s) that supported conclusion	Primary test result(s) plausible given implementation evidence?
Quality-of-care processes	Indeterminate	• No statistically significant or substantively important effect; power was marginal to detect an effect on the single outcome in the domain	Yes
Quality-of-care outcomes	Statistically significant favorable effect	• The impact estimate for 30-day unplanned hospital readmissions was statistically significant and the combined impact estimate across both outcomes in the domain was both statistically significant and larger than the substantive threshold	Yes
Service use	Indeterminate	• No statistically significant or substantively important effect; power was marginal to detect an effect on the single outcome in the domain	Yes
Spending	Indeterminate	• No statistically significant or substantively important effect; power was marginal to detect an effect on the single outcome in the domain	Yes

Source: Table IV.5.

CSHP = Center for State Health Policy; HCIA = Health Care Innovation Award.

V. DISCUSSION AND CONCLUSIONS

CSHP used its $14.3 million HCIA award to provide care management/care coordination services to frequent users of hospital services. Multidisciplinary, community-based care teams at four implementation sites connected participants to appropriate clinical and social services, helped them manage their conditions, and worked to overcome socioeconomic obstacles to care. These activites were intended to reduce patients' need for acute care and increase use of appropriate primary and specialty care. CSHP's goal was to reduce the average annual cost of care by 14.8 percent by the end of the award.

The results from our impact evaluation suggest that the intervention succeeded in improving quality-of-care outcomes for participants who were Medicare FFS beneficiaries. In particular, we observed a statistically significant 34.4 percent reduction in 30-day unplanned readmissions for Medicare FFS beneficiaries served by the four sites, relative to other Medicare FFS beneficiaries with similar characteristics. Improvement in the quality-of-care outcomes domain is plausible given the program's efforts to educate patients about appropriate use of primary and specialty care and to help them overcome barriers to care.

We could not determine whether the program had an effect on Medicare FFS beneficiaries' outcomes in the other three evaluation domains: quality-of-care processes, service use, and spending. The outcomes in these domains were not statistically or substantively better for

Medicare FFS patients served by the four sites than those for other Medicare patients with similar characteristics. The indeterminate effects in these three domains do not appear to be due to major problems implementing the intervention. Although each site faced unique implementation challenges, all four adapted and operated the program in a way that was consistent with the core design. This was reflected by several measures, including the following:

- The four sites hired a total of 35.3 FTE new care team and support staff (73 percent of their target), who provided care management/care coordination services throughout the award.
- The four sites provided care management/care coordination services to 1,068 participants (63 percent of their target).
- The four sites successfully established partnerships with local primary care providers, health plans, social service agencies, and community organizations to facilitate participant identification and coordination of medical, behavioral health, and social services. At the two sites that did not rely on hospital electronic health records for participant identification, these partners referred potential participants to the program; and at all four sites, these partnerships helped care team members connect enrollees to needed medical and social services.

However, it is difficult to say whether the lack of measured effects in the quality-of-care processes, service use, and spending domains means that the program did not have substantively important impacts in these three domains, or whether our evaluation failed to detect impacts that did occur. As shown in Table IV.5, the statistical power to detect substantively important effects was poor to moderate for all outcome. This means that even if the program did have impacts the size of the substantive threshold in one of these domains, we could have failed to detect it. Because we cannot say with confidence whether the lack of observed impacts on quality-of-care processes, service use, and spending was due to a lack of program impacts or to limitations in the evaluation, we considered both of the following possibilities.

Implementation findings suggest the lack of measurable effects in three of the outcome domains—all but quality-of-care outcomes—could be due to a combination of several factors that complicated the delivery of program services. Even though the program was generally implemented as planned, the 62 percent overall graduation rate and the initiation of post-graduation services at two of the sites reflected the difficulty of achieving long-term, independent behavioral change among program participants. The low graduation rate was consistent with reports from program staff that the complexity of participants' medical, social, psychological, cultural, and other needs impeded care teams' ability to help participants overcome obstacles to care. Our data on the baseline characteristics of the treatment group beneficiaries support this hypothesis. These beneficiaries had more than seven chronic conditions and an HCC risk score nearly four times the national average at program enrollment (Table IV.2). Program participants' needs were exacerbated by major social service and health system limitations, such as lack of safe and affordable housing. Together, these factors suggest that it might be difficult to change participants' behavior and health status enough to see measurable reductions in utilization and spending.

It is also possible, however, that data limitations or other constraints in the evaluation design could explain the indeterminate effects. For example, as noted previously, the evaluation was only poorly or marginally powered to detect effects in all four evaluation domains. Another potential limitation was that we were unable to match the treatment and comparison groups on important but unobservable factors, such as willingness to change and degree of social and family support for achieving change, which might be associated with evaluation outcomes or with the differential mortality observed between the treatment and comparison groups. It is unclear how differential mortality could have influenced the evaluation findings, if at all. Finally, treatment effects might have differed for certain populations. We were unable to report separate subgroup outcomes due to low statistical power—for example, by program site or among beneficiaries with different risk levels. Furthermore, because the evaluation was limited to Medicare FFS beneficiaries (including dually eligible beneficiaries), we could not include participants covered by Medicaid only, or those who were uninsured; the intervention might have had different effects on those participants but we were not able to assess whether that was the case.

This page has been left blank for double-sided copying.

REFERENCES

Boards of Trustees, Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds. "2013 Annual Report of the Boards of Trustees of the Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds." Table V.D1. Washington, DC: Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds, 2013. Available at http://downloads.cms.gov/files/TR2013.pdf. Accessed August 13, 2014.

Bradley, Katharine, Purvi Sevak, Cara Stepanczuk, Boyd Gilman, Greg Peterson, Catherine DesRoches, Sandi Nelson, Laura Blue, Keith Kranker, Kate Stewart, Frank Yoon, and Lorenzo Moreno. "Second Annual Report: Findings for Rutgers Center for State Health Policy." In Moreno, Lorenzo, Boyd Gilman, Greg Peterson, Catherine DesRoches, Sheila Hoag, Linda Barterian, Laura Blue, Katherine Bradley, Emily Ehrlich, Kristin Geonnotti, Lauren Hula, Keith Kranker, Rumin Sarwar, KeriAnn Wells, Joseph Zickafoose, Sandi Nelson, Kate Stewart, and Frank Yoon. "Evaluation of Health Care Innovation Awards (HCIA): Primary Care Redesign Programs." Second annual report to CMS. Volume II: Individual program summaries. Princeton, NJ: Mathematica Policy Research, December 11, 2015.

Centers for Medicare & Medicaid Services. "Attachment C: CMS CCW Task Order 10 Clinical Conditions Reference List." Baltimore, MD: CMS, 2013. Available at https://www.ccwdata.org/cs/groups/public/documents/document/clin_cond_refer.pdf. Accessed November 19, 2014.

Centers for Medicare & Medicaid Services. "CSV Flat Files—Revised: Readmissions Complications and Deaths—National.csv." Baltimore, MD: CMS, 2014. Available at https://data.medicare.gov/data/hospital-compare. Accessed August 14, 2014.

Chronic Conditions Data Warehouse. "Table A.1. Medicare Beneficiary Counts for 2003 – 2012." Baltimore, MD: Centers for Medicare & Medicaid Services, 2014a. Available at https://www.ccwdata.org/cs/groups/public/documents/document/ccw_website_table_a1.pdf. Accessed November 19, 2014.

Chronic Conditions Data Warehouse. "Table B.2. Medicare Beneficiary Prevalence for Chronic Conditions for 2003 Through 2012." Baltimore, MD: Centers for Medicare & Medicaid Services, 2014b. Available at https://www.ccwdata.org/cs/groups/public/documents/document/ccw_website_table_b2.pdf. Accessed November 19, 2014.

DeNavas-Walt, Carmen, Bernadette D. Proctor, and Jessica C. Smith. "Income, Poverty, and Health Insurance Coverage in the United States: 2012." U.S. Census Bureau, Current Population Reports, pp. 60–245. Washington, DC: U.S. Government Printing Office, 2013.

Gerhardt, Geoffrey, Alshadye Yemane, Keri Apostle, Allison Oelschlaeger, Eric Rollins, and Niall Brennan. "Evaluating Whether Changes in Utilization of Hospital Outpatient Services Contributed to Lower Medicare Readmission Rate." *Medicare & Medicaid Research Review,* vol. 4, no. 1, 2014, pp. E1–E13.

Gilman, Boyd, Sheila Hoag, Lorenzo Moreno, Greg Peterson, Linda Barterian, Laura Blue, Kristin Geonnotti, Tricia Higgins, Mynti Hossain, Lauren Hula, Rosalind Keith, Jennifer Lyons, Brenda Natzke, Brenna Rabel, Rumin Sarwar, Rachel Shapiro, Victoria Peebles, Cara Stepanczuk, KeriAnn Wells, and Joseph Zickafoose. "Evaluation of the Health Care Innovation Awards (HCIA): Primary Care Redesign Programs. First Annual Report, Volumes I and II." Princeton, NJ: Mathematica Policy Research, November 14, 2014.

Hansen, Ben B. "Full Matching in an Observational Study of Coaching for the SAT." *Journal of the American Statistical Association,* vol. 99, no. 467, 2004, pp. 609–618.

Health Indicators Warehouse. "Average Age of Medicare Beneficiaries (mean)." Hyattsville, MD: National Center for Health Statistics, HIW, 2014a. Available at http://www.healthindicators.gov/Indicators/Average-age-of-Medicare-beneficiaries-mean_308/Profile/ClassicData. Accessed November 19, 2014.

Health Indicators Warehouse. "Hospital Inpatient Medicare Admissions (per 1,000 beneficiaries)." Hyattsville, MD: National Center for Health Statistics, HIW, 2014b. Available at http://www.healthindicators.gov/Indicators/Hospital-inpatient-Medicare-admissions-per-1000-beneficiaries_2001/Profile/ClassicData. Accessed August 13, 2014.

Institute of Education Sciences. "What Works Clearinghouse: Procedures and Standards Handbook, Version 3.0." Washington, DC: U.S. Department of Education, IES, 2014. Available at http://ies.ed.gov/ncee/wwc/DocumentSum.aspx?sid=19. Accessed September 15, 2014.

Kaiser Family Foundation State Health Facts. "Dual Eligibles as a Percent of Total Medicare Beneficiaries." Data source: FY 2011 MSIS and CMS-64 reports. Menlo Park, CA: Kaiser Family Foundation, 2011. Available at http://kff.org/medicaid/state-indicator/duals-as-a-of-medicare-beneficiaries. Accessed February 26, 2016.

Peikes, Deborah, Stacy Dale, Eric Lundquist, Janice Genevro, and David Myers. "Building the Evidence Base for the Medical Home: What Sample and Sample Size Do Studies Need? White Paper." AHRQ Publication No.11-0100-EF. Rockville, MD: Agency for Healthcare Research and Quality, October 2011.

Rosenbaum, Paul R. "A Characterization of Optimal Designs for Observational Studies." *Journal of the Royal Statistical Society, Series B,* 1991, pp. 597–610.

Stuart, Elizabeth A. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science,* vol. 25, no. 1, 2010, pp. 1–21.

Truven Health Analytics. "AHRQ Quality Indicators, Prevention Quality Indicators v5.0 Benchmark Data Tables." Prepared for the Agency for Healthcare Research and Quality, U.S. Department of Health and Human Services. Santa Barbara, CA: Truven Health Analytics, March 2015. Available at http://www.qualityindicators.ahrq.gov/Downloads/Modules/PQI/V50/Version_50_Benchmark_Tables_PQI.pdf. Accessed August 18, 2015.

U.S. Census Bureau. "2008–2012 American Community Survey, Median Household Income." Washington, DC: U.S. Census Bureau, 2012.

This page has been left blank for double-sided copying.

CHAPTER 8

SANFORD HEALTH

Jelena Zurovac, KeriAnn Wells, Greg Peterson, Boyd Gilman, Laura Blue, Keith Kranker, Kate Stewart, Sheila Hoag, and Lorenzo Moreno

This page has been left blank for double-sided copying.

SANFORD HEALTH

CHAPTER SUMMARY

Introduction. Sanford Health received a $12.1 million Health Care Innovation Award (HCIA) to implement One Care, a medical home model, in 33 of its practices in Minnesota, North Dakota, and South Dakota. Key goals of the initiative were to integrate behavioral health care and care management services into primary care. Investments in workforce development and expansion of health information technology (IT) supported these goals. By the end of the award, Sanford Health aimed to reduce potentially preventable admission and outpatient emergency department (ED) visit rates by 20 percent and reduce total cost of care by 3 percent for Medicare, Medicaid, and Children's Health Insurance Program (CHIP) beneficiaries with targeted conditions. Sanford Health also aimed to improve quality-of-care outcomes such as optimal care for asthma, diabetes, and hypertension. Sanford Health's HCIA award ended on June 30, 2015.

Objectives. (1) To describe the design and implementation of Sanford Health's HCIA-funded intervention, including the role of clinicians in the intervention and the extent to which anticipated changes in providers' behavior occurred; (2) to assess impacts of the intervention on patients' outcomes and Medicare Part A and B spending during the award; and (3) to use both implementation and impact findings to identify possible explanations for the observed impacts.

Methods. We reviewed Sanford Health's program documents and self-monitoring metrics, conducted site visits and interviews with Sanford Health leadership and program staff, and surveyed participating trainees and clinicians. We estimated program impacts as the differences in outcomes during the intervention period for Medicare fee-for-service (FFS) beneficiaries attributed to 22 nonpediatric treatment practices with baseline data and for Medicare FFS beneficiaries attributed to 91 matched comparison practices, adjusting for any differences in outcomes between the two groups during a one-year baseline period.

Program design and implementation. The intervention had three components: (1) integration of behavioral health services into primary care; (2) provision of care management services; and (3) expansion of health IT to facilitate screening and management of health conditions via electronic health record (EHR)-based protocols, decision support, and quality measurement. Integration of behavioral health into primary care involved increased screening and provision of services for patients with anxiety, depression, and substance abuse (alcohol and drug abuse). Sanford Health provided care management services to patients with asthma, diabetes, heart failure, hypertension, and obesity. Sanford Health's intervention was largely implemented as planned, with some delays incorporating screenings into the EHR, implementing remote monitoring devices, and outreach to the Native American population in the Bemidji region. By the end of the award, Sanford Health engaged 290 staff across all 33 participating practices (pediatric and nonpediatric) in care teams and workforce development.

Clinicians' perceptions of intervention effects on the care they provide. Sanford Health's program required clinicians to buy into team-based care, such as using new behavioral

health screenings, communicating regularly with care teams, and referring patients to registered nurse (RN) health coaches and behavioral health triage therapists (BHTTs) when appropriate. The findings from the clinician survey administered at the end of the program suggest that more than 60 percent of clinicians believed that the program improved quality and patient-centeredness of care at participating practices. Data collected during site visits suggest that physicians increasingly referred patients to RN health coaches and BHTTs and valued newly available resources to address patients' behavioral health needs.

Impacts on patients' outcomes. The impact estimates indicate that, during the three years of the award, the intervention improved patients' outcomes in quality-of-care process and service use domains, but not in quality-of-care outcomes or spending domains. Specifically, the program had a statistically significant, favorable effect on quality-of-care processes (driven by improvements on the diabetes measure of 8.6 percent, with a one-sided *p*-value of 0.05) and service use (driven by the favorable effect on ED visits of 4.9 percent, with a one-sided *p*-value of 0.06).

Conclusion. The evidence indicates that, during the three-year award, Sanford Health had some success in improving patients' outcomes. The intervention improved outcomes in quality-of-care process and service use domains, driven by statistically significant improvements in receipt of recommended diabetes processes of care and statistically significant reductions in ED visits. The lack of statistically significant improvements in quality-of-care outcomes and spending domains appears not to be due to a failure to implement the program as planned. Rather, the lack of improvements in these outcomes might be due to (1) the intervention not being sufficiently intensive to generate substantively large effects and (2) the content of the intervention not being amenable to reduction in the analyzed outcomes.

Summary of intervention and impact results for Sanford Health

Intervention description			
Awardee description		Large integrated health system serving 100 communities in 9 states	
Award amount ($ millions)		$12.1 million	
Award extended beyond June 2015?		No	
Locations		Minnesota, North Dakota, and South Dakota (urban, suburban, and rural)	
Target population		All patients served by 33 of Sanford Health's practices, focusing on patients with at least 1 of 8 targeted conditions (asthma, anxiety, depression, diabetes, heart disease, hypertension, obesity, and substance abuse [alcohol and drug abuse])	
Interventions		Integrating behavioral health into primary care: • Screenings for behavioral health conditions • Short-term counseling and/or referrals Care management for five medical health conditions[a] • Patients coached on self-management skills; symptoms and progress monitored Expanded health information technology	
Metrics of intervention delivered		• 290 staff members helped implement intervention • Hired 18 behavioral health triage therapists • Increased share of patients identified as having depression from 13 to 17% and with anxiety from 10 to 14%	
Impact evaluation methods			
Core design		Difference-in-differences model with matched comparison group	
Treatment group	Definition	Medicare FFS beneficiaries attributed to 22 nonpediatric participating practices with baseline data	
	# of beneficiaries during primary test period[b]	12,950 to 18,238	
Comparison group definition		Medicare FFS beneficiaries attributed to 91 matched comparison practices	
Impact results: Quality-of-care processes domain			
Ambulatory care visit within 14 days of discharge (% of beneficiaries/quarter)		Comparison mean[c]	62.3%
		Impact estimate (% difference)	+<0.1 pp (+0.1%)
Received all four recommended diabetes processes of care (% of beneficiaries/year)		Comparison mean[c]	44.7%
		Impact estimate (% difference)	+3.8 pp (+8.6%)**
Combined impact estimate[d]		+4.3%**	
Impact conclusion[e]		**Statistically significant favorable effect**	
Impact results: Quality-of-care outcomes domain			
30-day unplanned hospital readmissions (#/1,000 beneficiaries/quarter)		Comparison mean[c]	10.9
		Impact estimate (% difference)	-0.1 (-1.3%)
Inpatient admissions for ACSCs (#/1,000 beneficiaries/quarter)		Comparison mean[c]	12.7
		Impact estimate (% difference)	+1.7 (+13.6%)
Combined impact estimate[d]		+6.2%	
Impact conclusion[e]		No substantively large effect	
Impact results: Service use domain			
All-cause inpatient admissions (#/1,000 beneficiaries/quarter)		Comparison mean[c]	82.5
		Impact estimate (% difference)	+1.5 (+1.8%)
Outpatient ED visits (#/1,000 beneficiaries/quarter)		Comparison mean[c]	138.9
		Impact estimate (% difference)	-6.8 (-4.9%)*
Combined impact estimate[d]		-1.6%	
Impact conclusion[e]		**Statistically significant favorable effect**	
Impact results: Spending domain			
Medicare Part A and B spending ($/beneficiary/month)		Comparison mean[d]	$898
		Impact estimate (% difference)	+$13 (+1.5%)
Impact conclusion[e]		Indeterminate effect	

Note: See the Sanford Health chapter for details on the intervention, impact methods, and impact results.

[a] The five conditions were asthma, diabetes, heart failure, hypertension, and obesity.

[b] Number of beneficiaries in the full treatment group across the quarters in the primary test period.

Summary of intervention and impact results for Sanford Health *(continued)*

[c] The comparison mean is the estimate of the outcome the treatment group beneficiaries would have had if they had not received the intervention. It is equal to the mean for the treatment group over the intervention quarters (in the primary test period) minus the impact estimate.

[d] The combined estimate is the average across all the individual estimates in the domain, where the impact estimate for each individual outcome is expressed as percentage change relative to the comparison group.

[e] We drew conclusions at the domain level based on the results of pre-specified primary tests, secondary tests (robustness checks), and consistency with implementation evidence. For each domain, we could draw one of five conclusions: (1) Statistically significant favorable effect (the highest level of evidence), (2) Substantively important (but not statistically significant) favorable effect, (3) Substantively important (but not statistically significant) unfavorable effect, (4), No substantively large effect, and (5) Indeterminate effect. Section V.A.8 of this report describes the decision rules we used to reach each of these possible conclusions.

*Significantly different from zero at the .10 level, one-tailed test.

**Significantly different from zero at the .05 level, one-tailed test.

***Significantly different from zero at the .01 level, one-tailed test.

ACSC = ambulatory care-sensitive condition; ED = emergency department; FFS = fee-for-service; pp = percentage point.

I. INTRODUCTION

This report presents findings from the evaluation of Sanford Health's Health Care Innovation Award (HCIA), with a focus on program impacts on patients' outcomes. Section II provides an overview of Sanford Health's HCIA-funded intervention and the design of the impact evaluation. Section III describes the design and implementation of the intervention, including how the program could be expected to affect study outcomes through changes in patients' and providers' behavior. In Section IV, we assess the evidence of the extent to which planned changes in providers' behavior occurred. Section V describes our methods for, and results and conclusions from, estimating program impacts on patients' outcomes in four domains: quality-of-care processes, quality-of-care outcomes, service use, and spending. Section VI draws conclusions by synthesizing the impact and implementation findings.

II. OVERVIEW OF SANFORD HEALTH'S HCIA-FUNDED INTERVENTION AND THE IMPACT EVALUATION

A. Sanford Health's HCIA-funded intervention

Sanford Health, one of the largest integrated care systems in the country, received $12.1 million in HCIA funding to implement One Care, a medical home model, in 33 of its primary care practices in Minnesota, North Dakota, and South Dakota over a three-year period (Table II.1). Participating practices joined the program at different times and, to simplify our impact analysis, we grouped them into two cohorts: (1) practices that joined from April 1, 2013, to December 31, 2013; and (2) practices that joined from January 1, 2014, to December 31, 2014. Sanford Health's HCIA award ended on June 30, 2015.

One Care aimed to transform primary care delivery through a multifaceted intervention that included three interrelated components: (1) integration of behavioral health services into primary care, (2) care management for patients with targeted conditions, and (3) expansion of health information technology (IT). A team-based approach to care and investments in workforce development supported these intervention components.

Even though there was no direct patient enrollment, the award closely focused on patients with one or more of eight targeted conditions. The award focused on improved identification and treatment of targeted behavioral health conditions (anxiety, depression, and substance abuse [alcohol and drug abuse]) and on care management services for patients with targeted medical conditions (asthma, diabetes, heart failure, hypertension, and obesity).

Sanford Health set goals for Medicare, Medicaid, and Children's Health Insurance Program (CHIP) patients with one or more of eight targeted conditions. For this population and by the end of the award, Sanford Health aimed to reduce (1) potentially preventable admission rates by 20 percent, (2) emergency department (ED) visit rates by 20 percent, and (3) total cost of care by 3 percent. Sanford Health also aimed to improve quality-of-care process outcomes such as optimal care for asthma, diabetes, and hypertension, but did not set specific targets.

To implement the intervention, Sanford Health hired new staff and retrained existing staff to create integrated care teams that provide comprehensive, patient-centered care. The new staff

included behavioral health triage therapists (BHTTs), registered nurses (RNs), health coaches, and panel managers. Some practices also hired addiction navigators, who offered peer support to patients with alcohol and drug abuse, and one employed a community health worker (CHW). The existing staff, who received new training, included physicians, nurses, medical assistants, and practice managers. Sanford Health equipped care teams with new tools, such as screening instruments and clinical guidelines, to improve the quality of care. According to Sanford Health's theory of action, these integrated care teams would use their new skills and tools to more effectively coordinate care and engage patients to manage their conditions, in turn leading to improved outcomes and lower health care expenditures. (Section III.A.3 describes the awardee's theory of action in detail.)

Table II.1. Summary of Sanford Health's HCIA program and our evaluation for estimating its impacts on patients' outcomes

	Program description
Award amount	$12,142,606
Award start date	June 2012
Implementation date	Staggered: 21 practices started on April 1, 2013 (cohort one), and 12 practices started on January 1, 2014 (cohort two)[a]
Award end date	June 30, 2015
Awardee description	Sanford Health, headquartered in Sioux Falls, South Dakota, is one of the largest integrated health systems in the nation, serving more than 100 communities in nine states.
Intervention overview	Sanford Health integrated behavioral health care, care management, and expanded health IT into 33 primary care practices, of which 24 serve mainly adults and 9 serve mainly children.
Intervention components	A team-based approach to care and workforce development supported each of the intervention components below: 1. **Integration of behavioral health services into primary care.** BHTTs screened patients for behavioral health conditions, provided referrals, and short-term counseling to patients with anxiety, depression, and substance abuse (alcohol and drug). 2. **Provision of care management.** RN health coaches provided care management for the following targeted health conditions: diabetes, asthma, hypertension, heart failure, and obesity. RN health coaches also worked on increasing use of the patient portal and telemonitoring for relevant populations. Panel managers/care coordination assistants supported RN health coaches with data and contributed to quality measurement. 3. **Expansion of health IT.** Health IT facilitated screening and management of health conditions via EHR-based instruments and practice guidelines for targeted conditions, including decision support. Electronic disease registries facilitated panel management and quality measurement.
Target population	All patients served by 33 of Sanford Health's practices, focusing on patients with at least one of following eight health conditions: anxiety, asthma, diabetes, depression, heart failure, hypertension, obesity, and substance abuse (alcohol and drug abuse).
Target impacts on patients' outcomes	For Medicare, Medicaid, and CHIP beneficiaries with one or more of eight targeted conditions: • Reduce preventable admissions by 20 percent • Reduce ED visits by 20 percent • Reduce overall costs of care by 3 percent • Improve quality-of-care process outcomes such as optimal care for asthma, diabetes, and hypertension (extent of improvement not specified)
Workforce development	Through March 2015, Sanford Health hired 57 new full-time equivalent staff and provided 7,818 person-training sessions (counting trainees for each training attended). By June 2015, 290 staff across 33 practices were engaged in care teams and workforce development. Sanford Health used HCIA funds to pay the salaries of 13 panel managers/care coordination assistants, to reimburse for nonbillable services for 18 BHTTs, and to pay for HCIA-related training.
Locations	Minnesota, North Dakota, and South Dakota (urban, suburban, and rural areas)

Table II.1 *(continued)*

	Impact evaluation
Core design	Difference-in-differences with matched comparison group
Treatment group	Adult Medicare FFS beneficiaries whom we attributed to the 22 nonpediatric treatment practices with baseline data and had diagnosis and/or procedure codes for one or more of eight targeted chronic conditions; attribution was done using the same method that CMMI uses for the CPC Initiative.[b]
Comparison group	Medicare FFS beneficiaries whom we attributed to matched comparison practices in the same states as the treatment practices.
Intervention component(s) included in impact evaluation	**All components described above.** Sanford Health expected the program components to work jointly in benefiting the targeted patients attributed to the treatment practices.
Extent to which the treatment group reflects the awardee's target population (for the component(s) evaluated)	**Moderate**. Our treatment group consists exclusively of Medicare FFS beneficiaries with one or more of eight targeted conditions at the start of the intervention period, whereas the awardee targeted all Medicare, Medicaid, and CHIP beneficiaries with those conditions. Limitations in Medicare managed care data and lags in Medicaid and CHIP data availability prevented us from conducting tests of effectiveness on these populations. For that reason, we excluded nine pediatric practices from the evaluation. We also excluded two practices that did not have baseline data. Further, the intervention group did not include Medicare beneficiaries who did not have a behavioral targeted condition at the start of the intervention period, but screened positive during the award period as part of award-related screening, and subsequently might have received intervention services. Finally, even though Sanford Health expected that provision of team-based patient-centered care might benefit all patients at participating practices, the focus of the award was on patients with one or more of eight targeted conditions.
Study outcomes, by domain	1. **Quality-of-care processes.** Preventive care for diabetes and 14-day ambulatory care follow-up after a hospitalization 2. **Quality-of-care outcomes.** 30-day unplanned readmissions and inpatient admissions for ambulatory care-sensitive conditions 3. **Service use.** All-cause inpatient admissions and outpatient ED visits 4. **Spending.** Medicare Part A and B spending

Sources: Review of Sanford Health reports, including its original application, operational plan, and 13 narrative reports to the Centers for Medicare & Medicaid Services.

[a] Sanford Health grouped practices into four phases of implementation: April to August 2013, September to December 2013, January to June 2014, and July to December 2014. To simplify our quantitative analysis, we grouped practices into two cohorts: (1) those that joined the program from April 1, 2013, to December 31, 2013 (cohort one); and (2) those that joined the program from January 1, 2014, to December 31, 2014 (cohort two).

[b] In each baseline and intervention month, we attributed beneficiaries to the primary care practice whose providers (physicians, nurse practitioners, or physician assistants) provided the plurality of primary care services in the past 24 months. Sanford Health provided data on which providers worked in the treatment practices and when. In each period (baseline and intervention), we assigned each beneficiary to the first treatment practice he or she was attributed to in the period, and continued to assign him or her to that practice for all quarters in the period.

BHTT = behavioral health triage therapist; CHIP = Children's Health Insurance Plan; CMMI = Center for Medicare & Medicaid Innovation; CPC = Comprehensive Primary Care; ED = emergency department; EHR = electronic health record; FFS = fee-for-service; HCIA = Health Care Innovation Award; IT = information technology; RN = registered nurse.

B. Overview of impact evaluation

To estimate program impacts on patients' outcomes, we compared outcomes for Medicare fee-for-service (FFS) beneficiaries served by 22 nonpediatric primary care practices that participated in the HCIA intervention and had baseline data (treatment practices) to outcomes for beneficiaries served by 91 matched comparison practices, adjusting for any differences in outcomes between these two groups before the intervention began. Our treatment group

consisted exclusively of Medicare FFS beneficiaries with one or more of eight chronic conditions, whereas Sanford Health's targeted population comprised all Medicare, Medicaid, and CHIP beneficiaries with those conditions. Limitations in Medicare managed care administrative data and lags in Medicaid and CHIP data availability prevented us from conducting tests of effectiveness on the Medicaid or CHIP populations. As a result, we excluded 9 pediatric practices from the evaluation. We also excluded 2 practices that were newly founded during the intervention period and thus did not have baseline data. Table II.1, bottom panel, summarizes our impact evaluation design.

We selected the 91 comparison practices for the evaluation from the pool of 997 potential comparison practices located in Minnesota, North Dakota, and South Dakota, the three states with participating Sanford Health practices. We selected practices that were similar to the 22 treatment practices in terms of practice characteristics and characteristics of their Medicare FFS patients before the intervention began.

We estimated impacts on outcomes, as measured in Medicare FFS claims data, which we grouped into four domains: (1) quality-of-care processes, (2) quality-of-care outcomes, (3) service use, and (4) spending. Across the HCIA awardees in primary care redesign (PCR), we designed our impact evaluations to identify promising interventions or intervention components—consistent with the evaluation goals from the Center for Medicare & Medicaid Innovation (CMMI) to find programs that could be scaled or retested as part of a future model test. Before conducting the analysis, we specified a series of primary tests, describing the evidence we would need to conclude that the program was effective, and the awardee and CMMI reviewed these tests. Each test specified a population, outcome, period, expected direction of effect, and threshold that we count as substantively important. The purpose of these primary tests was to focus the impact evaluation on hypotheses that would provide the most robust evidence about program effectiveness. We used the results from the primary tests and robustness checks to draw conclusions about program impacts in each of the four evaluation domains. Because we sought to identify promising interventions rather than only those programs with unequivocally demonstrated success, we conducted one-sided statistical tests (that is, testing only for program benefits) and used a threshold for statistical significance of 0.1, which is not as strict as the conventional standard of 0.05.

Our impact evaluation design reflects the effects of all three intervention components, but only for part of Sanford Health's HCIA target population. Because the treatment group included all targeted FFS Medicare beneficiaries attributed to the 22 nonpediatric practices with baseline data, these beneficiaries as a group (although not necessarily individually) were exposed to all three components of the intervention. However, the treatment group did not include Medicare managed care beneficiaries, Medicaid, and CHIP beneficiaries, key members of the awardee's target population. Further, the intervention group did not include Medicare beneficiaries who did not have a behavioral targeted condition at the start of the intervention period, but had screened positive during the award period as part of award-related screening and subsequently might have received intervention services. We excluded such beneficiaries from the treatment group because we could not identify the beneficiaries in the comparison group who would represent the appropriate counterfactual – that is, comparison group members who would have screened

positive had they been in the treatment group. (Also, because Sanford's program might affect the composition of the attributed beneficiaries, we excluded from the intervention group Medicare beneficiaries who did not have a medical targeted condition at the start of the intervention period, but who developed one during the intervention period.) Finally, even though Sanford Health expected that provision of team-based patient-centered care might benefit all patients at participating practices, the focus of the award was on patients with one or more of eight targeted conditions.

III. PROGRAM IMPLEMENTATION

In this section, we first provide a detailed description of Sanford Health's HCIA-funded intervention, highlighting the program's design and its theory of action. Second, we assess the evidence on the extent to which the intervention was implemented as planned based on measures of program enrollment, service delivery, staffing, training, and timeliness. Third, we summarize the facilitators and barriers associated with implementation effectiveness.

We based our evaluation of Sanford Health's program implementation on a review of the awardee's quarterly reports to CMMI and self-monitoring program metrics, telephone discussions and follow-up communications with program administrators, and information collected during site visit interviews with selected staff in June 2014 and April 2015. We visited a sample of internal medicine, family medicine, and pediatrics practices in Fargo, North Dakota; Sioux Falls, South Dakota; and Thief River Falls, Minnesota. We did not visit practices in Bemidji, Minnesota, because it was too difficult to reach. During site visits we spoke to awardee leadership, physicians, BHTTs, RN health coaches, panel managers/care coordination assistants (CCAs), practice managers, and an addiction navigator. We did not verify the quality of the performance data reported by the awardee in its self-measurement and monitoring reports.

A. Program design and adaptation

1. Target population and patient identification, recruitment, and enrollment

In this section, we describe how Sanford Health selected practices to participate in the HCIA intervention and identified target patients for care management and behavioral health services.

Identification of practices for participation. Sanford Health selected 33 Sanford Health-owned primary care practices in four regions of three adjacent states representing urban, suburban, and rural communities. Of the 33 participating practices, 16 specialized in family medicine, 8 in internal medicine, and 9 in pediatrics. Participating practices were distributed across four regions: Sioux Falls with 15, Fargo with 11, Bemidji with 4, and Thief River Falls with 3 practices. (The impact evaluation focused on 15 family medicine and 7 internal medicine practices that served adults, spanning all four regions.) Practices in the Bemidji region also focused on outreach to Native American communities whose residents face substantial health disparities, such as higher rates of chronic conditions and behavioral health diagnoses than seen in the general population, complicated by a health system that does not always align with Native American communities' values.

Target population of patients. There was no direct patient enrollment; team-based, patient-centered care was available to all patients. The program specifically targeted improvements in services delivered to patients with one or more of the eight targeted conditions: asthma, anxiety, depression, diabetes, heart disease, hypertension, obesity, and substance abuse (alcohol and drug abuse). Sanford Health identified these patients using diagnosis codes in administrative data. Drawing from four psychometrically validated instruments, Sanford Health developed and implemented the 6-item Behavioral Health Screener (BH-6) to identify more patients with anxiety, depression, and substance abuse (alcohol and drug abuse).

2. Intervention components

Sanford Health's intervention had three components. Two components—integration of behavioral health services into primary care and care management for patients with at least one of the medical targeted conditions—delivered services directly to patients at participating practices. The other component—expansion of health IT—provided tools for staff and patients to improve the quality and efficiency of care. To support these components, Sanford Health implemented a series of trainings for participating staff.

Some of the staff interviewed during site visits noted that Sanford Health was preparing to integrate behavioral health and quality improvement activities before the HCIA award. Some sites had started implementing some award-related interventions, but generally, little was implemented beyond preparations and building buy-in. For example, an interviewee in Thief River Falls noted that Sanford Health had hired one RN health coach before the award. Practices in Sioux Falls were an exception: they reported having RN health coaches in place before the award. Discussions with the awardee suggested that HCIA enabled Sanford Health to refine these roles, scale them up in conjunction with other program components such as team-based care, and move much more quickly than without the funds. Despite this limited prior experience in certain aspects of the intervention, the HCIA funding provided a substantial investment in these activities.

Integration of behavioral health into primary care. Before HCIA, some Sanford Health practices in the Fargo region had some behavioral health services in the primary care setting, but did not employ BHTTs. With HCIA funding, Sanford Health introduced this position, hiring new staff and retraining existing staff to serve as BHTTs. BHTTs met with patients with existing behavioral health diagnoses, patients who screened positive on the BH-6, and patients who exhibited signs of behavioral health issues during encounters. BHTTs triaged and assessed patients to determine the severity of their anxiety, depression, and substance abuse (alcohol and drug abuse). BHTTs offered short-term counseling (up to six sessions) for lower-risk patients and referred higher-risk patients, such as those with schizophrenia, to specialists. During our site visits, many staff suggested that Sanford Health identified more patients with behavioral health conditions and thus providers made more referrals to behavioral health specialists. However, these referrals were described as more appropriate than pre-HCIA referrals because they were for patients with more severe symptoms, whereas patients with less severe symptoms were managed in the primary care setting. Many providers characterized identification and treatment of behavioral health issues as necessary precursors to improving physical health. For example, patients who were depressed or engaged in alcohol or drug abuse were less likely to effectively

manage other health conditions, such as diabetes. BHTTs billed payers for reimbursable services such as counseling and assessment; nonreimbursable time, such as triaging patients and time spent in trainings, were HCIA-funded.

Care management. Sanford Health also scaled up the RN health coach position, which existed in Sioux Falls practices before the award. Sanford Health hired and retrained existing nurses, focusing trainings on topics such as motivational interviewing and trauma-informed care. (Trauma-informed care training helped providers to be more attuned to the social history of the patient, including issues such as domestic violence.) RN health coaches also completed an intensive Chronic Care Professional training for continuing medical education credits. RN health coaches focused on care management for patients with targeted medical conditions. They monitored patients' conditions, helped patients set achievable goals, and monitored patients' progress toward meeting those goals. Staff at several practices reported that they initially built infrastructure such as disease registries and panel management protocols for asthma, diabetes, and/or hypertension before moving on to heart failure and obesity.

Sanford Health provided some patients with remote blood pressure cuffs and scales configured to automatically transmit patients' vital signs into the electronic health record (EHR), to assist RN health coaches' care management efforts. Some RN health coaches also provided follow-up care to patients discharged from the hospital, although this service was not emphasized as a component of award activities. Sanford Health also integrated panel managers/CCAs into care teams. (In Sioux Falls, panel managers were rebranded as CCAs to reflect their Medical Assistant credential.) Panel managers/CCAs produced population management reports to identify which patients might benefit from outreach, such as patients with diabetes overdue for hemoglobin A1c testing. They also reviewed patients' information to help teams with previsit planning, such as determining which patients should see a BHTT or RN health coach and summarizing incoming patients' medical records. Other than time spent in trainings, RN health coaches were not compensated with HCIA funds, although panel managers'/CCAs' salaries were HCIA-funded.

Expansion of health IT. Sanford Health enhanced its EHR to include clinical practice guidelines for targeted conditions, incorporating behavioral health into nationally recognized guidelines. For instance, Sanford Health modified the National Institute of Health's guidelines for asthma care to incorporate the behavioral health assessment. In April 2014, Sanford Health also built the BH-6 into its EHR and into MyChart, the patient portal available at all Sanford Health practices, including practices not participating in One Care. Before April 2014, many practices were administering a paper-based BH-6. Making the BH-6 electronic at the 24 nonpediatric practices enabled clinicians to screen patients more efficiently, because screenings could be completed before the appointment, either at home via MyChart or with a tablet in the waiting room. In addition, screening results were automatically available in the EHR for clinicians to review. Sanford Health also built disease registries, which panel managers/CCAs used to monitor patients with specific conditions.

3. Theory of action

Based on an extensive review of information collected from awardee documents, site visit discussions, and telephone discussions with awardee leadership, we developed a theory of action to depict the mechanisms through which the program was expected to improve the outcomes we selected for the impact evaluation. (See Table II.1 for the list of these outcomes). Sanford Health's theory of action included four steps:

1. **Developed Sanford Health's workforce into integrated primary care teams.** Sanford Health One Care practices hired new staff and retrained existing staff to build care teams that integrated behavioral health staff and RN health coaches. Care team staff included primary care physicians, RN health coaches, BHTTs, and panel managers/CCAs. Some practices' care teams also included addiction navigators and one included a CHW. Addiction navigators offered peer support to patients with alcohol or drug abuse, and the CHW primarily conducted outreach to Native American patients who lived near practices in Bemidji. The CHW also led Better Choices Better Health chronic disease self-management groups, held at community centers.

   Teams communicated regularly, often huddling daily or weekly, to discuss patients and to increase team cohesion. Core team meetings and informal socializing also helped teams build consensus. Improved coordination and collaboration in integrated care teams was expected to help improve the quality of care by ensuring that patients received needed services and by reducing duplication of services.

2. **Trained care teams and equipped them with tools to provide high quality care.** The Clinical Skills Development Team (CSDT), composed of more than 50 multidisciplinary clinicians from participating practices, developed clinical practice guidelines, training curricula, and screenings to support practice transformation. CSDT members included physician champions and core team members, leading transformation within practices and facilitating staff buy-in. Clinical practice guidelines facilitated clinicians' ability to adhere to consistent, nationally recognized standards and to identify behavioral health issues.

   Sanford Health incorporated training modules into its online learning platform and provided classroom trainings. Trainings helped clinicians and staff develop a common understanding of One Care components and goals. (We discuss training in more detail and present the results of the trainee survey in Section III.B.4.)

   Other tools provided to care teams included the BH-6 instrument, disease registries, patient synopses (all three incorporated into the EHR), MyChart, and remote monitoring devices (as described in Section III.A.2). Some participating practices also used the Patient Activation Measure (PAM) to assess patients' knowledge, skills, and confidence in managing their conditions. These trainings and tools supported multiple aspects of patient-centered care, including early identification of behavioral health issues, panel management, care coordination, patient engagement, and chronic condition management.

3. **Better equipped care teams, coordinated care, and engaged patients more effectively, filling in care gaps and helping patients to better manage their care.** Primary care clinicians incorporated behavioral health screening into their workflows, heightening their

awareness about how to identify and treat behavioral health conditions and increasing their referrals to BHTTs and behavioral health specialists. Clinicians referred patients with targeted medical conditions to RN health coaches. When possible, clinicians made a warm handoff by introducing patients to BHTTs and RN health coaches rather than always relying on referrals. Similarly, RN health coaches handed off patients to BHTTs when patients exhibited behavioral health issues. Sanford Health staff described these personal introductions as facilitators of patient engagement and adherence. Using skills developed in trainings funded by the award, BHTTs and RN health coaches assessed, treated, monitored, and referred patients with targeted conditions, whereas panel managers/CCAs summarized patients' data for care teams. Teams also engaged patients via MyChart, through which patients could access information, communicate with providers, and complete screenings such as the BH-6.

4. **Activated patients receiving venue-appropriate care were expected to have fewer hospital admissions and ED visits, thus reducing total costs of care.** The awardee proposed that early identification and treatment of behavioral health conditions would facilitate patients' ability to improve their lifestyles and manage co-occurring medical conditions, such as asthma and diabetes. Early identification of behavioral health conditions would also lead to earlier intervention via short-term counseling in the primary care setting or referral to a specialist. Improved management of chronic conditions in the primary care setting would prevent patients' health conditions from advancing in severity and requiring a higher-level of care in a hospital or ED. New efforts to proactively identify, treat, and manage behavioral and other medical conditions would lead to improved scores on quality metrics.

   Having multiple points of contact in the primary care setting would result in fewer unnecessary ED visits. Self-management education was expected to reduce the need for use of acute care services by improving early management of emerging health problems. Although not emphasized as part of award activities, RN health coaches' management of post-discharge care might reduce hospital readmissions. Ultimately, healthier patients would require less care, including costlier specialized and inpatient care, reducing average costs per patient.

Text box III.1. Examples from Sanford Health illustrating the program's theory of action

"In June, we gave a young man with chronic heart failure a profile scale from Sanford One Care telemotivation [telemonitoring] program. Since then he has been able to monitor his weight more closely at home and was able to address a recent increase of 30 pounds. We have kept his mother informed with visits to his home for additional support. We were also able to connect him with behavioral health resources. This was a gentleman who, by bringing him a scale, we were able to establish rapport and start to take significant steps forward to improve his health. He even allowed the community paramedic [to] join him at a cardiac rehab appointment which he was unwilling to engage in prior to this recent encouragement and support."

"This story is best told in the words of one of our behavioral health triage therapists: 'Last week a daughter of one of our patients called. I had not yet met, nor had contact with the patient in question before, but she wanted to do an 'intervention' with her father who had been heavily drinking for the past 20 years. I spoke with her about some options, gave her support, and linked her to the addiction navigator to answer additional questions and discuss options. I talked with her throughout the week and after their family meeting to provide support. Today I received a thank you card in the mail with this message written in it:

Thank you for your guidance and help last week. Without you, we wouldn't have been able to get into the doctor at that critical point when my dad was open to it. He admitted himself into Tallgrass on Saturday—a 30-day treatment program. We all had our part in dad's acceptance and you were an integral part. Thank you.'"

Source: Sanford Health's 12th quarterly report to the Centers for Medicare & Medicaid Services

4. Intervention staff and workforce development

Table III.1 provides key details about staff hired and trained for the HCIA-funded intervention. Sanford Health engaged all types of staff in One Care. For example, office managers and administrators participated in core team meetings and provided administrative support and licensed practical nurses (LPNs) administered the BH-6. However, most trainings and new care processes targeted physicians or advanced practice providers, RN health coaches, BHTTs, and panel managers/CCAs. Some practices also employed addiction navigators and one practice employed a CHW.

Table III.1. Intervention staff, their credentials, and responsibilities

Staff members	Credentials	Staff/team responsibilities	Adaptations from originally planned roles
Clinicians (primary care, behavioral health)	MD, NP, or PA	• Participated in CSDT to develop and implement clinical guidelines and trainings • Participated in staff engagement activities such as core team meetings and team huddles • Communicated regularly with care teams, especially BHTTs and RN health coaches, and handed off patients when appropriate	No
BHTT	Social workers or licensed professional mental health counselors	• Triaged patients with behavioral health issues after referral from clinicians or RN health coaches • Conducted behavioral health assessments • Offered short-term therapy in the primary care setting for patients with anxiety, asthma, and substance abuse (alcohol and drug abuse) • Coordinated referrals to behavioral health specialists	No
RN health coaches	RN	• Provided chronic condition management education to patients with targeted conditions (asthma, diabetes, heart failure, hypertension, and obesity) • Provided ongoing support to patients via telephone calls and electronic messages	No
Panel managers/ CCAs	• Lay people (Fargo, North Dakota; Bemidji and Thief River Falls, Minnesota) • Medical assistants (Sioux Falls, South Dakota)	• Managed patients' data to assist care teams with previsit planning • Used registries to identify patients with chronic conditions • Reached out to patients to administer screenings and schedule visits	No
Addiction navigators	Lay people	• Offered peer support to patients with alcohol and/or drug abuse • Aligned patients with appropriate addiction treatment • Supported patients after addiction treatment	Yes; awardee placed the first addiction navigator outside the practice, but shifted to internal placement to increase team's integration and referrals

Table III.1 *(continued)*

Staff members	Credentials	Staff/team responsibilities	Adaptations from originally planned roles
CHW	CHW certification (Bemidji)	• Built relationships with local Native American communities • Served as a liaison between Native American communities and practices • Convened Better Choices Better Health groups to educate the community about chronic condition management • Conducted nonclinical home visits with patients who were recently discharged from the ED and had no primary care physician • Assisted patients in navigating the health care system, especially aligning patients with primary care	Yes; awardee planned to place cultural advisors on care teams at the point of care, but realized they should instead focus on outreach to communities to help bridge cultural divides

Sources: Interviews and document review.

BHTT = behavioral health triage therapist; CCA = care coordination assistant; CHW = community health worker; CSDT = Clinical Skills Development Team; ED = emergency department; MD = doctor of medicine; NP = nurse practitioner; PA = physician assistant; RN = registered nurse.

B. Implementation effectiveness

In this section, we examine the evidence on implementation effectiveness—that is, we analyze measures of the intervention delivered and, when possible, compare those measures to the services the awardee intended to deliver. We assess the evidence on implementation effectiveness in five areas: (1) program enrollment, (2) service delivery, (3) staffing, (4) training, and (5) implementation timeliness. To conduct this analysis, we used data from site visits and interviews with program administrators and selected frontline staff and self-reported metrics included in Sanford Health's self-monitoring and measurement reports to CMMI. We report total enrollment and staffing measures through June 2015, the end of the award. Through March 2015, three months before the end of the award, we report quarterly enrollment, the percentage of patients with behavioral health conditions, the percentage of patients using MyChart, and implementation metrics from trainee survey data. We also report the number of encounters with RN health coaches for adult patients with chronic conditions from January 2013 to December 2014, the only period for which these data were available.

To the extent possible, we restricted implementation effectiveness results to the 22 nonpediatric practices included in the impact evaluation in order to directly relate implementation effectiveness and quantitative impact results. For several sets of results, it was not possible to report information for the 22 nonpediatric practices included in the impact evaluation. We report the number of staff hired and actively working in care teams by position and the percentage of patients using MyChart for all 33 participating practices. We report the number of patients served per quarter (including Medicare, Medicaid, commercially insured, and others); percentage of patients screened for behavioral health issues; percentage of patients with identified behavioral health conditions; and number of encounters with RN health coaches for

24 nonpediatric practices. We report trainee survey results for 22 nonpediatric practices included in the impact evaluation.

1. Program enrollment

As previously noted, Sanford Health did not enroll patients into One Care. Even though the award focused on patients with one or more of eight chronic conditions, all patients who sought care at participating practices could potentially benefit from improvements in team-based, patient-centered care. The 24 nonpediatric practices reported serving an average of 69,843 adult patients per quarter (Medicare, Medicaid, commercially insured, and others) from January 2013 to March 2015, the only period for which these data were available for adult patients.

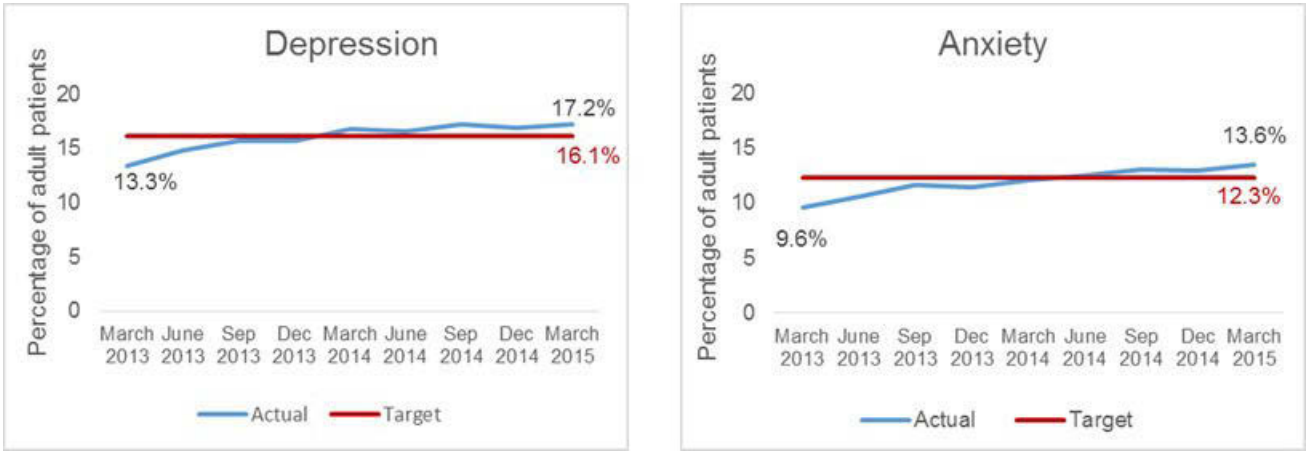
2. Service-related measures

Integration of behavioral health into primary care. Conversations during our site visits suggested that Sanford Health planned to administer BH-6 screening to all adult patients with qualifying encounters, which included new patient encounters and routine annual well visits. At 24 participating nonpediatric practices, Sanford Health observed an increase in the proportion of adult patients with a qualifying encounter who were screened from 28 percent in April to June 2014 to 56 percent from January to March 2015) (data not shown). The percentage of patients with depression and anxiety exceeded targets Sanford Health set based on national prevalence rates about halfway through the award (Figure III.1a). However, Sanford Health did not observe a similar increase in the number of patients identified with alcohol and drug abuse disorders (Figure III.1b).

Care management. For 24 participating nonpediatric practices, Sanford Health reported 31,320 encounters with RN health coaches for adult patients with chronic conditions from January 2013 to December 2014, the only period for which these data were available. However, the awardee described this number as "grossly underreporting" the number of RN health coach encounters due to challenges with extracting and integrating data from multiple EHRs. Care management activities also encompassed use of disease registries to track patients and remote monitoring of blood pressure and body mass index. Sanford Health also tracked the percentage of patients receiving optimal diabetes and asthma care, indicating a focus on these conditions.

Expansion in health IT. Among all 33 participating practices, Sanford Health observed a steady increase in the number of patients using MyChart. From January 2013 to March 2015, the percentage of patients who accessed the portal increased from 11 to 32 percent. However, this was still much lower than the awardee's goal of 70 percent adoption (data not shown). Figures for nonpediatric practices alone are not available.

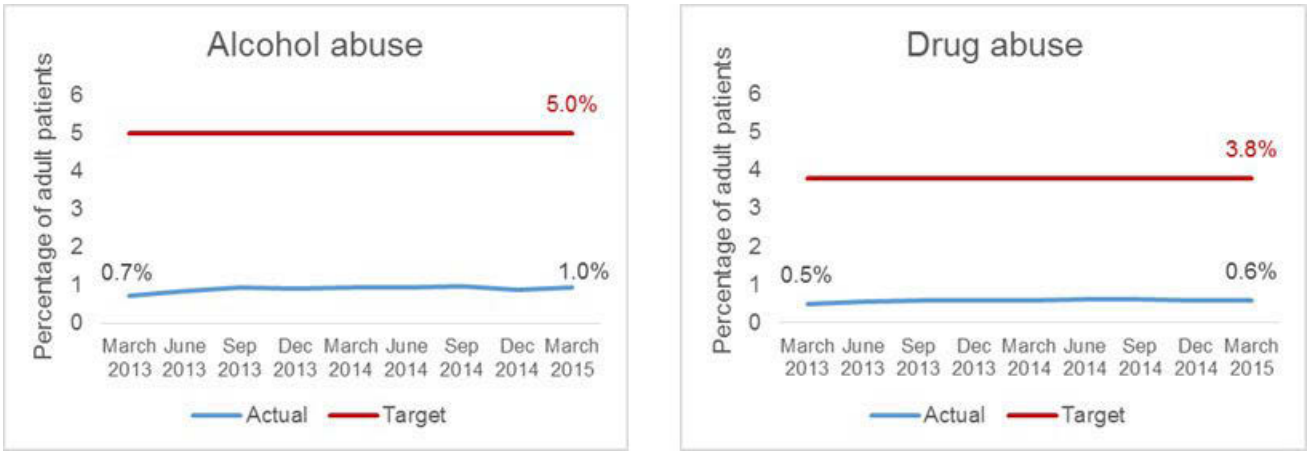
Figure III.1a. Percentage of adult patients who screened positive for depression or anxiety



Source: Sanford Health's measurement and monitoring report submitted to the Centers for Medicare & Medicaid Services in June 2015.

Note: Sanford set target rates for percentage of patients screening positive based on national prevalence rates.

Figure III.1b. Percentage of adult patients who screened positive for alcohol or drug abuse disorder



Source: Sanford Health's measurement and monitoring report submitted to the Centers for Medicare & Medicaid Services in June 2015.

Notes: Sanford set target rates for percentage of patients screening positive based on national prevalence rates.

3. Staffing measures

By June 2015, across all 33 participating practices, Sanford Health engaged 290 staff in seven key roles on care teams and workforce development (Table III.2), nearly meeting its initial target of 325 staff. Engaged staff included both HCIA-funded and non-HCIA-funded new hires and existing staff. They accounted for 56.95 full-time equivalent (FTE) new hires, exceeding its target of 50 new FTE hires. Most of the new hires were panel managers/CCAs, BHTTs, and practice administrators. Sanford Health allocated HCIA funding for staff time spent developing trainings and protocols and to supplement staff's salaries for nonreimbursable work. Panel managers/CCAs' salaries were fully HCIA-funded and BHTTs salaries were partially HCIA-funded. All other staff received HCIA-funded compensation only for time spent developing trainings and protocols and participating in trainings. Staffing data were not available for the 22 nonpediatric practices included in our impact evaluation.

Table III.2. Number of care team staff at 33 participating practices, by type of position

Type of position	Number of staff
Physicians	175
Registered nurse health coaches	38
Advanced practice providers	32
Behavioral health triage therapists	18
Panel managers/care coordination assistants	13
Peer support advocates	4
Community health workers	1
Others	9[a]
Total	290

Source: Sanford Health's final narrative progress report submitted to the Centers for Medicare & Medicaid Services in June 2015.

Note: Inclusive of all staff trained and actively working in care teams across all 33 participating practices, both HCIA-funded and nonfunded. Data are as of June 30, 2015.

[a] Sanford Health's progress report did not show a breakdown of the "others" category.

4. HCIA-funded training

Sanford Health created a training curriculum to support practice transformation. Topics included chronic disease management, integration of care teams, motivational interviewing, trauma-informed care, cultural mindfulness, and PAM. RN health coaches completed Chronic Care Professional training, a four-part series for which nurses received a continuing education certificate upon successful completion of an examination. All staff—including physicians, administrators, nurses, behavioral health staff, and panel managers/CCAs—participated in training, although most trainings targeted RN health coaches and BHTTs.

To assess perspectives of HCIA-funded staff who received training, we administered the HCIA Primary Care Redesign Trainee Survey from January to March 2015 (21 to 23 months

after cohort one practices began implementation and 12 to 14 months after cohort two practices began implementation). These dates corresponded to the end of the quantitative evaluation primary test period for the 15 cohort one practices and the beginning of the primary test period for the 7 cohort two practices included in the impact evaluation (See Section V.A.6 for a description of the primary test period). Of the 281 staff who participated in Sanford Health's HCIA-funded program at the time of the survey at all 33 participating practices, 193 responded to the trainee survey (leading to a response rate of 69 percent). The trainees included the following staff roles as defined in the survey, listed in the order of the number of respondents: LPNs, RNs, physicians, receptionists/registrars, others, health coaches, and several other categories with fewer than 11 respondents. Respondents could select more than one category. Among staff working at the 22 nonpediatric practices included in the impact evaluation, 145 staff responded to the survey. The response rate for staff at nonpediatric practices is not available; we do not know how many respondents worked in nonpediatric practices.

More than 80 percent of the 145 respondents at the 22 nonpediatric practices reported receiving formal training (data not shown), defined as web- or classroom-based learning. (Informal training included training that occurred during staff meetings, team huddles, and at the point of care.) Of staff who reported receiving formal training, most participated in trainings about screening tools, cultural mindfulness, and motivational interviewing (Table III.3). The types of training reported in Table III.3 reflect the composition of the survey respondents. For example, given that LPNs made up a fifth to a quarter of respondents and that their role in the award was to administer BH-6 screening, it is not surprising that BH-6 and other screening training was the most commonly reported type of training. In addition, screening was one of the key program activities in general.

Consistent with Sanford Health's approach of creating a training program for new and existing staff, 77 percent of respondents who received formal training described it as "additional training for an existing position." Most respondents who received formal training characterized the trainings as good or excellent (74 percent), strongly or somewhat agreed that the topics covered in formal trainings were relevant (85 percent), somewhat or strongly agreed that the training would be useful to their work (88 percent), and somewhat or strongly agreed that the training helped improve their performance or complete new job responsibilities (75 percent) (data not shown).

Table III.3. Percentage and number of surveyed staff who received each type of Sanford Health One Care training, for 22 nonpediatric practices included in the impact evaluation

Type of training	Percentage of staff (number)
BH-6 and other screenings	74.4% (87)
Cultural mindfulness	66.7% (78)
Motivational interviewing	53.0% (62)
Medical home and chronic disease management	46.2% (54)
Mental/behavioral health integration	40.2% (47)
Patient Activation Measure	38.5% (45)
Team-based care	35.9% (42)
Chronic care professional	34.2% (40)
Full clinic presentation	22.2% (26)
Trauma-informed care	21.4% (25)
Clinical practice guideline training	18.0% (21)
Adaptive leadership training	--[a]

Source: Mathematica's analysis of trainee survey data. Data are limited to respondents at the 22 nonpediatric practices included in the impact evaluation. The trainees included the following staff roles as defined in the survey, listed in the order of the number of respondents: LPNs, RNs, physicians, receptionists/registrars, others, health coaches, and several other categories with fewer than 11 respondents.

[a] Not reported because fewer than 11 respondents participated in this type of training.

BH-6 = 6-item Behavioral Health Screener; LPN = licensed practical nurse; RN = registered nurse.

Table III.4 summarizes the perceptions about training for staff who received formal or informal training. Of the 99 trainees, most thought that training had a positive effect on the quality of care they provided, patient-centeredness of care, ability to respond in a timely way to patients' needs, relay relevant information to the care team, access the care they need, explain information about patients care to patients and their families in lay terms, work with a diverse set of patients, and help patients access nonmedical services. (Table III.4). Rarely did trainees perceive negative effects of the training. Nineteen or more respondents did not answer the questions shown in Table III.4, which accounts for most of the responses not displayed in the table. A handful of trainees perceived negative impacts on cost-effectiveness of care and on providers' ability to respond in a timely way to patients' needs; however, in both cases, the number of trainees reporting a negative effect was fewer than 11.

Table III.4. Trainee survey respondents' perceptions of the effects of training on their care in 22 nonpediatric practices included in the impact evaluation

Survey question		Percentage (number) of respondents who reported the training had a positive effect on this dimension of their care[a]	Percentage (number) of respondents who reported the training had a no effect on this dimension of their care or that it was too soon to tell[a]
Please indicate the impact you believe the training you received for the One Care or medical home program has had on the following aspects of care you provide to patients enrolled in Sanford Health	Equity	63% (62)	16% (16)
	Patient-centeredness	62% (61)	16% (16)
	Quality of care	60% (59)	20% (20)
	Ability to respond in a timely way to patients' needs	55% (54)	21% (21)
	Efficiency/cost-effectiveness of care	38% (38)	35% (35)
Please indicate whether the training you received has had a positive or negative effect on your ability to:	Relay relevant information to the care team	60% (59)	21% (21)
	Access the care they need	59% (58)	18% (18)
	Explain information about patients' care to patients and their families in lay terms	55% (54)	24% (24)
	Work with a diverse set of patients	54% (53)	25% (25)
	Help patients access nonmedical services	51% (50)	26% (26)
	Help patients take control of their own care	49% (49)	30% (30)
	Use data to evaluate my performance to improve the services I provide to patients	48% (48)	29% (29)

Source: Mathematica's analysis of trainee survey data.

[a] The denominators include 99 trainees who reported that they received some training (formal or informal) for the Sanford Health One Care or Medical Home program and that their role in the One Care program was physician, nurse practitioner, registered nurse, licensed practical nurse, patient navigator, community health worker, behavioral health staff, or health coach.

Table III.5 displays care management activities of 99 trainee survey respondents in 22 nonpediatric practices included in the impact evaluation. In general, the way trainees reported spending their time aligned with the One Care model. More than half of trainees reported that they routinely educated patients about managing their own care, and more than a quarter said that they spent at least two hours per day on the task. We observed similar results for other health coaching and care management activities, such as calling patients to check on medications and counseling patients on exercise, nutrition, and how to stay healthy. These responses aligned with the program's emphasis on health coaching and care management. More than a third (35 percent) of respondents reported routinely spending time coaching patients. Nearly 20 percent of trainees spent more than two hours per day on the task. More than a quarter (27 percent) reported routinely spending time in team meetings and care conferences, which we would have expected to be higher given Sanford Health's emphasis on training and staff engagement. Even though the activities shown in Table III.5 are most relevant to the work of RN health coaches, we could not show their responses alone because many questions had fewer than 11 respondents.

Table III.5. Trainee survey respondents' care management activities in 22 nonpediatric practices included in the impact evaluation

Activity	Percentage (and number) of respondents who reported that they:[a]	
	Personally help to manage patients' care through this activity *routinely*	Spend more than 2 hours on this activity on a typical work day
Educate patients about managing their own care	61% (60)	27% (27)
Execute standing orders for medication refills, ordering tests, or delivering routine preventive services	58% (57)	24% (24)
Call patients to check on medications, symptoms, or help coordinate care between visits	59% (58)	24% (24)
Counsel patients on exercise, nutrition, and how to stay healthy	53% (52)	22% (22)
Patient coaching	35% (35)	19% (19)
Attend team meetings/care conferences	27% (27)	--[b]
Follow up on transitions of care	19% (19)	--[b]
Assist patients with accessing nonmedical services such as housing, job training, or supplemental nutrition services (for example, SNAP benefits)	16% (16)	--[b]
Attend medical appointments with patients	--[b]	--[b]
Conduct home visits with patients	--[b]	--[b]

Source: Mathematica's analysis of trainee survey data.

[a] The denominators include 99 trainees who reported that they received some training (formal or informal) for the Sanford Health One Care or Medical Home program and that their role in the One Care program was physician, nurse practitioner, registered nurse, licensed practical nurse, patient navigator, community health worker, behavioral health staff, RN health coach, or health educator.

[b] Not reported because fewer than 11 respondents reported yes.

RN = registered nurse; SNAP = Supplemental Nutrition Assistance Program.

5. Program timeline

Sanford Health implemented most project activities in accordance with its planned timeline, including incorporating clinical guidelines, hiring and training staff, and convening core teams at each practice. Three elements of the intervention took longer than expected: incorporating new screening tools, integrating new staff and processes into Bemidji practices, and implementing remote monitoring devices.

Integrating BH-6 into the EHR took longer than expected due to technical challenges. Sanford Health originally planned to incorporate screening tools into the EHR by July 2013, but it was completed in the second quarter of 2014 after a successful pilot at one participating practice. Sanford Health also faced delays implementing the PAM, reporting limited use in One Care practices, and at the end of the award considered creating a task force to support ongoing PAM implementation.

In Bemidji practices, it took longer than expected to conduct outreach to local Native American communities. Rather than incorporating cultural advisors directly into care teams as originally intended, Sanford Health focused on building cultural awareness and sensitivity among staff and providing outreach to the Native American community. As of June 2015, a CHW in a Bemidji practice was engaging Native American patients via Better Choices Better Health workshops, held in community centers. Sanford Health also completed an inspirational video and distributed it to all 33 participating practices. The video featured the Native American Olympic gold medal winner Billy Mills, who has diabetes. The video took longer than expected to complete due to licensing delays related to the use of Olympic footage.

The remote blood pressure cuffs and scales took longer than expected to distribute. Sanford Health started distributing them in the last calendar quarter of 2014. Challenges with the devices included small cuffs; difficulties transmitting data due to lack of Internet service or inability to register devices; and patients choosing not to automatically transmit data. RN health coaches reported that they began collecting data by telephone or email in early 2015 and the process ramped up in the following months. Despite the challenges and delays with implementation, RN health coaches found the collected data to be helpful in managing patients' conditions.

C. Summary of facilitators and barriers to implementation

Several factors facilitated implementation of Sanford Health's HCIA-funded intervention, whereas other factors hindered implementation. We described those factors in detail in the second annual report (Wells et al. 2016). Here we summarize key facilitators and barriers, along with any new information since the second annual report that supports those facilitators or barriers (Table III.6). Some of the survey questions discussed in this section do not focus on specific barriers and facilitators; rather, they provide contextual information about the clinical environment in which the program was implemented.

Table III.6. Summary of key facilitators and barriers to the implementation of Sanford Health's program

Item	Description based on findings in the second annual report	Additional supporting data not available in the second annual report, if applicable
	Facilitators (domain)	
Staff engagement (implementation process)	Staff engagement was consistently cited as an element of successful implementation. Practices engaged staff at all levels, including registrars, LPNs, administrative staff, RN health coaches, BHTTs, and physicians. To increase buy-in from the ground up, leadership designated physician champions, many of whom helped develop trainings, to effectively engage other physicians at their sites. Practices developed core teams of multidisciplinary stakeholders to facilitate program implementation by fostering buy-in and reinforcing the purpose of the intervention. Regular core team meetings, informal coffee breaks, and intentional seating of BHTTs and RN health coaches in office space frequented by clinicians helped build team integration among various staff types and role-specific meetings helped hone skills within individual roles.	Most trainee survey respondents agreed or strongly agreed that management was supportive of them (88 percent) and that they felt encouraged by supervisors to offer suggestions and improvements (81 percent). Clinician survey results suggested that clinicians were engaged with care teams, with 96 percent agreeing or strongly agreeing that the care team relayed relevant information in a timely manner.
Practice-level flexibility (program characteristic)	Sanford Health staff cited practice-level flexibility as a facilitating factor. For instance, variations in spatial layout, number of staff, and number of appointments per day influenced teams' ability to formally huddle on a scheduled basis, and some teams instead communicated informally. Although all practices administered the BH-6 and increased behavioral health integration, some practices chose to address some medical conditions first, especially asthma, diabetes, and hypertension.	None
Perceived relative advantage (program characteristic)	Staff at all levels agreed that integrated, team-based care was better than traditional volume-based care for patients and physicians. Many staff regarded BHTTs as critical to practice transformation and noted that their addition to the primary care team facilitated comprehensive care. Physicians' buy-in increased when they began to see successes among patients who had interacted with RN health coaches and BHTTs. New patient reporting, such as reports showing upward trends in patients' asthma control test scores, also facilitated support for the One Care model. Staff at all levels expressed satisfaction with their jobs and unwillingness to return to nonintegrated care.	More than half (57 percent) of trainees said that clinicians' resistance was not a barrier to the program, suggesting that most clinicians bought into the One Care model.
Patient engagement (implementation process)	Staff viewed patients' engagement as critical to achieving program goals. RN health coaches found motivational interviewing particularly useful for helping patients set manageable goals. Staff at some practices found PAM very helpful to determine how to best serve a patient based on his or her motivation and knowledge. Patients' engagement also included educating them about their conditions and resources available in the community, and offering patients the MyChart platform to communicate with staff and monitor their health. Staff attributed this new patient engagement approach to greater perceived success in patients' goal attainment, but emphasized that progress could be slow.	Most clinicians agreed or strongly agreed that information regarding patients' care was explained to patients and their families in lay terms (97 percent) and that when communicating with patients, members of the care team allowed enough time for questions (92 percent). Clinicians also reported that patients could request

Table III.6 *(continued)*

Item	Description based on findings in the second annual report	Additional supporting data not available in the second annual report, if applicable
		appointments or referrals online (100 percent), request prescription refills online (99 percent), and email clinicians about a medical question or concern (100 percent). Most clinicians (89 percent) responded that it was extremely important to improve patients' capacity to manage their own care.
Prior history (internal factors)	Sioux Falls practices had RN health coaches in place before the award, whereas Fargo internal medicine practices had limited experience with behavioral health (but no BHTTs). In both regions, this experience provided the institutional knowledge to integrate new team members and enabled regions to learn from one another. Internal medicine practices in the Fargo and Minnesota regions were also certified Minnesota medical homes, and staff at these practices saw One Care as harmonious with their medical home model. Practices without applicable prior experience, such as family medicine practices in Fargo, did not have the benefit of prior experience, but they were able to learn from Fargo internal medicine clinics, which were often collocated.	None
	Barriers (domain)	
Payment models (external environment)	Sanford Health's clinician payment model created challenges for sustaining One Care. Physicians and administrators cited challenges transitioning to value-based care under a volume-based payment model. Sustaining new nonbillable services required Sanford Health to absorb new costs associated with One Care, such as nonbillable BHTT triage services, in its operational budget. At the end of the award, Sanford Health was exploring alternative compensation models to help mitigate this barrier, such as weighting value more heavily in physicians' salaries.	Most clinicians responded that the level of reimbursement was not adequate for the time required to provide optimal patient-centered care, with 34 percent characterizing it as somewhat limiting and 48 percent characterizing it as limiting a great deal.
Cultural attitudes about alcohol and drug abuse (external environment)	Care team members felt that they lagged in identifying those with alcohol and drug abuse issues. They attributed this lag to cultural attitudes about alcohol, including both patients' and physicians' attitudes. Physicians were reluctant to focus on alcohol and drug abuse unless they felt they had the resources to address the issue with their patients. Hiring the addiction navigator in Sioux Falls was a response to this perceived gap.	None

Note: We reviewed four domains associated with implementation experience: (1) program characteristics, (2) implementation process, (3) internal factors, and (4) external environment. Implementation research suggests that barriers and facilitators within these domains are important determinants of implementation effectiveness.

BH-6 = 6-item Behavioral Health Screener; BHTT = behavioral health triage therapist, LPN = licensed practical nurse; PAM = Patient Activation Measure; RN = registered nurse.

Five factors were particularly important in facilitating program implementation. First, extensive staff engagement facilitated implementation because staff at all levels developed a common understanding of One Care and communicated regularly to continuously improve care. Second, practice-level flexibility facilitated implementation, enabling practices to customize the intervention to their specific environments, such as selecting target conditions to address first and deciding whether and how often to huddle with teams. Third, staff agreed that One Care represented an improvement over previous approaches to care by integrating behavioral health staff. Fourth, patient engagement via motivational interviewing, use of PAM, and MyChart also facilitated implementation, helping RN health coaches and other staff work with patients to set achievable goals and more effectively self-manage conditions. Finally, staff noted that practices with pre-HCIA experience with certain intervention components facilitated implementation and learning across practices.

Two important barriers to implementation included payment models and cultural attitudes about alcohol and drug abuse. Staff reported feeling challenged to provide value-based care in a largely volume-based payment model that encouraged short visits with many patients and discouraged nonreimbursable activities such as team huddles, BHTT triage, and panel management. Staff also reported challenges discussing alcohol and drug abuse with patients, who were reluctant to disclose and/or alter consumption.

D. Conclusions about the extent to which the program, as implemented, reflect core design

Sanford Health implemented One Care largely as intended. By June 2015, Sanford engaged 290 staff in seven key positions on care teams and in workforce development, nearly achieving Sanford Health's goal of 325 staff. Sanford Health successfully provided various trainings to intervention staff, which aligned with the award's emphasis on workforce development. Most trainees described trainings as relevant and useful to their work. Trainee survey results suggest that about 60 percent of trainees routinely educated patients about managing their own care, and that more than a quarter of those who routinely educated patients spent at least two hours per day on the task. Sanford Health also observed an increase in patients' use of MyChart, suggesting greater patient engagement.

All 33 One Care practices successfully integrated BHTTs into primary care, who screened patients for behavioral health conditions and provided behavioral health services to patients with depression, anxiety, and alcohol and drug abuse. Sanford Health met its goals for identifying patients with depression and anxiety, but not for identifying patients with alcohol and drug abuse. Sanford Health experienced some delays incorporating behavioral health screenings, implementing remote monitoring devices, and introducing outreach to Native American patients in Bemidji. However, these delays did not significantly impede implementation of the program's core design.

One Care practices also successfully incorporated RN health coaches to manage patients' medical conditions. RN health coaches cited motivational interviewing as facilitating their ability to effectively engage patients to manage their conditions, but emphasized that progress could be

slow. Even though remote monitoring devices faced some early delays and technical challenges, staff reported that they facilitated management of obesity and hypertension.

IV. CLINICIANS' PERCEPTIONS OF PROGRAM EFFECTS ON THE CARE THEY PROVIDE TO PATIENTS

This section describes the available evidence on the extent to which Sanford Health's intervention had its intended effects on changing clinicians' behavior as a way to achieve desired impacts on patient outcomes. As described in Section III.A.3, the program's theory of action required that clinicians (1) develop and/or participate in staff trainings, (2) develop and/or use clinical practice guidelines that incorporate behavioral health, and (3) coordinate with RN health coaches and BHTTs to help patients manage their medical and behavioral conditions. We use data from two rounds of the HCIA Primary Care Redesign Clinician Survey and from Sanford Health on clinician engagement in One Care to assess changes in providers' behavior and conclude whether the anticipated changes occurred. The analysis relies on self-reported survey responses and reflects clinicians' perceptions of the program on providers' behavior and patients' care, rather than measuring quantitatively direct program effects on their care. We supplemented these survey results with qualitative information we collected during interviews with care team members.

A. Clinician survey

Survey methods. We administered the clinician survey in two rounds (fall 2014 and summer 2015). Administration of the first round of the survey corresponded to the middle of the primary test period for the quantitative evaluation for cohort one practices and slightly before the primary test period for cohort two practices. The second round corresponded to the end of the primary test period for practices in both cohorts. We sent the survey to clinicians at all 33 participating practices. A total of 123 and 128 clinicians participating in Sanford Health's HCIA program responded to the survey during the first and second rounds, respectively (a response rate of 67 percent in round 1 and 65 percent in round 2). To align with our impact evaluation, which focuses on adult Medicare beneficiaries, we included responses from clinicians in 22 nonpediatric practices with baseline data. This resulted in a sample size of 99 respondents for round 1 and 80 for round 2. The response rate for nonpediatric staff working at nonpediatric practices is not available.

Survey results. Many surveyed clinicians reported being somewhat or very familiar with the HCIA program (74 percent in round 1 and 76 percent in round 2). Nearly all surveyed clinicians reported working in care teams, a central component of One Care (91 percent in round 1 and 90 percent in round 2) (data not shown). Of the clinicians who reported that they were at least somewhat familiar with the HCIA program, the proportion reporting positive impacts of the intervention on all dimensions of care increased during the intervention (Table IV.1). By the second round of the survey, 66 percent of clinicians familiar with the program reported that the intervention improved patient-centeredness of care, 62 percent reported that the intervention improved quality of care, 59 percent reported that the intervention improved their ability to respond to patients' needs in a timely way (possibly related to support from RN health coaches and BHTTs), and 57 percent reported that the intervention improved access to information

available for clinical decision making (possibly related to panel manager/CCAs' support in using data more effectively). About half of respondents reported a positive effect on efficiency and safety of care and fewer than half on equity of care.

Table IV.1. Clinicians' perceptions of the effects of the program on the care they provided to patients

	Percentage (and number) of clinicians familiar with HCIA reporting that the HCIA had the following effect on the care they provided to patients enrolled in their practice in the past year			
	First round of survey (17 to 19 months after cohort one and 8 to 10 months after cohort two implementation) N = 73		Second round of survey (25 to 27 months after cohort one and 16 to 18 months after cohort two implementation) N = 61	
Dimension of care	Positive impact	No impact or too soon to tell	Positive impact	No impact or too soon to tell
Patient-centeredness	51% (37)	44% (32)	66% (40)	33% (20)
Quality	48% (35)	47% (34)	62% (38)	38% (23)
Ability to respond in a timely way to patients' needs	48% (35)	48% (35)	59% (36)	25% (41)
Information available for clinical decision making	n.a.[a]	n.a.[a]	57% (35)	41% (25)
Efficiency	32% (23)	53% (39)	49% (30)	38% (23)
Safety	41% (30)	55% (40)	49% (30)	48% (29)
Equity	33% (24)	62% (45)	43% (26)	51% (31)

Source: Clinician Survey Round 1 (field period September to November 2014) and Round 2 (field period May to July 2015)

Note: The numbers (and percentages) are limited to clinicians who reported that they were at least somewhat familiar with the HCIA program.

[a] The first survey round did not ask this question.

n.a. = not applicable.

B. Sanford Health data on clinician behavior

During our site visits, BHTTs and RN health coaches reported that clinicians increasingly referred patients to them, as clinicians bought into team-based care with integrated behavioral health and care management. Physicians and advanced practice providers also emphasized the importance of BHTTs and RN health coaches in providing comprehensive primary care. Several physicians expressed an unwillingness to return to the pre-HCIA model of care, often citing the relative advantage of having BHTTs to address patients' behavioral health needs. We did not collect information on providers' use of the newly developed clinical practice protocols.

C. Conclusions about intermediate program effects on clinician behavior

Based on available information, the HCIA-funded initiative appears to have had moderate effects on how clinicians provide care. About three-quarters of surveyed clinicians said they

were aware of the One Care program, most of whom said they believed that the program improved the patient-centeredness and quality of care. Sanford Health's workforce development and staff engagement efforts increased buy-in among many clinicians, leading to greater participation in team-based activities, such as referrals to BHTTs and RN health coaches. However, about a quarter of clinicians reported that they were not familiar with the HCIA initiative. About half believed that the program had no impact on safety and equity, or that it was too soon to tell. Finally, we do not have information to assess the extent to which clinicians used clinical practice protocols or increased referrals to RN health coaches and BHTTs, both important activities according to the awardee's theory of action.

V. PROGRAM IMPACTS ON PATIENTS' OUTCOMES

This section of the report draws conclusions, based on available evidence, about the impacts of Sanford Health's HCIA program on patients' outcomes in four domains: quality-of-care processes, quality-of-care outcomes, service use, and spending. We first describe the methods for estimating impacts (Section V.A) and then the characteristics of the 22 nonpediatric HCIA treatment practices at the start of the intervention (Section V.B). We next demonstrate that the treatment practices were similar at the start of the intervention to the practices we selected as a comparison group, which is important for limiting potential bias in impact estimates (Section V.C). Finally, in Section V.D, we describe the quantitative impact estimates, their plausibility given implementation findings, and our conclusions about program impacts in each domain. The findings in this report update the impact results for Sanford Health from the Second Annual Report (Wells et al. 2015), by adding cohort two practices, extending the outcome period by 6 months, and adding new outcomes.

A. Methods

1. Overview

We estimated program impacts on patients' outcomes as the difference in outcomes for Medicare FFS patients served by the 22 nonpediatric primary treatment practices with baseline data and those served by 91 nonpediatric matched comparison practices, adjusting for any differences in outcomes between these groups during the year before the intervention began. We prespecified primary tests, describing the evidence we would need to conclude that the program was effective, and the awardee and CMMI reviewed these. Each test specified a population, outcome, time period, expected direction of effect, and threshold that we counted as substantively important. The purpose of these primary tests was to focus the impact evaluation on hypotheses that would provide the most robust evidence about program effectiveness. We used the results from the primary tests and secondary tests (robustness checks) to draw conclusions about program impacts in each of the four evaluation domains. The remaining subsections describe each component of the impact evaluation in more detail.

2. Treatment group definition

Treatment practices joined the program at different times and, to simplify our analysis, we grouped them into two cohorts: (1) practices that joined the program from April 1, 2013, to December 31, 2013 (cohort one); and (2) practices that joined the program from January 1, 2014,

to December 31, 2014 (cohort two). The treatment group consists of Medicare FFS patients served by the 22 nonpediatric treatment practices in four baseline quarters before the intervention began (for cohort one practices April 1, 2012, to March 31, 2013; for cohort two practices January 1, 2013, to December 31, 2013). We followed cohort one practices for nine intervention quarters (April 1, 2013, to June 30, 2015) and cohort two practices for six quarters (January 1, 2014, to June 30, 2015). Our treatment group consists exclusively of Medicare FFS beneficiaries served by practices that treated adults, whereas the awardee's target population consists of all Medicare, Medicaid, and CHIP beneficiaries served by treatment practices. Limitations in Medicare managed care administrative data and lags in Medicaid and CHIP data availability prevented us from conducting tests of effectiveness on these populations. As a result, we excluded nine pediatric practices from the evaluation. We also excluded two practices that were newly founded during the intervention period and thus did not have baseline data.

We constructed the treatment group in four steps.

1. First, we attributed beneficiaries to practices using the same decision rule that CMMI uses for the Comprehensive Primary Care Initiative. Specifically, in each baseline and intervention month, we attributed beneficiaries to the primary care practice whose providers (physicians, nurse practitioners, or physician assistants) provided the plurality of primary care services in the past 24 months. If there was a tie, we attributed beneficiaries to the practice they visited most recently. Sanford Health provided data on which providers worked in the treatment practices and when.
2. Second, in baseline and intervention periods, we assigned each patient to the first treatment practice he or she was attributed to in that period, and continued to assign him or her to that practice for all quarters in the period. This assignment rule, which is distinct from the attribution method, ensured that, during the intervention period, patients did not exit the treatment group solely because the intervention succeeded in reducing their service use (including visits at treatment practices). The definition for the baseline period corresponds to that of the intervention period so that, across the two periods, interpretation of the population changes over time is comparable.
3. Third, we limited the analytic population to beneficiaries targeted by Sanford Health's program. The program specifically targeted improvements in services delivered to patients with one of eight chronic health conditions: anxiety, asthma, diabetes, depression, heart failure, hypertension, obesity, and substance abuse (alcohol or drug abuse). Sanford Health identified patients with these conditions based on diagnosis codes in administrative data.

   In this report, we present results for the group of beneficiaries with at least one of the eight targeted conditions, and refer to its members as *targeted beneficiaries*. We identified beneficiaries who had one or more of the eight conditions at the start of baseline and intervention periods. We used the Chronic Conditions Data Warehouse (CCW) (2016b) algorithms to identify patients with all conditions, except obesity, for which we used the list of diagnoses provided by Sanford Health. To flag patients with alcohol and drug abuse, we used the CCW's draft algorithms published at the end of 2014 after public comments were incorporated. The CCW algorithms generally corresponded well to the criteria used by Sanford Health to identify these conditions. For alcohol and drug abuse, the list of

conditions corresponded well, except that CCW used procedure codes related to treatment of alcoholism and diagnosis and treatment of conditions caused by alcoholism, whereas Sanford Health's algorithm used only diagnosis codes. Given that procedure codes are likely duplicative of diagnoses codes, we believe the Sanford Health and our identification criteria correspond well. Sanford Health generally used a look-back period of 12 months, whereas the CCW algorithm looks further back for some of the conditions.

4. Fourth, we applied additional restrictions to define the final analysis sample in each quarter. A beneficiary assigned to a treatment practice in a quarter was included in the analysis sample for that quarter if he or she (1) had observable outcomes for at least one day in the quarter; and (2) lived for at least one day in one of the states with participating practices (Minnesota, North Dakota, or South Dakota) or neighboring states (Iowa or Nebraska).

3. Comparison group definition

The comparison group consists of Medicare FFS beneficiaries whom we assigned to 91 matched comparison practices in each of the baseline and intervention quarters. The comparison practices were similar to the treatment practices during the baseline period on factors that can influence patients' outcomes and factors that influence the decision to participate in the program. This section describes how we constructed the matched comparison group; Section V.C shows the balance we achieved between the two groups on the matching variables.

We identified the 91 comparison practices in four steps:

1. First, we identified a pool of 997 nonpediatric, nonparticipating potential comparison practices located in the three states with participating Sanford Health practices: Minnesota, North Dakota, and South Dakota.
2. Second, we developed matching variables, defined at the start of the intervention, for all treatment and potential comparison practices. These variables included characteristics of the practices (for example, the number of providers and whether the practice was owned by a hospital or health system); characteristics of all Medicare FFS beneficiaries assigned to the practices (for example, average Medicare spending in the past year and the percentage of attributed beneficiaries who are Native American); and characteristics of assigned Medicare FFS beneficiaries with at least one of the targeted conditions. We did not include measures of quality-of-care processes in matching because, when we completed matching (spring 2015), these measures were not yet available. When assigning Medicare beneficiaries to practices, we used the same attribution and practice assignment logic that we used for the treatment practices, as described previously. Section V.C describes the matching variables in detail.
3. Third, we narrowed the potential comparison pool of practice from 997 to 465 by excluding potential comparison practices with characteristics not observed among the treatment group. We excluded (1) Indian Health Service practices; (2) practices that did not accept Medicaid; (3) practices not owned by a hospital or health system or part of a medical group; and (4) practices with very high or very low values for key matching variables, such as practice

size and service utilization. After we applied these restrictions, 465 potential comparison practices remained available for matching.

4. Finally, we used propensity-score methods to select 91 comparison practices from the pool of 465 that were similar to the 22 nonpediatric treatment practices on the matching variables. The propensity score is the predicted probability, based on all of a practice's matching variables, that a given practice was selected for treatment (Stuart 2010). It collapses all of the matching variables into a single number for each practice that can be used to assess how similar practices are to one another. By matching each treatment practice to one or more comparison practices with similar propensity scores, we generated a comparison group that was similar, on average, to the comparison group on the matching variables.

We required each treatment practice to match to at least one, but no more than five, comparison practices and that the overall ratio of comparison to treatment practices be at least 3:1. This matching ratio increased the statistical certainty in the impact estimates (relative to 1:1 overall matching ratio), because it created a more stable comparison group against which to compare the treatment group's experiences.

After completing the matching, we assigned Medicare FFS beneficiaries to the comparison practices in each intervention quarter using the same rules we used for the treatment group (see Section V.A.2).

We included in the potential comparison group Sanford Health's nonparticipating practices, because they were more likely to be similar to participating practices than non-Sanford Health practices, given that Sanford Health serves a large part of the region and owns many practices. Through the matching process, we ended up selecting eight Sanford Health's nonparticipating practices to be included in the final set of 91 comparison practices. Two concerns arise for allowing Sanford Health's nonparticipating practices to serve as comparisons: (1) even though they were similar to participating practices on observable characteristics, nonparticipating Sanford Health practices differed in that they were not selected for participation in the award—that is, they might differ from participants on unobservable characteristics; and (2) it is possible that Sanford Health's intervention had been extended to nonparticipating practices. If true, both these concerns could contribute to a bias in impact estimates. However, we do not believe that the risk of bias is substantial. There is no evidence that the 33 participating practices were selected based on motivation or another unobservable characteristic that could affect outcomes. Further, there is no evidence that the HCIA-funded intervention was extended to Sanford Health's nonparticipating practices during the award period. Therefore, the risk of bias in impact estimates is small.

4. Construction of outcomes and covariates

We used Medicare claims from April 1, 2009, to June 30, 2015, for beneficiaries assigned to the treatment and comparison practices to develop two types of variables: (1) outcomes, defined for each person in each baseline or intervention quarter; and (2) covariates, which describe a beneficiary's characteristics at the start of the baseline and intervention periods and are used in the regression models for estimating impacts to adjust for beneficiaries' characteristics before the

period began. We used covariates defined at the start of each period, without updating them each quarter, to avoid controlling in each intervention quarter for previous quarters' program effects, as this would bias the effect estimates away from detecting true impacts. Appendix 1 provides details on the methods we used to construct these variables.

Outcomes. For each person, we calculated seven outcomes that we grouped into four domains:

1. Domain: Quality-of-care processes
    a. Diabetes quality-of-care composite (binary variable for each beneficiary); calculated as whether a beneficiary with diabetes had had all four recommended tests—lipid profile, hemoglobin A1c test, dilated eye exam, and nephropathy screening—during the previous 12 months
    b. Ambulatory-care follow-up visit within 14 days of a hospital discharge (binary variable for each beneficiary); calculated as whether all of a patient's discharges in a quarter were followed by an ambulatory visit with a primary care or specialist physician within 14 days of the discharge
2. Domain: Quality-of-care outcomes
    a. Inpatient admissions (number/quarter) for ambulatory care-sensitive conditions (ACSCs)
    b. Number of inpatient admissions followed by an unplanned readmission within 30 days (number/quarter)
3. Domain: Service use
    a. All-cause inpatient admissions (number/quarter)
    b. Outpatient ED visit rate (number/quarter); outpatient ED visits are defined as ED visits or observational stays that do not end in a hospital admission
4. Domain: Spending
    a. Total Medicare Part A and B spending (dollars/month)

Four of these outcomes—all but ACSC admissions and the two quality-of-care process measures—are outcomes that CMMI has specified as core for the evaluations of all HCIA programs. Our definition of the readmission measure, however, differs from CMMI's standard definition. CMMI typically defines readmissions as the proportion of inpatient admissions that end in an unplanned readmission. Instead, we analyzed impacts on the *number* of these unplanned readmissions across all beneficiaries per quarter because this enables us to look at the total impact on readmissions across the treatment group, rather than readmissions contingent on an inpatient admission. We made this decision, in consultation with CMMI, because the intervention might also affect the number of and type of admissions.

All outcomes are quarter-specific—meaning that we calculated them for each baseline and intervention quarter separately—except for the quality-of-care process measure for diabetes. Because this measure assesses whether a beneficiary received recommended preventive care

services over a year-long period, we calculated this measure over the full years rather than quarters: namely, over the baseline year (that is, the period corresponding to the four baseline quarters), and over the first year of the primary test period (corresponding to the last four intervention quarters). We avoided calculating this measure for overlapping periods, meaning that no measurement year included services provided in another measurement year.

Finally, we defined all outcomes for all treatment and comparison group members, except for the two measures of quality-of-care processes. We calculated the measure of 14-day follow-up after discharge among only those patients with at least one hospital discharge in the relevant quarter. We calculated the diabetes composite measure among beneficiaries ages 18 to 75 with diabetes at the beginning of the period (baseline or intervention period).

Covariates. The covariates include (1) 22 indicators for whether a patient has each of the following chronic conditions: alcohol abuse, Alzheimer's and related dementia, anxiety, asthma, atrial fibrillation, bipolar disorder, cancer, chronic kidney disease, chronic obstructive pulmonary disease, depression, diabetes, drug abuse, heart failure, hip fracture, hyperlipidemia, hypertension, ischemic heart disease, obesity, osteoporosis, rheumatoid arthritis, schizophrenia, and stroke); (2) HCC score; (3) demographics (age, gender, and race or ethnicity); (4) whether a beneficiary is dually eligible for Medicare and Medicaid; and (5) original reason for Medicare entitlement (old age, disability, or end-stage renal disease). We defined all covariates as of the start of the relevant period (baseline or intervention).

5. Regression model

We used a regression model to implement the difference-in-differences design for estimating impacts. For each outcome, the model estimated the relationship between the outcome and a series of predictor variables, assuming that each of the predictor variables has a linear (additive) relationship with the outcome. The predictor variables included patient-level covariates (defined in Section V.A.4); whether the patient was assigned to a treatment or a comparison practice; an indicator for each practice (which accounted for differences between practices in their patients' outcomes at baseline); indicators for each post-intervention quarter (or, for the diabetes measure, for the final post-intervention quarter of the year-long measurement period); and an interaction of a beneficiary's treatment status with each post-intervention quarter (or, for the diabetes measure, the final post-intervention quarter of the year-long measurement period). Given that the pre-intervention differences in the percentage of dually eligible beneficiaries in treated versus selected comparison practices was larger than our goal of 0.25 standard deviations, we also included as predictor variables interactions between the indicator for dual eligibility and patient-level covariates, to help control for the differences between the treated and matched comparison populations.

The estimated relationship between the interaction term and the outcome in a given quarter is the impact estimate for that quarter (or, for the diabetes measure, for the year ending with that quarter). It measures the average difference between outcomes for beneficiaries assigned to the treatment and comparison practices during that period, subtracting out any differences between these groups during the four baseline quarters. By providing separate impact estimates for each intervention quarter (or year, for the diabetes measure), the model enables the program's impacts

to change the longer the practices are enrolled in the program. We can also test impacts over discrete sets of quarters or years, which is needed to implement the primary tests discussed in the next section. Finally, the model quantifies the uncertainty in the impact estimates, allowing for statistical tests that determine whether observed differences in outcomes between the treatment and comparison groups are likely due to chance. The model uses robust standard errors to account for clustering of outcomes across quarters for the same beneficiary and a dummy variable for each practice (fixed effects) to account for clustering of outcomes for beneficiaries assigned to the same practice. Appendix 2 provides details on the regression methods, including descriptions of the weights each beneficiary receives in the model.

6. Primary tests

Table V.1 shows the primary tests for Sanford Health, by domain. Each test specifies a population, outcome, time period, expected direction of effect, and threshold that we count as substantively important. The purpose of these primary tests is to focus the impact evaluation on hypotheses that will provide the most robust evidence about program effectiveness (see Appendix 4 for detail and a description of how we selected each test). We provided both the awardee and CMMI an opportunity to comment on the primary tests.

Our rationale for selecting these primary tests is as follows:

- **Outcomes.** Sanford Health's central goal was to reduce inpatient admissions for ACSCs; ED visit rates; and spending for Medicare, Medicaid, and CHIP beneficiaries. It also sought to improve clinical and intermediate outcomes that were not easily measured in claims, such as quality of life, functional status measures such as severity of targeted mental health conditions, and the number of encounters and screenings. The primary tests focus on outcomes that Sanford Health aimed to affect that were also measurable in Medicare claims data: admissions for ACSCs, all-cause hospital admissions, ED visits, and Medicare Part A and B spending. We also examine the effects on unplanned readmissions. Finally, we included two quality-of-care process measures that, based on the Sanford Health's theory of action, we think the program could improve: a composite measure for whether a beneficiary with diabetes received all of four recommended processes of care during the year (HbA1c test, lipid profile, dilated eye exam, and nephropathy screening) and receipt of a follow-up ambulatory care visit with a primary care or specialist provider within 14 days of hospital discharge. Even though Sanford Health did not set explicit targets for these two measures, the award targeted patients with diabetes and monitored the percentage of patients with diabetes whose care was optimally managed, based on treatment goals per the Minnesota Community Measurement specifications; for example, these included target values for HbA1c, low-density lipoprotein cholesterol, and blood pressure measurement. We included the unplanned readmissions measure in our primary tests because, even though transitional care was not a focus of Sanford Health's award, transitional care was one of the roles of RN health coaches; further, the awardee expressed that the readmission rate was a relevant outcome for the impact evaluation.
- **Time period.** Sanford Health expected small impacts during the first year and sizeable impacts in the second year of program implementation. For cohort one practices (those that

joined the program from April 1 to October 31, 2013), our primary tests cover the fifth through ninth quarters of the intervention (I5 through I9), corresponding to the period from April 1, 2014, to June 30, 2015, when the award ended. For cohort two practices (those that joined the program from January 1 to December 31, 2014), our primary tests cover the fifth and sixth intervention quarters (I5 and I6), corresponding to the period from January 1 to June 30, 2015, when the award ended. Most of the measures are defined quarterly so, to estimate impacts over the specified time period, we averaged the impact estimates for each quarter from I5 through I9. In contrast, the process of care measure for diabetes is defined over a year. For the diabetes quality-of-care process measure, which is defined annually, we estimated impacts only for cohort one practices. We estimated them over the last four quarters available (I6 through I9), which corresponds to the last year of the program. We did not estimate impacts on the diabetes measure for cohort two practices because we had only two quarters of primary test period data available for those practices before the intervention ended on June 30, 2015, rather than the full year needed to calculate outcomes.

- **Population.** For all but the quality-of-care process measures, the primary test population includes beneficiaries with one of eight targeted conditions: anxiety, asthma, diabetes, depression, heart failure, hypertension, obesity, and drug/alcohol abuse. The program targeted this group and provided more intensive services to them than to other patients and specified expected impacts for this targeted population. For the diabetes process of care measure, we limited the population to beneficiaries with diabetes who were observable in FFS claims for all 12 months of the measurement year and were 18 to 75 years old during that period. For the 14-day follow-up measure, we limited the sample in each quarter to those who had at least one index hospitalization during the quarter for which we could observe whether the person had a 14-day follow-up visit.

  Due to limitations in Medicare managed care data and lags in Medicaid and CHIP data, we did not include these populations in our primary tests.

- **Direction (sign) of the impact estimate.** For the quality-of-care process measures, we expect the impact estimate to be positive, signaling an increase in the percentage of people receiving recommended care. For all other outcomes, we expect the impact estimate to be negative, indicating a reduction in service use or overall expenditures.

- **Substantive thresholds.** Some impact estimates could be large enough to be policy relevant (to CMMI and other stakeholders) even if they are not statistically significant; for this reason, we have prespecified thresholds for what we call substantive importance. We express the threshold as a percentage change from the counterfactual—that is, the outcomes that beneficiaries in the treatment group would have had if they had not received the HCIA-funded intervention. We set a substantively important threshold of 15 percent for ED visits and inpatient admissions for ACSCs, which is 75 percent of Sanford Health's original estimate of 20 percent for these two outcomes. For Medicare Part A and B spending, we set a substantive threshold of 2.25 percent, or 75 percent of the anticipated 3 percent. The 15 percent threshold for the process-of-care measures is extrapolated from the literature (Peikes et al. 2011) because Sanford Health did not specify by how much it expected to improve these outcomes.

Table V.1. Specification of the primary tests for Sanford Health

Domain (number of tests in the domain)[a]	Outcome (units)	Time period for impacts (controlling for baseline differences)[b]	Population	Substantive threshold (expected direction of effect)[c]
Quality-of-care processes (2)	Received all four recommended diabetes processes of care in the year (binary [yes or no]/beneficiary/year)	Last year of intervention (corresponding to quarters 6 through 9)[d]	Medicare FFS beneficiaries ages 18 to 75 with diabetes and assigned to cohort one treatment practices	15.0 (+)
	All inpatient admissions within a quarter were followed by an ambulatory care visit with a primary care or specialist provider within 14 days (binary [yes or no]/beneficiary/year)	Average over intervention quarters 5 through 9 for cohort one practices; average over intervention quarters 5 through 6 for cohort two practices[e]	Medicare FFS beneficiaries with a targeted condition and assigned to treatment practices who had at least one hospital stay in the quarter	15.0 (+)
Quality-of-care outcomes (2)	Inpatient admissions for ambulatory care-sensitive conditions (#/beneficiary/quarter)	Average over intervention quarters 5 through 9 for cohort one practices; average over intervention quarters 5 through 6 for cohort two practices[e]	Medicare FFS beneficiaries with a targeted condition and assigned to treatment practices	15.0 (-)
	30-day unplanned hospital readmission rate (#/beneficiary/quarter)	Average over intervention quarters 5 through 9 for cohort one practices; average over intervention quarters 5 through 6 for cohort two practices[e]	Medicare FFS beneficiaries with a targeted condition and assigned to treatment practices	15.0 (-)
Service use (2)	All-cause inpatient admissions (#/beneficiary/quarter)	Average over intervention quarters 5 through 9 for cohort one practices; average over intervention quarters 5 through 6 for cohort two practices[e]	Medicare FFS beneficiaries with a targeted condition and assigned to treatment practices	15.0 (-)
	Outpatient ED visit rate (#/beneficiary/quarter)	Average over intervention quarters 5 through 9 for cohort one practices; average over intervention quarters 5 through 6 for cohort two practices[e]	Medicare FFS beneficiaries with a targeted condition and assigned to treatment practices	15.0 (-)
Spending (1)	Medicare Part A and B FFS spending ($/beneficiary/month)	Average over intervention quarters 5 through 9 for cohort one practices; average over intervention quarters 5 through 6 for cohort two practices[e]	Medicare FFS beneficiaries with a targeted condition and assigned to treatment practices	2.25 (-)

Table V.1 *(continued)*

Note: Sanford's One Care program targeted improvements in services delivered to patients with one of eight chronic health conditions: anxiety, asthma, diabetes, depression, heart failure, hypertension, obesity, and substance abuse. We grouped practices into two cohorts: (1) those that joined the program from April 1, 2013, to December 31, 2013 (cohort one); and (2) those that joined the program from January 1, 2014, to December 31, 2014 (cohort two).

[a] We adjusted the *p*-values from the primary test results for the multiple comparisons made within each domain, but not across domains.

[b] The regression models for estimating impacts controlled for pre-intervention differences between the treatment and comparison groups.

[c] The substantive threshold is the impact as a percentage of the counterfactual. The counterfactual is the outcome the treatment group would have had in the absence of the HCIA-funded intervention.

[d] For the diabetes quality-of-care process measure, which is defined annually, we estimated impacts only for cohort one practices. We did not estimate impacts on the diabetes measure for cohort two practices because we have only two quarters of primary test period data available for those practices before the intervention ended on June 30, 2015.

[e] For all but the diabetes quality-of-care process measure, we took the average across the quarterly impact estimates (quarters 5 through 9 for cohort one practices and quarters 5 and 6 for cohort two practices).

ED = emergency department; FFS = fee-for-service; HCIA = Health Care Innovation Award

7. Secondary tests (robustness checks)

We also conducted secondary quantitative tests to help corroborate the findings from the primary tests. This is important because some of the differences observed between the treatment and comparison groups in the primary test results could reflect limitations of the non-experimental impact evaluation design or random fluctuations in the data. We have greater confidence in the primary results if they are generally consistent with the expected broader pattern of results from the secondary tests.

Specifically, we estimated the program's impacts on all measures in the quality-of-care outcomes, service use, and spending domains for the targeted beneficiaries during an additional intervention period—the first year after the practice joined the intervention. Because Sanford Health expected small impacts in the first year and substantial impacts in the second year, the following pattern would be consistent with an effective program: smaller impacts in the first versus the second year of the program. In contrast, finding very large differences in outcomes (favorable or unfavorable) in the first year but not the second could suggest a limitation in the comparison group, not true program impacts.

8. Synthesizing evidence to draw conclusions

Within each domain, we drew one of five conclusions about program effectiveness, based on the primary test results, the results of secondary tests, and the plausibility of those findings given the implementation evidence:

1. Statistically significant favorable effect (the highest level of evidence)
2. Substantively important (but not statistically significant) favorable effect
3. Substantively important (but not statistically significant) unfavorable effect
4. No substantively large effect
5. Indeterminate effect

We cannot conclude that a program has a statistically significant unfavorable effect because, in consultation with CMMI, we decided to use one-sided statistical tests that do not test for evidence of unfavorable effects. We used one-sided tests to increase the probability that, if a program truly did have impacts, we would be able to detect them.

Appendix 3 describes our decision rules for each of the five possible conclusions. In short, we concluded that a program had a statistically significant favorable effect in a domain if (1) at least one primary test result in the domain was favorable and statistically significant, after adjusting the statistical tests to account for multiple tests (if applicable) within a domain; or (2) the average impact estimate across all primary tests in the domain was favorable and statistically significant. In both cases, we also had to determine that the primary test results were plausible given the results of the secondary tests and implementation evidence. We concluded that a program had a substantively important favorable effect if the average impact estimate in the domain was substantively important but not statistically significant, and if the result was plausible given the secondary tests and implementation evidence. In contrast, if the average

impact estimate was unfavorable (opposite the hypothesized direction), larger than the substantive threshold, and unfavorable effects were plausible given the other evidence, we concluded the program had a substantively important unfavorable effect. If the tests in a domain did not meet any of these criteria, we drew one of two conclusions. First, if the tests for at least one outcome in the domain (or all outcomes in the domain together) had sufficient statistical power to detect an impact of the size of the substantive threshold with at least 75 percent probability, we concluded that there was not a substantively large effect because we are reasonably confident that we would have detected such a large effect had there been one. Alternatively, if the power was not sufficient (less than 75 percent) to detect this type of impact, we concluded the impact in the domain was indeterminate. Indeterminate means either that the program truly did not have effects that were substantively large or that it did, but our statistical tests were not able to detect them.

B. Characteristics of the treatment group at baseline

This section describes the characteristics of the treatment group at the start of the intervention (April 1, 2013, for cohort one practices and January 1, 2014, for cohort two practices). We also show this information in the second column of Table V.2. (Table V.2 serves a second purpose—to show the equivalence of the treatment and comparison practices at the start of the intervention—which we describe in Section V.C.)

Characteristics of the practices overall. Because Sanford Health is a health system, all 22 practices included in the impact evaluation are regarded as being owned by a system. Eighty-six percent of treatment practices were located in urban areas, with 32 percent located in health professional shortage areas. Nearly all treatment practices (96 percent) had providers receiving payment from the Centers for Medicare & Medicaid Services for meaningful use of EHRs. Treatment practices had on average 11 providers and a vast majority of providers in these practices had primary care as their specialty (86 percent).

Characteristics of the practices' Medicare FFS beneficiaries. Treatment patients' characteristics were similar to the nationwide Medicare FFS averages. The HCC risk score was 1.1, close to the national average of 1.0. Patients in the treatment practices had nearly identical hospital admission rates as the national averages during the baseline period. The mean outpatient ED visit rate (115/1,000 beneficiaries/quarter) was higher than the national average of 105, whereas 30-day unplanned readmission rates and Medicare Part A and B spending were lower than the national averages. Targeted beneficiaries (those with at least one of eight targeted chronic conditions) had somewhat higher health care needs during the baseline period than all Medicare FFS beneficiaries assigned to treatment practices. Their mean HCC risk scores were somewhat higher than the mean for all treatment group members (1.2 versus 1.1). Further, they had about 17 percent higher all-cause inpatient admissions, 14 percent more outpatient ED visits, and 14 percent higher Medicare Part A and B spending.

Table V.2. Characteristics of treatment and comparison practices before the intervention start date (April 1, 2013, for cohort one and January 1, 2014, for cohort two practices)

Characteristics of practices	Treatment practices (N = 22)	Matched comparison group (N = 91)	Absolute difference[a]	Standard-ized difference[b]	Medicare FFS national average
Exact match variables[c]					
Indian Health Service practice (%)	0.0	0.0	0.0	n.a.	n.a.
Practice accepts Medicaid (%)	100.0	100.0	0.0	n.a.	n.a.
Located in an urban area (%)	86.4	86.4	0.0	n.a.	n.a.
Propensity-score matching variables[d]					
Characteristics of the practices overall					
Owned by a hospital or health system (%)[e]	100.0	90.0	10.0	0.485*	n.a.
Practice is part of a medical group (%)[e]	0.0	10.0	-10.0	-0.485*	n.a.
MAPCP demonstration participation (%)	54.5	53.9	-0.6	0.012	n.a.
Practice size (number of providers)	10.5	10.6	-0.1	0.006	n.a.
Meaningful use of EHRs (%)	95.5	94.4	1.1	0.051	n.a.
Providers in practice with a primary care specialty (%)	85.7	81.1	4.6	0.182	n.a.
Characteristics of practices' locations					
In primary care health professional shortage area (%)	31.8	32.1	-0.3	-0.006	n.a.
Characteristics of all Medicare FFS beneficiaries assigned to practices during the baseline year (April 1, 2012, to March 31, 2013, for cohort one and January 1, 2013, to December 31, 2013, for cohort two practices)					
Number of beneficiaries	1,109	892	217	0.334**	n.a.
HCC risk score	1.05	1.04	0.01	0.069	1.0
All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	74.43	74.29	0.14	0.010	74[f]
Outpatient ED visit rate (#/1,000 beneficiaries/quarter)	114.65	115.71	-1.07	-0.036	105[g]
Medicare Part A and B spending ($/beneficiary/month)	738	764	-26	-0.193	860[h]
30-day unplanned hospital readmission rate (%)	13.7	13.5	0.2	0.042	16.0[i]
30-day unplanned hospital readmission rate (#/1,000 beneficiaries/quarter)	9.35	9.47	−0.13	−0.031	n.a.
Inpatient admissions for ambulatory care-sensitive conditions (#/1,000/beneficiary/quarter)[j]	11.91	12.56	−0.65	−0.148	11.8[k]
Disability as original reason for Medicare entitlement (%)	22.7	22.4	0.3	0.040	16.7[l]
Percentage dually eligible for Medicare and Medicaid	15.7	13.9	1.8	0.289	22[m]
Age (years)	71.86	72.05	-0.20	-0.066	71[n]
Female (%)	60.1	58.0	2.1	0.290	54.7[l]
Percentage Native American or Alaska Native (%)	1.6	0.9	0.7	0.686***	n.a.

Table V.2 *(continued)*

Characteristics of practices	Treatment practices (N = 22)	Matched comparison group (N = 91)	Absolute difference[a]	Standard-ized difference[b]	Medicare FFS national average
Characteristics of targeted Medicare FFS patients attributed to practices during the baseline year (April 1, 2012, to March 31, 2013, for cohort one and January 1, 2013, to December 31, 2013, for cohort two practices)					
Number of targeted beneficiaries[j]	830	657	174	0.346**	n.a.
HCC risk score	1.18	1.17	0.01	0.015	1.0
All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	86.97	86.85	0.12	0.007	74
Outpatient ED visit rate (#/1,000 beneficiaries/quarter)	131.07	130.89	0.18	0.005	105
Medicare Part A and B spending ($/beneficiary/month)	841	867	-26	-0.170	860
30-day unplanned hospital readmission rate (#/1,000 beneficiaries/quarter)[j]	11.34	11.50	-0.16	-0.032	n.a
Inpatient admissions for ambulatory care-sensitive conditions (#/1,000/beneficiary/quarter)[j]	14.60	15.42	-0.82	-0.158	11.8
Variables not included in matching					
Characteristics of Medicare FFS patients assigned to panels during the baseline year who met diagnosis, age, and/or service use restrictions					
Receipt of all four recommended processes of care measures, among those with diabetes ages 18 to 85 at cohort one practices (%)[p]	45.9	44.7	1.2	0.106	NA
Receipt of an ambulatory care visit within 14 days of all hospital discharges in the quarter, among those with at least one discharge in the quarter (%)	56.1	62.4	-6.3	-0.596***	NA

Sources: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at CMS. Zip code (whether an urban zip code or health professionals shortage area) was merged from the American Community Survey ZIP Code Characteristics. Data on meaningful use of EHRs were merged from CMS.

Notes: The comparison group means are weighted based on the number of matched comparison practices per treatment practice. For example, if four comparison practices are matched to one treatment practice, each of the four comparison practices has a matching weight of 0.25.

Absolute differences might not be exact due to rounding.

[a] The absolute difference is the difference in means between the matched treatment and comparison groups.

[b] The standardized difference is the difference in means between the matched treatment and comparison groups divided by the pooled standard deviation of the variable. The standard deviation is calculated among the pooled treatment and matched comparison groups.

[c] Exact match means that Indian Health Service practices were excluded from our comparison practices; all practices also had to accept Medicaid, and we required that practices match on rural versus urban location.

[d] Variables that matched using a propensity score, which captures the relationship between a practice's characteristics and its likelihood of being in the treatment group.

Table V.2 *(continued)*

[e] Because we were unable to match within the 0.25 standard on several essential variables when requiring all comparison practices to be owned by a health system, we matched on whether a practice was owned by a health system *or* a medical group. The rationale is that medical groups, like health systems, can provide resources to practices that are not available to much smaller, independent practices.

[f] Health Indicators Warehouse (2014b).

[g] Gerhardt et al. (2014).

[h] Boards of Trustees (2013).

[i] Centers for Medicare & Medicaid Services (2014).

[j] These measures were included in the table for descriptive purposes but were not included in the matching model.

[k] This rate is for individuals ages 65 and above (Truven Health Analytics 2015).

[l] Chronic Conditions Data Warehouse (2016a, Table A.1).

[m] Health Indicators Warehouse (2014c).

[n] Health Indicators Warehouse (2014a).

[o] Targeted beneficiaries are those with one or more of eight chronic health conditions targeted by the One Care program: anxiety, asthma, diabetes, depression, heart failure, hypertension, obesity, and substance abuse.

[p] We report balance at baseline on the diabetes process of care measure only for cohort one practices, because only cohort one practices are included in the analysis of impacts on this measure, given that only cohort one practices had a full year of follow-up during the primary test period. Cohort two practices joined too late to have a full year of follow-up during the primary test period.

*/**/*** Significantly different from zero at the 0.10/0.05/0.01 levels, respectively, two-tailed test. No differences were significantly different from zero at the 0.01 level.

CMS = Centers for Medicare & Medicaid Services; ED = emergency department; EHR = electronic health record; FFS = fee-for-service; HCC = Hierarchical Condition Category; MAPCP = Multi-Payer Advanced Primary Care Practice.

NA = not available.

n.a. = not applicable.

C. Equivalence of treatment and comparison groups at baseline

Demonstrating that the treatment and comparison groups are similar at the start of the intervention is important for the evaluation design. This similarity increases the credibility of a key assumption underlying difference-in-differences models—that the change over time in outcomes for the comparison group is the same change that would have happened for the treatment group, had the treatment group not received the intervention.

Table V.2 shows that the 22 treatment practices and the 91 selected comparison practices were similar at the start of the intervention on most matching variables. By construction, there were no differences between the two groups on the exact matching variables—whether a practice was an Indian Health Service practice, whether a practice accepted Medicaid, and whether a practice was located in an urban area. There were some differences between treatment group beneficiaries and matched comparison group beneficiaries on the variables we matched through propensity scores, but the standardized differences across the matching variables were almost all within our target of 0.25 standardized differences, and most were within 0.15 standardized differences (the 0.25 target is an industry standard; for example, see Institute of Education Sciences 2014).

On average, the treatment practices had somewhat more attributed Medicare FFS beneficiaries overall (1,109 versus 892) and for the high-risk participants (830 versus 657). However, in discussion with CMMI, we determined that—although these two variables fell outside our preferred standard—it was reasonable to accept the selected comparison group for two reasons. First, both sets of practices were large—the magnitude of the difference between treatment and selected comparison practice was not large; also, both were owned by a hospital/health system or a medical group, indicating that both sets of practices had resources not otherwise available to much smaller, independent practices. Second, we could account for differences in practice size through regression weights in our impact analyses.

We considered the percentage of Native American beneficiaries to be important for matching because Sanford Health selected some practices to participate in the program based on their proximity to Native American reservations. In discussion with CMMI, we determined that it was reasonable to accept the imbalance on this variable, because (1) the magnitude of the difference was small (0.7 percentage points) and (2) the percentage of Native American population was very small at both treatment and comparison practices (1.6 versus 0.9 percent, respectively).

We successfully matched on whether a practice was owned by a health system *or* a medical group, but not on ownership by a health system. We do not think that not matching exactly on ownership by a health system will introduce bias in impact estimates. The rationale is that medical groups, like health systems, can provide resources to practices that are not available to independent practices. Including potential comparisons practices owned by medical groups enabled us to improve balance on several essential variables. With regard to the imbalance on the proportion of attributed female beneficiaries, we believe that this characteristic is unlikely to bias our impact estimates because we control for gender directly using patient-level covariates.

The 15 cohort one treatment practices were also found to be similar to their selected comparisons; it is important that cohort one treatment practices are similar to selected comparisons because (1) we analyzed outcomes only for cohort one practices for the seventh through ninth intervention quarters (I7 through I9); and (2) the impact on the diabetes process of care measure was measured only for cohort one practices, because we had a full year of the primary test period only for these practices. We reported detailed balance results for cohort one in Wells et al. (2015). One difference in balance stood out; for cohort one practices and their matched comparisons, we found an imbalance in the percentage of practices located in primary care health professional shortage areas. However, because we used practice-fixed effects to capture all time-invariant practice characteristics, this controlled for the imbalance on the percentage of practice located in health professional shortage areas, given that this characteristic could change only marginally over the length of the intervention period.

The treatment and comparison practices also differed in baseline performance on one of the two quality-of-care process measures, which—as described in Section V.A.3—we did not include in the propensity-score matching algorithm because the measures were not available at the time of matching. These measures assess preventive care for those with diabetes and 14-day follow-up ambulatory care visits for those with a recent hospital discharge. For the follow-up

within 14 days of a hospital discharge, the differences between the treatment and comparison groups exceeded our thresholds (with a standardized difference of 0.60). The difference-in-differences model used to estimate impacts assumes that this difference in baseline performance would persist into the intervention period in the absence of the intervention itself.

D. Beneficiaries' outcomes and intervention impacts

In this section, we first present sample sizes and mean outcomes, by quarter, for the treatment and comparison groups. These mean outcomes provide context for understanding the difference-in-differences estimates that follow; however, the differences in mean outcomes are not regression-adjusted and not impact estimates by themselves. Next, we present the results of the primary tests, by domain. Then, we present the results of the secondary tests (robustness checks) and assess whether the primary test results are plausible given the secondary test results and the implementation evidence. We end with conclusions about program impacts in each domain.

1. Sample sizes

The sample sizes for impact estimation differ depending on the outcome. We present sample sizes by domain.

Quality-of-care processes (Table V.3)

- The **diabetes preventive care composite measure** is defined among Medicare FFS beneficiaries with diabetes ages 18 to 75 and assigned to one of 15 cohort one practices. The sample size for the treatment group and the weighted comparison group ranges from 1,659 to 1,977 across the baseline and intervention years. This population accounts for about 11 percent of the total Medicare FFS sample in the treatment and comparison groups.
- The **14-day follow-up measure** is defined among Medicare FFS beneficiaries assigned to cohort one or cohort two practices, with at least one hospital stay in the quarter. For the treatment group, the sample size ranges from 829 to 1,303 beneficiaries across the baseline and intervention quarters (accounting for nearly 7 percent of all treatment beneficiaries in each quarter). For the comparison group, the sample ranges from 818 to 1,309 across the baseline and intervention quarters (accounting for a similar proportion of the total comparison group).

Quality-of-care outcomes, service use, and spending. The sample sizes for all outcomes in these three domains are the same. In the first baseline quarter (B1), the treatment group includes 18,090 beneficiaries assigned to the 22 nonpediatric participating practices with baseline data and the comparison group includes 18,239 beneficiaries assigned to the 91 nonpediatric comparison practices (Table V.4). These analysis populations, which are limited to targeted beneficiaries, comprise about three-quarters of all Medicare FFS beneficiaries assigned to the practices. The sample sizes increased during the first two quarters of the baseline and intervention periods and decreased slowly thereafter. This means that after two or three quarters, more beneficiaries moved out of the sample (due to death, moving from the region, or switching from FFS to managed care) than were added.

Table V.3. Unadjusted mean outcomes (quality-of-care processes) observed only among select Medicare FFS beneficiaries, by treatment status and quarter

		Number of Medicare FFS beneficiaries (practices)			Mean outcomes		
Period	Quarter(s)	T	C (not weighted)	C (weighted)	T	C	Difference (%)
Among those with diabetes and ages 18 to 75 in cohort one practices and their matched comparisons, the percentage who received all four recommended diabetes processes of care in the year (%/year)							
Baseline	B1–B4[a]	1,977 (15)	5,156 (61)	1,970	47.1	43.9	3.2 (7.3%)
Intervention primary test period	I6–I9[a] (last full year of the award)	1,731 (15)	4,394 (61)	1,659	48.5	41.5	7.0 (16.9%)
Among beneficiaries with at least one inpatient admission in the quarter in cohort one and cohort two practices and their matched comparisons, the percentage of all beneficiaries whose inpatient admissions in the quarter were followed by an ambulatory care visit with a primary care or specialist provider within 14 days of discharge (%/quarter)							
Baseline	B1	1,211 (22)	3,581 (91)	1,309	56.2	58.7	-2.5 (-4.3%)
	B2	1,259 (22)	3,435 (91)	1,279	53.0	64.2	-11.3 (-17.5%)
	B2	1,211 (22)	3,517 (91)	1,261	58.5	60.5	-2.0 (-3.3%)
	B4	1,303 (22)	3,470 (91)	1,271	57.8	62.4	-4.6 (-7.3%)
Intervention	I1	1,207 (22)	3,327 (89)	1,219	61.9	63.5	-1.6 (-2.5%)
	I2	1,263 (22)	3,209 (91)	1,161	60.4	64.1	-3.7 (-5.7%)
	I3	1,141 (22)	3,101 (90)	1,130	57.8	63.0	-5.2 (-8.3%)
	I4	1,254 (22)	3,191 (90)	1,121	61.6	63.1	-1.5 (-2.4%)
	I5	1,200 (22)	3,176 (91)	1,134	58.7	64.5	-5.8 (-9.1%)
	I6	1,148 (22)	3,009 (90)	1,051	62.3	66.8	-4.5 (-6.7%)
	I7	829 (15)	2,225 (61)	818	63.7	65.7	-2.0 (-3.0%)
	I8	947 (15)	2,419 (61)	918	64.3	66.8	-2.5 (-3.7%)
	I9	891 (15)	2,219 (60)	840	62.5	67.8	-5.3 (-7.8%)

Source: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at the Centers for Medicare & Medicaid Services.

Table V.3 *(continued)*

Notes: The baseline quarters are measured relative to the start of the baseline period on April 1, 2012, for cohort one practices and January 1, 2013, for cohort two practices. For example, for cohort one practices, the first baseline quarter (B1) runs from April 1, 2012, to June 30, 2012. The intervention quarters are measured relative to the start of the intervention period on April 1, 2013. For example, the first intervention quarter (I1) runs from April 1, 2013, to June 30, 2013. In each period (baseline or intervention), the treatment group each quarter includes beneficiaries assigned to a treatment practice by the start of the quarter and who met other sample criteria—that is, they were enrolled in FFS Medicare Parts and B with Medicare as the primary payer, and lived for at least one day in one of the states with participating practices (Minnesota, North Dakota, or South Dakota) or neighboring states (Iowa or Nebraska). In addition, for the diabetes measure, we required beneficiaries to be observable for the full 12 months covered by the measure. In each period (baseline or intervention), the comparison group includes all beneficiaries assigned to a comparison practice by the start of the quarter and who met the other sample criteria. See text for details.

The outcome means were weighted such that (1) each treatment beneficiary gets a weight of 1; and (2) each comparison beneficiary gets a weight that is the product of two weights: (a) a matching weight, equal to the reciprocal of the total number of comparison practices matched to the same treatment practice as the beneficiary's assigned practice, and (b) a practice size weight, which equals the average number of beneficiaries assigned to the matched treatment panel during the four baseline quarters divided by the average number of beneficiaries assigned to the beneficiary's comparison practice over those quarters. The difference between the treatment and comparison groups in a quarter is calculated by subtracting the mean outcome for the comparison group from the mean outcome for the treatment group. The percentage difference equals that difference divided by the mean outcome for the comparison group.

[a] The quality-of-care process measures were calculated over year-long periods, corresponding to the baseline and the last four intervention quarters.

FFS = fee-for-service.

Table V.4. Unadjusted mean outcomes (quality-of-care outcomes, service use, and spending) measured for all Medicare FFS beneficiaries, by treatment status and quarter

	Number of Medicare FFS beneficiaries (practices)			Inpatient admissions for ambulatory care-sensitive conditions (#/1,000/quarter)			30-day unplanned hospital readmissions (#/1,000/quarter)			All-cause inpatient admissions (#/1,000/quarter)			Outpatient ED visit rate (#/1,000/quarter)			Medicare Part A and B spending ($/beneficiary/month)		
Q	T	C (no wgt)	C (wgt)	T	C	Diff (%)	T	C	Diff (%)	T	C	Diff (%)	T	C	Diff (%)	T	C	Diff (%)
Baseline period (April 1, 2012–March 31, 2013) for cohort one and January 1, 2013–December 31, 2014 for cohort two practices)																		
B1	18,090 (22)	50,203 (91)	18,239	15.1	16.7	-1.6 (-9.4%)	10.9	11.9	-1.1 (-8.8%)	84.6	90.2	-5.6 (-6.2%)	121.9	131.8	-9.9 (-7.5%)	$825	$869	$-44 (-5.1%)
B2	18,362 (22)	50,690 (91)	18,421	12.6	14.3	-1.7 (-11.8%)	10.7	12.0	-1.3 (-10.9%)	86.9	88.0	-1.1 (-1.2%)	133.2	132.3	0.9 (0.7%)	$825	$860	$-35 (-4.0%)
B3	18,503 (22)	50,888 (91)	18,501	14.6	15.7	-1.1 (-7.0%)	11.9	10.7	1.1 (10.7%)	83.3	87.0	-3.6 (-4.2%)	131.2	129.6	1.5 (1.2%)	$843	$874	$-31 (-3.5%)
B4	18,120 (22)	49,363 (91)	17,915	15.6	17.2	-1.6 (-9.1%)	11.5	12.1	-0.6 (-4.8%)	91.2	89.1	2.2 (2.4%)	121.4	124.7	-3.2 (-2.6%)	$885	$871	$14 (1.6%)
Intervention period (April 1, 2013–June 30, 2015 for cohort one practices and January 1, 2014–June 30, 2015 for cohort two practices)																		
I1	18,909 (22)	49,268 (91)	17,844	12.7	15.5	-2.7 (-17.6%)	7.5	10.1	-2.6 (-25.9%)	77.5	84.9	-7.5 (-8.8%)	122.4	126.1	-3.7 (-2.9%)	$824	$871	$-48 (-5.5%)
I2	19,019 (22)	49,684 (91)	17,990	13.9	12.6	1.4 (10.8%)	10.6	11.2	-0.6 (-5.4%)	84.8	82.1	2.6 (3.2%)	125.5	138.7	-13.2 (-9.5%)	$885	$854	$31 (3.6%)
I3	18,953 (22)	49,560 (91)	17,908	12.5	14.0	-1.6 (-11.2%)	9.9	10.3	-0.4 (-3.9%)	77.2	80.7	-3.5 (-4.3%)	126.8	131.2	-4.5 (-3.4%)	$872	$877	$-5 (-0.5%)
I4	18,524 (22)	48,582 (91)	17,498	15.2	15.1	0.1 (0.5%)	10.5	10.8	-0.3 (-2.9%)	86.4	81.5	5.0 (6.1%)	127.0	131.7	-4.7 (-3.6%)	$885	$850	$35 (4.1%)
I5	18,238 (22)	47,607 (91)	17,147	13.9	13.7	0.2 (1.5%)	10.6	11.3	-0.7 (-6.1%)	83.3	83.6	-0.3 (-0.4%)	135.6	136.6	-0.9 (-0.7%)	$893	$898	$-5 (-0.6%)
I6	17,990 (22)	47,035 (91)	16,909	13.2	11.2	2.0 (18.1%)	9.9	9.9	0.0 (0.1%)	80.0	78.4	1.6 (2.0%)	135.7	147.5	-11.8 (-8.0%)	$882	$911	$-28 (-3.1%)
I7	13,639 (15)	33,523 (61)	12,794	13.9	14.8	-1.0 (-6.6%)	10.5	11.3	-0.8 (-7.1%)	79.4	82.5	-3.1 (-3.8%)	127.5	136.3	-8.7 (-6.4%)	$885	$919	$-34 (-3.7%)
I8	13,189 (15)	32,463 (61)	12,358	16.2	17.4	-1.1 (-6.6%)	11.3	12.0	-0.7 (-6.2%)	89.8	93.3	-3.5 (-3.8%)	130.3	145.1	-14.8 (-10.2%)	$932	$923	$9 (1.0%)
I9	12,950 (15)	32,017 (61)	12,176	14.6	13.8	0.8 (6.0%)	11.9	12.9	-1.0 (-8.1%)	87.6	87.8	-0.2 (-0.3%)	131.1	141.2	-10.0 (-7.1%)	$962	$954	$8 (0.8%)

Source: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at the Centers for Medicare & Medicaid Services.

Notes: The baseline quarters are measured relative to the start of the baseline period on April 1, 2012. For example, for cohort one practices, the first baseline quarter (B1) runs from April 1, 2012, to June 30, 2012. The intervention quarters are measured relative to the start of the intervention period on April 1, 2013. For example, the first intervention quarter (I1) runs from April 1, 2013, to June 30, 2013. In each period (baseline or intervention), the treatment group each quarter includes all beneficiaries assigned to a treatment practice by the start of the quarter and who met other sample criteria—that is, they were enrolled in FFS Medicare Parts A and B with Medicare as primary payer, and lived for at least one day in one of the states with participating practices (Minnesota, North Dakota, or South Dakota) or neighboring states (Iowa or Nebraska). In each period, the comparison group includes all beneficiaries assigned to a comparison practice by the start of the quarter and who met the other sample criteria. See text for details.

Table V.4 *(continued)*

The outcome means were weighted such that (1) each treatment beneficiary gets a weight of 1; and (2) each comparison beneficiary gets a weight that is the product of two weights: (a) a matching weight, equal to the reciprocal of the total number of comparison practices matched to the same treatment practice as the beneficiary's assigned practice, and (b) a practice size weight, which equals the average number of beneficiaries assigned to the matched treatment practice during the four baseline quarters divided by the average number of beneficiaries assigned to the beneficiary's comparison practice over those quarters. The difference between the treatment and comparison groups in a quarter is calculated by subtracting the mean outcome for the comparison group from the mean outcome for the treatment group. The percentage difference equals that difference divided by the mean outcome for the comparison group.

B = baseline; C = comparison; Diff = difference; ED = emergency department; FFS = fee-for-service; I = intervention; no wgt = unweighted; Q = quarter; T = treatment; wgt = weighted.

2. Mean outcomes for the treatment and comparison groups, by domain and quarter

Quality-of-care processes. During the baseline year, 47.1 percent of treatment and 43.9 percent of comparison beneficiaries with diabetes ages 18 to 75 received all four recommended processes of care. This percentage increased slightly to 48.5 in the second program year for the treatment group whereas it declined to 41.5 for the comparison group (Table V.3).

For both the treatment and comparison groups, 53.0 to 64.2 percent of beneficiaries who had any hospital stay in a baseline quarter had all of those stays followed by an ambulatory care visit within 14 days of discharge. This percentage increased modestly during the intervention period, so that by the ninth intervention quarter (I9) the value was 67.8 percent for the comparison group and 62.5 for the treatment group (Table V.3).

Quality-of-care outcomes. For both treatment and comparison groups, the rates of ambulatory care-sensitive admissions declined slightly over the intervention period. The differences between the groups fluctuated, without being consistently positive or negative. The 30-day unplanned readmission rates (number per quarter) were much lower in the treatment group in I1 (7.5 versus 10.1 percent in treatment and comparison groups, respectively). In most quarters of the baseline and intervention periods, readmission rates were lower in the treatment group. Readmission rates increased in each of the last three quarters (I7 through I9). Given that these quarters include only cohort one practices, decreasing readmission rates over this period indicate that the two cohorts might have exhibited different outcome trends (Table V.4).

Service use. All-cause inpatient admissions declined for both the treatment and comparison groups from B1 through I6. For the treatment group, admissions declined again in I7, but then increased in the last two quarters. The differences between the groups in inpatient admissions fluctuated, without being consistently positive or negative.

ED visit rates were similar for the treatment and comparison group during the baseline period, but consistently higher for the comparison group compared with the treatment group during the intervention period. The number of ED visits fluctuated widely over the baseline and intervention periods, but generally trended up for both the treatment and the comparison group, increasing more for the comparison group (Table V.4).

Spending. Mean Medicare Part A and B spending increased over time for both treatment and comparison groups, with a larger increase in the last three quarters, which include only cohort one practices. There was no clear trend in the differences in mean Medicare Part A and B spending over time for the comparison group compared with the treatment group. The difference was -5.5 to +4.1 percent in all baseline and intervention quarters (Table V.4).

3. Results for primary tests, by domain

Overview. The primary tests conducted for this report cover the full primary test period for all measures. The test period covers quarters I5 through I9 for all measures, except the diabetes process of care measure, for which the test period covers quarters I6 through I9.

For two of the study domains—quality-of-care outcomes and spending—the regression-adjusted differences between the treatment and comparison groups during the primary test period were small, with one exception: the intervention was associated with a 13.6 percent increase in ambulatory care-sensitive admissions among the treatment group (Table V.5). No differences were statistically significant or larger than the substantive thresholds in either a favorable or an unfavorable direction. For the other two study domains—quality-of-care processes and service use—regression-adjusted differences between the treatment and comparison groups were statistically significant.

Quality-of-care processes. The likelihood of receiving recommended processes of care for diabetes was 8.6 percent higher for the treatment group than the estimated counterfactual, a favorable and statistically significant estimate. This was a modest improvement; however, it was smaller than the substantive threshold for this outcome of 15 percent. (Our estimated counterfactual—the outcome the treatment group members would have had in the absence of the HCIA intervention—is the treatment group mean during the intervention minus the difference-in-differences estimate.) The likelihood of receiving an ambulatory care visit within 14 days of hospital discharge was 0.1 percent higher in the treatment group than its estimated counterfactual, a favorable difference that was neither substantively large nor statistically significant. The combined estimate across both measures in the quality-of-care processes domain was 4.3 percent, a favorable point estimate that was not substantively large. The statistical power to detect substantively large effects was good (more than 99 percent) for both quality-of-care process measures individually and, in addition, combined across the measures.

Quality-of-care outcomes. The rate of ACSC admissions for the treatment group during the primary test period was 13.6 percent higher than our estimate of the counterfactual, a large (but not substantively large) unfavorable estimate. The rate of 30-day unplanned readmissions was 1.3 percent lower than our estimate of the counterfactual, a favorable, but not substantively large impact. After combining results across the two outcomes in this domain, the combined effect was 6.2 percent, smaller than the substantive threshold of 15 percent, but in the unfavorable direction. We cannot conclude whether this unfavorable result is statistically significant because the one-sided statistical tests we used tested only for improvements in outcomes.

The statistical power to detect effects the size of the substantive threshold was good for ACSC admissions (77.1 percent) and marginal for 30-day unplanned readmissions (64.1 percent). However, power was also good (80.7 percent) for the combined effect in the domain.

Service use. The treatment group's admission rate was 1.8 percent higher and not statistically significant or substantively large, and the outpatient ED visit rate was 4.9 percent lower and statistically significant. This modest improvement in ED visits was smaller than the substantive threshold of 15 percent. After combining results across the two outcomes in this domain, the outcomes for the treatment group were slightly better (1.6 percent lower) than the estimated counterfactual. Power to detect effects that were the size of the substantive thresholds was good (more than 99 percent) for both service use measures individually and, in addition, combined across the measures.

Table V.5. Results of primary tests for Sanford Health

Primary test definition					Statistical power to detect an effect that is[a]		Results			
Domain (# of tests in domain)	Outcome (units)	Time period for impacts	Population	Substantive threshold (impact as a percentage relative to the counterfactual[b])	Size of the substantive threshold	Twice the substantive threshold[c]	Treatment group mean	Regression-adjusted difference between the treatment and estimated counterfactual[b] (standard error)	Percentage difference[d]	*p*-value[e]
Quality-of-care process (2)	Received all four recommended diabetes processes of care in the year (binary [yes or no]/beneficiary/year)	Last year of intervention (corresponding to quarters 6 through 9)[f]	Medicare FFS beneficiaries assigned to cohort one treatment practices with diabetes and ages 18 to 75	+15.0%	98.6%	> 99.9%	48.5	3.8 (1.9)	8.6%	0.05
	All inpatient admissions within a quarter were followed by an ambulatory care visit with a primary care or specialist provider within 14 days (binary [yes or no]/beneficiary/year)	Average over intervention quarters 5 through 9 for cohort one practices; average over intervention quarters 5 through 6 for cohort two practices	Medicare FFS beneficiaries with a targeted condition assigned to treatment practices and who had at least on hospital stay in the quarter	+15.0%	> 99.9%	> 99.9%	62.3	0.0 (1.2)	0.1%	0.50
	Combined	Varies by test	Varies by test	+15.0%	> 99.9%	> 99.9%	n.a.	n.a.	4.3%	0.04
Quality-of-care outcomes (2)	Inpatient admissions for ambulatory care-sensitive conditions (#/1,000 beneficiaries/quarter)	Average over intervention quarters 5 through 9 for cohort one practices; average over intervention quarters 5 through 6 for cohort two practices	Medicare FFS beneficiaries with a targeted condition assigned to treatment practices	-15.0%	77.1%	99.7%	14.4	1.7 (0.9)	13.6%	0.94

Table V.5 *(continued)*

Primary test definition					Statistical power to detect an effect that is[a]		Results			
Domain (# of tests in domain)	Outcome (units)	Time period for impacts	Population	Substantive threshold (impact as a percentage relative to the counterfactual[b])	Size of the substantive threshold	Twice the substantive threshold[c]	Treatment group mean	Regression-adjusted difference between the treatment and estimated counterfactual[b] (standard error)	Percentage difference[d]	*p*-value[e]
	30-day unplanned readmissions (#/1,000 beneficiaries/quarter)	Average over intervention quarters 5 through 9 for cohort one practices; average over intervention quarters 5 through 6 for cohort two practices	Medicare FFS beneficiaries with a targeted condition assigned to treatment practices	-15.0%	64.1%	97.7%	10.8	-0.1 (1.0)	-1.3%	0.49
	Combined (%)	Average over intervention quarters 5 through 9 for cohort one practices; average over intervention quarters 5 through 6 for cohort two practices	Medicare FFS beneficiaries with a targeted condition assigned to treatment practices	-15.0%	80.7%	99.9%	n.a.	n.a.	6.2%	0.81
Service use (2)	All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	Average over intervention quarters 5 through 9 for cohort one practices; average over intervention quarters 5 through 6 for cohort two practices	Medicare FFS beneficiaries with a targeted condition assigned to treatment practices	-15.0%	> 99.9%	> 99.9%	84.0	1.5 (2.6)	1.8%	0.59
	Outpatient ED visits (#/1,000 beneficiaries/quarter)	Average over intervention quarters 5 through 9 for cohort one practices; average over intervention quarters 5 through 6 for cohort two practices	Medicare FFS beneficiaries with a targeted condition assigned to treatment practices	-15.0%	> 99.9%	> 99.9%	132.1	-6.8 (3.7)	-4.9%	0.06

Table V.5 *(continued)*

Primary test definition					Statistical power to detect an effect that is[a]		Results			
Domain (# of tests in domain)	Outcome (units)	Time period for impacts	Population	Substantive threshold (impact as a percentage relative to the counterfactual[b])	Size of the substantive threshold	Twice the substantive threshold[c]	Treatment group mean	Regression-adjusted difference between the treatment and estimated counterfactual[b] (standard error)	Percentage difference[d]	*p*-value[e]
	Combined (%)	Average over intervention quarters 5 through 9 for cohort one practices; average over intervention quarters 5 through 6 for cohort two practices	Medicare FFS beneficiaries with a targeted condition assigned to treatment practices	-15.0%	> 99.9%	> 99.9%	n.a.	n.a.	-1.6%	0.25
Spending (1)	Medicare Part A and B spending ($/beneficiary/month)	Average over intervention quarters 5 through 9 for cohort one practices; average over intervention quarters 5 through 6 for cohort two practices	Medicare FFS beneficiaries with a targeted condition assigned to treatment practices	-2.25%	37.4%	73.8%	$911	$13 (21)	1.5%	0.73

Source: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at the Centers for Medicare & Medicaid Services.

Note: The results for each outcome are based on a difference-in-differences regression model, as described in the text. Estimates are calculated for Medicare beneficiaries who are observable in the relevant time period: that is, beneficiaries who are enrolled in Medicare FFS (Part A and B), are alive, and have Medicare as their primary payer. Additional sample restrictions apply to the quality-of-care process measures; see text for details.

[a] The power calculation is based on actual standard errors from the analysis. For example, in the last row, a 2.25 percent effect on Medicare Part A and B spending (from the counterfactual of $911 – $13 = $898) would be a change of $20. Given the standard error of $21.00 from the regression model, we would be able to detect a statistically significant result 37.4 percent of the time if the impact was truly $20, assuming a one-sided statistical test at the $p = 0.10$ significance level.

[b] The substantive threshold is the impact as a percentage of the counterfactual. The counterfactual is the outcome the treatment group would have had in the absence of the HCIA-funded intervention. Our estimate of the counterfactual is the treatment group mean minus the regression-adjusted difference-in-differences estimate.

[c] We show statistical power to detect a very large effect (twice the size of the substantive threshold) because this provides additional information about the likelihood that we will find effects if the program is indeed effective. If the power to detect effects is less than 75 percent even for a very large effect, then the evaluation is extremely poorly powered for that outcome.

Table V.5 *(continued)*

[d] Percentage difference is calculated as the regression-adjusted difference between the treatment and comparison group, divided by the adjusted comparison group mean.

[e] *p*-values test the null hypothesis that the regression-adjusted difference-in-differences estimate is less than or equal to zero for outcomes in the quality-of-care processes domain, or greater than or equal to zero in all other domains (a one-sided test). Because it is a one-sided test, as the difference-in-differences estimate approaches infinity in an unfavorable direction (negative for process of care measures and positive for all other measures), the *p*-value approaches 1, whereas it would approach 0 in a two-sided test. We adjusted the *p*-values for the multiple (three) comparisons made within the quality-of-care processes domain, and (separately) for the two comparisons made within the quality-of-care outcomes domain, and for the two comparisons made within the service use domain.

[f] We estimated impacts as the average across intervention quarters 5 through 9 for all outcomes except for the quality-of-care process measure for diabetes. For that measure, we calculated outcomes instead over a year-long period (rather than quarters)—specifically, over the last four quarters of the intervention. The impact estimates apply to the same time period—that is, the year that corresponds to intervention quarters 5 through 9—but the estimate is not an average of quarterly estimates.

ED = emergency department; FFS = fee-for-service; HCIA = Health Care Innovation Award.

Spending. The treatment group averaged $911 per beneficiary per month in Part A and B spending during the fifth through ninth intervention quarters, a value 1.5 percent (or $13) higher than the estimated counterfactual. This difference was smaller than the substantive threshold of 2.25 percent. Statistical power to detect an effect the size of the substantive threshold was poor (37.4 percent).

Aggregate estimates for CMMI's core measures. The estimates presented for the CMMI core outcomes—that is, for 30-day unplanned readmissions, all-cause inpatient admissions, the outpatient ED visit rate, and Medicare Part A and B spending—have so far been expressed per 1,000 beneficiaries per quarter (or, for spending, per beneficiary per month). Table V.6 translates these rates or per-person-month estimates into estimates of aggregate impacts during the five-quarter-long primary test period presented in this report. We calculated these aggregate impacts by multiplying the point estimates by the average number of targeted Medicare beneficiaries in the treatment group and by the number of quarters or months during the primary test year. Although the point estimates are small for most of these measures, the aggregate estimates are large because they are scaled to the entire targeted Medicare population of roughly 15,000 beneficiaries per quarter and to the five quarters of the primary test period. For example, the results in Table V.5 show that the intervention was associated with a decrease in ED visits of 6.8 per 1,000 beneficiaries per quarter, or 4.9 percent relative to the estimated counterfactual. Across roughly 15,000 beneficiaries per quarter and five quarters of the primary test period, this translates into an aggregate reduction of 519 ED visits. The intervention was associated with an increase in Medicare Part A and B spending of only $13 per beneficiary per month, or 1.5 percent relative to the estimated counterfactual. This small increase per person per month translates into an aggregate cost of the program of nearly $3 million. The large point estimate for spending should be interpreted with caution because the estimate is not statistically significant (the *p*-values for aggregate estimates are the same as they are for the main results shown in Table V.5).

Table V.6. Results for primary tests for CMMI's core outcomes expressed as aggregate effects for all Medicare FFS beneficiaries in the treatment group

Outcome (units)	Aggregate impact estimate during the primary test period (I5–I9)	*p*-value
30-day unplanned readmissions (#)	-11	0.49
All-cause inpatient admissions (#)	+110	0.59
Outpatient ED visits (#)	-519	0.06
Medicare Part A and B spending ($)	+$2,979,556	0.73

Source: Authors' calculation, based on analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at the Centers for Medicare & Medicaid Services.

Note: To estimate the aggregate impact during the primary test year (intervention quarters 5 through 9) we (1) multiplied the per beneficiary per quarter estimate from Table V.5 by the average number of Medicare FFS beneficiaries in the treatment group during the four primary test quarters then (2) scaled the estimate to the primary test period by multiplying the resulting product by five (the number of quarters in the primary test period). The *p*-values are taken from Table V.5 and are therefore one-sided (testing that the program improved outcomes) and adjusted for multiple comparisons conducted within each outcome domain.

CMMI = Center for Medicare & Medicaid Innovation; ED = emergency department; FFS = fee-for-service.

4. Results for secondary tests

Estimates during the first intervention year (for cohort one April 1, 2013, to March 31, 2013; for cohort two January 1, 2014, to December 31, 2014). As shown in Table V.7, the differences in inpatient admissions and ACSC admissions for the treatment group and its estimated counterfactual were small (3.9 and 0.6 percent, respectively) and not statistically significant during the first year of the intervention (I1 through I4). The difference in readmissions between the treatment group and its counterfactual were larger during the first year of the intervention versus during the fifth through ninth quarters (6.6 versus 1.3 percent, respectively). However, the estimate during the first year of the intervention was not statistically significant or substantively large. Results indicate statistically significantly lower outpatient ED visit rates among the treatment group in the first intervention year of 4.2 percent, slightly smaller than the 4.9 percent decrease seen during the primary test period. Therefore, the secondary test results generally support the primary test results for ED visits by showing smaller impacts in the first program year, as the awardee expected.

Further, we found a substantively large and unfavorable, but statistically insignificant difference in Medicare Part A and B spending of 2.7 percent. Despite being larger than the substantive threshold of 2.25 percent, the difference was sufficiently small in magnitude so that it did not raise concerns about the credibility of the comparison group. For all other outcomes, we did not see substantively large differences during the first year of practice participation, a period during which we and the Sanford Health did not expect large program effects. These results increase our confidence in the comparison group that, in turn, gives us greater confidence in the primary test results and, eventually, the conclusions of the impact evaluation.

5. Consistency of impact estimates with implementation findings

The impact estimates in the primary tests are plausible given the implementation findings. The primary test results showed statistically significant improvements during the fifth through ninth quarters of the program on the receipt of recommended tests for diabetes and outpatient ED visits. The implementation evidence shows the program was active during this period. A statistically significant improvement in the diabetes process-of-care measure is consistent with the fact that practices cited diabetes as one of the first conditions on which they focused for building registries and panel management. Sanford Health increased identification of depression and anxiety in the beginning of 2014, surpassing target identification rates Sanford Health set based on national prevalence rates. This period corresponds to the first quarter of the primary test period for the 15 cohort one practices and start of the intervention period for the 7 cohort two practices. According to Sanford Health's theory of action, earlier identification of depression and anxiety results in earlier intervention, which is expected to result in fewer outpatient ED visits, whether for behavioral health conditions or for other conditions.

Essentially no change in the 14-day follow-up after inpatient admissions and a very small reduction in 30-day unplanned readmissions are consistent with the fact that the HCIA program did not emphasize post-discharge care management as part of the award activities.

Table V.7. Results of secondary tests for Sanford Health

Secondary test definition				Results			
Domain	Outcome (units)	Time period for impacts	Population	Treatment group mean	Regression-adjusted difference between treatment and the estimated counterfactual (standard error)	Percentage difference[a]	*p*-value[b]
Quality of care	Inpatient admissions for ambulatory care-sensitive conditions (#/1,000 beneficiaries/quarter)	Average over intervention quarters 1–4	Medicare FFS beneficiaries with a targeted condition assigned to treatment practices	13.6	0.5 (0.9)	3.9%	0.72
	30-day unplanned readmissions (#/1,000 beneficiaries/quarter)	Average over intervention quarters 1–4	Medicare FFS beneficiaries with a targeted condition assigned to treatment practices	9.6	-0.7 (0.9)	-6.6%	0.23
Service use	All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	Average over intervention quarters 1–4	Medicare FFS beneficiaries with a targeted condition assigned to treatment practices	81.5	0.5 (2.4)	0.6%	0.59
	Outpatient ED visits (#/1,000 beneficiaries/quarter)	Average over intervention quarters 1–4	Medicare FFS beneficiaries with a targeted condition assigned to treatment practices	125.4	-5.5 (3.4)	-4.2%	0.05
Spending	Medicare Part A and B spending ($/beneficiary/month)	Average over intervention quarters 1–4	Medicare FFS beneficiaries with a targeted condition assigned to treatment practices	$866	$23 ($19)	2.7%	0.88

Source: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at the Centers for Medicare & Medicaid Services.

Note: The results for each outcome are based on a difference-in-differences regression model, as described in the text. Estimates are calculated for Medicare beneficiaries who are observable in the relevant time period: that is, beneficiaries who are enrolled in Medicare FFS (Part A and B), are alive, and have Medicare as their primary payer. We defined targeted beneficiaries as those with one or more of eight chronic health conditions targeted by the One Care program: anxiety, asthma, diabetes, depression, heart failure, hypertension, obesity, and substance abuse. We defined the targeted beneficiaries among all treatment group members at the beginning of the baseline period (for outcomes in the baseline period) or intervention period (for outcomes in the intervention period).

[a] Percentage difference is calculated as the regression-adjusted difference between the treatment and comparison groups, divided by the adjusted comparison group mean.

[b] The *p*-values from the secondary test results were *not* adjusted for multiple comparisons within each domain or across domains.

ED = emergency department; FFS = fee-for-service.

6. Conclusions about program impacts, by domain

Based on all evidence currently available, we have drawn the following conclusions about program impacts during the five-quarter-long primary test period for cohort one practice and two-quarter-long primary test period for cohort two practices. Table V.8 summarizes these conclusions and their support.

- **The program had a statistically significant, modest, favorable improvement on quality-of-care process and service use measures.** The impact on the diabetes measure drove the favorable improvement for quality-of-care process measures. The likelihood of receiving recommended processes of care for diabetes was 8.6 percent higher for the treatment group than the counterfactual, a favorable and statistically significant estimate. In contrast, there was no effect on the 14-day follow-up measure; not only was that estimate not substantively large or statistically significant, but the point estimate was close to zero. The estimate for outpatient ED visits drove the favorable effect for service use. The treatment group's outpatient ED visit rate was 4.9 percent lower and statistically significant. In contrast, the estimate for inpatient admissions was neither substantively large nor statistically significant. The secondary test results support the primary test results for ED visits by showing smaller impacts in the first program year, as the awardee expected. These conclusions are also consistent with implementation findings, as described earlier.
- **The program had no substantively large effects on quality-of-care outcomes**. The estimates for the two outcomes in this domain—ACSC admissions and 30-day unplanned readmissions—were neither substantively large nor statistically significant. The statistical power to detect substantively large effects was good for ACSC admissions (77.1 percent) and for the two measures combined (80.7 percent), so it is unlikely that the program had substantively large effects that our tests failed to detect.
- **The program had an indeterminate effect on Medicare spending.** The primary test results were neither substantively large nor statistically significant. However, the statistical power was poor (37.4 percent) to detect effects the size of the substantive threshold, largely because the threshold was so small (2.25 percent). As a result, null findings from the primary test in this domain could be due to (1) the program truly not having a substantively large effect or (2) the program having a substantively large effect but our tests failing to detect it. Although statistically insignificant, the point estimate on inpatient admissions and the ACSC admissions was positive (indicating an increase).

Table V.8. Preliminary conclusions about the impacts of Sanford Health's HCIA program on patients' outcomes, by domain

		Evidence supporting conclusion		
Domain	Preliminary conclusion	Primary test result(s) that supported conclusion	Primary test result(s) plausible given secondary tests?	Primary test result(s) plausible given implementation evidence?
Quality-of-care process	Statistically significant favorable effect	• Statistically significant effect on the diabetes process-of-care measure	Yes	Yes
Quality-of-care outcomes	No substantively large effects	• No substantively large or statistically significant effects; well powered to detect effects in the domain overall (well powered for ACSC outcomes but not readmissions)	Yes	Yes
Service use	Statistically significant favorable effect	• Statistically significant effect on ED visits	Yes	Yes
Spending	Indeterminate	• No substantively large or statistically significant effects; poorly powered	Yes	Yes

Sources: Tables V.5 and V.7.

ACSC = ambulatory care-sensitive condition; ED = emergency department.

VI. DISCUSSION AND CONCLUSIONS

Sanford Health used its $12.1 million in HCIA funds to implement One Care, a medical home model that had three components: (1) integration of behavioral health services into primary care, (2) provision of care management services, and (3) expansion of health IT. To support these components, Sanford Health developed and trained care teams—including BHTTs and RN health coaches—to provide integrated, patient-centered care. Sanford Health also equipped care teams with new tools (such as screening instruments and clinical guidelines) and used health IT to support all intervention components. By the end of the award, Sanford Health aimed to reduce potentially preventable admission rates and ED visit rates by 20 percent and total cost of care by 3 percent for Medicare, Medicaid, and CHIP patients with targeted conditions. Sanford Health also expected to improve quality-of-care process outcomes such as optimal care for asthma, diabetes, and hypertension, but did not set specific targets.

The results from our impact evaluation suggest Sanford Health met some of these goals during the three-year award. Compared with beneficiaries in the matched comparison practices, the program had favorable and statistically significant effects on quality-of-care processes (driven by favorable estimates for the receipt of recommended diabetes processes of care) and service use (driven by favorable estimates for ED visits). The program did not have substantively large effects on quality-of-care outcomes and there is no evidence that the program reduced Medicare spending. The evaluation was well powered to detect substantively large impacts on

quality-of-care processes, service use, and quality-of-care outcomes, but not on Medicare spending.

The lack of improvements in quality-of-care outcomes and spending domains does not appear to be due to a failure to implement the intervention as planned. Despite some delays implementing the BH-6 screening, remote monitoring, and conducting outreach to the Native American population in Bemidji, Sanford Health successfully integrated behavioral health and care management services into primary care practices consistent with its core design. Sanford Health's training program and staff engagement efforts facilitated team cohesion and buy-in. The newly developed tools for care teams (such as clinical practice guidelines, electronic screenings, and disease registries for the targeted conditions) facilitated care management and behavioral health integration. However, we do not know the extent to which clinicians used the newly developed guidelines.

Several measures capture the generally successful implementation:

- Sanford Health engaged 290 staff in care teams and workforce development in the 33 participating practices, nearly meeting its goal of 325 staff.
- Sanford Health successfully increased the identification of patients with depression and anxiety, exceeding targets based on national prevalence rates. Early identification of these conditions led to increased referrals of high-acuity patients to specialists. Improvements in care for these beneficiaries might have contributed to the observed reduction in ED visits.
- More than 60 percent of clinicians at 22 nonpediatric practices who were familiar with the award believed that the intervention positively affected quality and patient-centeredness of care. About sixty percent reported that the award improved their ability to respond to patients' needs in a timely way. Clinicians attributed this improvement to the integration of new care team members, especially RN health coaches and BHTTs.

The lack of improvements in quality-of-care outcomes and spending domains might be due to any one of four factors. First, the intervention might not have been sufficiently intensive to generate substantively large effects on these outcomes. For example, it is possible that health coaches needed to meet more frequently with patients and to follow them for a longer period. We do not have data on the intensity of the intervention and, therefore, we cannot investigate this possibility further. Second, it is possible that the content of the intervention was not amenable to a reduction in some of the analyzed outcomes. We found statistically significant effects on outpatient ED visits, but not on inpatient stays, which usually entail higher acuity needs and thus are more difficult to prevent. Third, because some practices first built asthma, diabetes, and hypertension disease registries and panel management protocols before moving on to heart failure and obesity, they might have had fewer new tools to help improve outcomes for these two conditions.

Finally, it is possible that impacts take longer to accrue than we could observe in this study and would have grown larger if the program had continued. At the start of the primary test period for most practices included in the quantitative evaluation (15 cohort one practices), clinicians'

responses about the effect of the program on patient-centeredness, quality, and timeliness of care were divided. About half of the respondents said that the award had a positive effect on these aspects of care, whereas half said that the program did not have a positive effect or that it was too soon to tell. Even though the positive responses increased to 60 percent or more by the end of the award, it is possible that earlier improvements in care were required for the impacts on outcomes to accrue over the primary test period.

Even though Sanford Health clinicians and other staff acknowledged that patients' progress occurs slowly, the evaluation followed patients' outcomes for a long time: the evaluation covered five quarters of the primary test period for 15 cohort one practices and two quarters for 7 cohort two practices. Moreover, we started the primary test period in the fifth quarter of the intervention, already accounting for expectations that impacts would take time to accrue. If impacts take even longer to accrue, they would have to be very large during the additional months to generate favorable results over the entire primary test period for quality-of-care outcomes and spending, especially because current results show increases in several outcomes (inpatient admissions, ACSC admissions, and spending). Overall, these factors did not prevent Sanford Health from demonstrating statistically significant, modest improvements in the quality-of-care process and service use domains.

This page has been left blank for double-sided copying.

REFERENCES

Boards of Trustees, Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds. "2013 Annual Report of the Boards of Trustees of the Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds." Table V.D1. Washington, DC: Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds, 2013. Available at http://downloads.cms.gov/files/TR2013.pdf. Accessed August 13, 2014.

Centers for Medicare & Medicaid Services. "CSV Flat Files—Revised: Readmissions Complications and Deaths—National.csv." Baltimore, MD: CMS, 2014. Available at https://data.medicare.gov/data/hospital-compare. Accessed August 14, 2014.

Chronic Conditions Data Warehouse. "Table A.1.a Medicare Beneficiary Counts for 2005–2014." Baltimore, MD: Centers for Medicare & Medicaid Services, 2016a. Available at https://www.ccwdata.org/web/guest/medicare-tables-reports. Accessed June 29, 2016.

Chronic Conditions Data Warehouse. "Condition Categories." Baltimore, MD: Centers for Medicare & Medicaid Services, 2016b. Available at https://www.ccwdata.org/web/guest/condition-categories. Accessed August 5, 2016.

Gerhardt, Geoffrey, Alshadye Yemane, Keri Apostle, Allison Oelschlaeger, Eric Rollins, and Niall Brennan. "Evaluating Whether Changes in Utilization of Hospital Outpatient Services Contributed to Lower Medicare Readmission Rate." *Medicare & Medicaid Research Review,* vol. 4, no. 1, 2014, pp. E1–E13.

Health Indicators Warehouse. "Average Age of Medicare Beneficiaries (mean)." Hyattsville, MD: National Center for Health Statistics, HIW, 2014a. Available at http://www.healthindicators.gov/Indicators/Average-age-of-Medicare-beneficiaries-mean_308/Profile/ClassicData. Accessed November 19, 2014.

Health Indicators Warehouse. "Hospital Inpatient Medicare Admissions (per 1,000 beneficiaries)." Hyattsville, MD: National Center for Health Statistics, HIW, 2014b. Available at http://www.healthindicators.gov/Indicators/Hospital-inpatient-Medicare-admissions-per-1000-beneficiaries_2001/Profile/ClassicData. Accessed August 13, 2014.

Health Indicators Warehouse. "Medicare Beneficiaries Eligible for Medicaid (percent)." Hyattsville, MD: National Center for Health Statistics, HIW, 2014c. Available at http://www.healthindicators.gov/Indicators/Medicare-beneficiaries-eligible-for-Medicaid-percent_317/Profile/ClassicData. Accessed August 4, 2015.

Institute of Education Sciences. "What Works Clearinghouse: Procedures and Standards Handbook, Version 3.0." Washington, DC: U.S. Department of Education, IES, 2014. Available at http://ies.ed.gov/ncee/wwc/DocumentSum.aspx?sid=19. Accessed September 15, 2014.

Peikes, Deborah, Stacy Dale, Eric Lundquist, Janice Genevro, and David Myers. "Building the Evidence Base for the Medical Home: What Sample and Sample Size Do Studies Need? White Paper." AHRQ Publication No.11-0100-EF. Rockville, MD: Agency for Healthcare Research and Quality, October 2011.

Stuart, Elizabeth A. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science,* vol. 25, no. 1, 2010, pp. 1–21.

Truven Health Analytics. "AHRQ Quality Indicators, Prevention Quality Indicators v5.0 Benchmark Data Tables." Prepared for the Agency for Healthcare Research and Quality, U.S. Department of Health and Human Services. Santa Barbara, CA: Truven Health Analytics, March 2015. Available at http://www.qualityindicators.ahrq.gov/Downloads/Modules/PQI/V50/Version_50_Benchmark_Tables_PQI.pdf. Accessed August 18, 2015.

Wells, KeriAnn, Jelena Zurovac, Catherine DesRoches, Boyd Gilman, Greg Peterson, Sandi Nelson, Laura Blue, Keith Kranker, Kate Stewart, Frank Yoon, and Lorenzo Moreno. "Findings for Sanford Health One Care." In Moreno, Lorenzo, Boyd Gilman, Greg Peterson, Catherine DesRoches, Sheila Hoag, Linda Barterian, Laura Blue, Katherine Bradley, Emily Ehrlich, Kristin Geonnotti, Lauren Hula, Keith Kranker, Rumin Sarwar, KeriAnn Wells, Joseph Zickafoose, Sandi Nelson, Kate Stewart, and Frank Yoon. "Evaluation of Health Care Innovation Awards (HCIA): Primary Care Redesign Programs." Second annual report to CMS. Volume II: Individual program summaries. Princeton, NJ: Mathematica Policy Research, December 11, 2015.

CHAPTER 9

TransforMED

Sean Orzol, Rosalind Keith, Mynti Hossain, Michael Barna, Greg Peterson, Boyd Gilman, Laura Blue, Keith Kranker, Kate Stewart, Sheila Hoag, and Lorenzo Moreno

This page has been left blank for double-sided copying.

TransforMED

CHAPTER SUMMARY

Introduction. TransforMED received a $20.8 million Health Care Innovation Award (HCIA) to implement a patient-centered medical neighborhood (PCMN) program. The PCMN intervention was focused almost exclusively on implementing health information technology (IT), with TransforMED providing software and technical assistance to 90 primary care practices recruited by 15 participating health systems in 15 states. TransforMED aimed to reduce the cost of health care for Medicaid and Medicare fee-for-service (FFS) patients by 4 percent (or $49.5 million) by the end of the award. The organization planned to achieve this aim by reducing patients' need for acute care—such as inpatient admissions and emergency department (ED) visits—and improving coordination of care among providers within the PCMN community.

Objectives. This report (1) describes the design and implementation of TransforMED's HCIA-funded program, including the role of clinicians in the program and the extent to which anticipated changes in clinician behavior occurred; (2) estimates impacts of the program on patients' outcomes and Medicare Part A and B spending during the three years of the award; and (3) uses both implementation and impact findings to identify possible explanations for the observed impacts.

Methods. We reviewed TransforMED's program documents and self-monitoring metrics, conducted interviews with TransforMED leadership and program staff, and surveyed participating clinicians and practice staff. To estimate impacts, we compared outcomes for Medicare FFS patients served by 87 of the 90 treatment practices (those for which we identified Medicare FFS beneficiaries in each program quarter) to outcomes for Medicare FFS beneficiaries served by 286 matched comparison practices, located in the same regions as the PCMN practices, adjusting for any differences in outcomes between the two groups during a one-year baseline period. We estimated impacts in three evaluation domains: quality-of-care processes, service use, and spending.

Program design and implementation. The program had two components: (1) providing population health management and cost-reporting software to practices so they could more effectively use data to improve clinical processes (for example, by monitoring utilization and spending across their patients, identifying patients who would benefit from preventive services, and sending automated emails encouraging patients to schedule recommended follow-up visits); and (2) technical assistance to practices and participating health systems on how to use the software, and to promote practice transformation.

Despite not implementing some aspects of the program to the extent anticipated (such as the cost-reporting functions), the program was implemented largely as planned. TransforMED successfully recruited 15 health systems and 90 practices without any major delays. In addition to implementing these two planned components, TransforMED trained all practices on using quality improvement processes (in the form of plan-do-study-act cycles) to make quality-of-care process improvements that aligned with patient-centered medical home (PCMH) concepts. Practice staff surveyed about the training they received for the PCMN program generally

reported that the training had improved their ability to provide care in a way that aligned with PCMH concepts.

Clinicians' perceptions of program effects on the care they provided to participants. For most of TransforMED's program goals, the program design did not require clinicians to change the way they provided care. However, TransforMED provided participating practices with a tool to compare physicians' efficiency (based on cost and utilization metrics) against a national peer group and to produce efficiency scores for physicians by certain diagnoses and procedures. This indicates that the program intended to change clinicians' behaviors related to referrals to specialists. The available evidence suggests that TransforMED did not engage clinicians as planned; the quarterly narratives TransforMED submitted to the Centers for Medicare & Medicaid Services (CMS) suggest that clinicians' engagement was a challenge throughout the award. Further, fewer than half of the clinicians surveyed reported that they thought the program improved the quality, timeliness, safety, patient-centeredness, and equity of care, and the availability of information for decision making.

Impacts on patients' outcomes. The program had statistically significant favorable impacts on service use during the last year of the award. Specifically, we estimated the program decreased our composite service use measure by 5.5 percent ($p = 0.03$, after adjusting for multiple tests in this evaluation domain). The decrease in service use captures a 7.1 percent reduction in the inpatient admission rate and a 5.7 percent decrease in the outpatient ED visit rate among the Medicare FFS population. Secondary tests (robustness checks) confirmed the plausibility of the findings on service use. However, the impact estimates suggest that the intervention did not improve patients' outcomes in the two other evaluation domains; there was no evidence of statistically significant or substantively large effects in either the quality-of-care processes or spending domains.

Conclusion. These results indicate that TransforMED's HCIA-funded program reduced service use for Medicare FFS beneficiaries, but also show that the intervention did not have statistically significant favorable impacts in the quality-of-care processes or spending domains. These results suggest that providing practices with population health management and cost-reporting software—along with technical assistance for how to use them—can complement practices' own PCMH transformation efforts and add meaningfully to their impacts on service use. These favorable findings likely would not replicate, however, in settings where providers lack incentives to use the IT systems and technical assistance in the same way that providers do when participating in broader PCMH efforts to transform primary care delivery.

Summary of intervention and impact results for TransforMED

Intervention description			
Awardee description		National learning and dissemination contractor (subsidiary of the American Academy of Family Physicians)	
Award amount ($ millions)		$20.8 million	
Award extended beyond June 2015?		No	
Locations		Multistate (urban, suburban, and rural)	
Target population		All patients served by 90 primary care practices that were part of 15 health systems	
Interventions		Health IT to help practices function as part of a patient-centered medical neighborhood • Software for managing health of patient panel and identifying cost drivers • Technical assistance (learning collaboratives and monthly calls) to use new health IT	
Metrics of intervention delivered		• 78 of 90 practices implemented population health management software • 96% of practices identified a health coach to serve as an expert for population management in each practice.	
Impact evaluation methods			
Core design		Difference-in-differences model with matched comparison group	
Treatment group	Definition	Medicare FFS beneficiaries attributed to 87 participating practices	
	# of beneficiaries during primary test period[a]	93,213 to 97,994, depending on the quarter	
Comparison group definition		Medicare FFS beneficiaries attributed to 286 matched comparison practices	
Impact results: Quality-of-care processes domain			
Ambulatory care visit within 14 days of discharge (% of beneficiaries/quarter)		Comparison mean[b]	61.2%
		Impact estimate (% difference)	+0.8 pp (+1.3%)
Received recommended lipid test, for patients with IVD (% of beneficiaries/year)		Comparison mean[b]	75.1%
		Impact estimate (% difference)	+1.4 pp (+1.9%)
Received all four recommended diabetes processes of care (% of beneficiaries/year)		Comparison mean[b]	44.6%
		Impact estimate (% difference)	+0.5 pp (+1.2%)
Combined impact estimate[c]		+1.5%	
Impact conclusion[d]		No substantively large effect	
Impact results: Service use domain			
All-cause inpatient admissions (#/1,000 beneficiaries/quarter)		Comparison mean[b]	82.6
		Impact estimate (% difference)	-5.8 (-7.1)**[e]
Outpatient ED visits (#/1,000 beneficiaries/quarter)		Comparison mean[b]	144.7
		Impact estimate (% difference)	-8.2 (-5.7)**[e]
Combined impact estimate[c]		-5.5**[f]	
Impact conclusion[d]		**Statistically significant favorable effect**	
Impact results: Spending domain			
Medicare Part A and B spending ($/beneficiary/month)		Comparison mean[b]	$910
		Impact estimate (% difference)	-$10 (-1.1%)[e]
Combined impact estimate[c]		0.40%[g]	
Impact conclusion[d]		No substantively large effect	

Note: See the TransforMED chapter for details on the intervention, impact methods, and impact results.

[a] Number of beneficiaries in the full treatment group across the quarters in the primary test period.

[b] The comparison mean is the estimate of the outcome the treatment group beneficiaries would have had if they had not received the intervention. It is equal to the mean for the treatment group over the intervention quarters (in the primary test period) minus the impact estimate.

[c] The combined estimate is the average across all the individual estimates in the domain, in which the impact estimate for each individual outcome is expressed as percentage change relative to the comparison group.

[d] We drew conclusions at the domain level based on the results of prespecified primary tests, secondary tests (robustness checks), and consistency with implementation evidence. For each domain, we could draw one of five conclusions: (1) Statistically significant favorable effect (the highest level of evidence), (2) Substantively important (but not statistically significant) favorable effect, (3) Substantively important (but not statistically significant) unfavorable effect, (4), No substantively large effect, and (5) Indeterminate effect. Section V.A.8 of this report describes the decision rules we used to reach each of these possible conclusions.

Summary of intervention and impact results for TransforMED *(continued)*

[e] We also conducted a primary test of all-cause inpatient admissions, outpatient ED visits, and spending for a high-risk subset of the full sample. The results of those tests, which supported the result for the full population shown in this table, are reported in the chapter.

[f] The combined impact estimate for the service use domain comprises the estimates of all-cause inpatient admissions and outpatient ED visits for the entire sample (reported above) and for the high-risk sample (reported in the chapter only).

[g] The combined impact estimate for the spending domain comprises the estimates of Medicare Part A and B spending for the entire sample (shown above) and for the high-risk sample (reported in the chapter only).

*Significantly different from zero at the .10 level, one-tailed test.

**Significantly different from zero at the .05 level, one-tailed test.

***Significantly different from zero at the .01 level, one-tailed test.

ED = emergency department; FFS = fee-for-service; IT = information technology; IVD = ischemic vascular disease; pp = percentage point.

I. INTRODUCTION

This report presents findings from the evaluation of TransforMED's Health Care Innovation Award (HCIA), with a focus on program impacts on patient outcomes. Section II provides an overview of TransforMED's HCIA-funded intervention and the design of the impact evaluation. Section III describes the design and implementation of the program, including how the program could be expected to affect study outcomes through changes in patient and clinician behavior. In Section IV, we assess the evidence of the extent to which planned changes in clinician behavior occurred. Section V describes our methods for, and results and conclusions from, estimating program impacts on patient outcomes in three domains: quality-of-care processes, service use, and spending. Section VI discusses our findings and describes our conclusions drawn from synthesizing the impact and implementation findings.

II. OVERVIEW OF TRANSFORMED'S HCIA-FUNDED INTERVENTION AND THE IMPACT EVALUATION

A. TransforMED's HCIA-funded intervention

TransforMED, a subsidiary of the American Academy of Family Physicians that guided efforts to transform primary care practices to patient-centered medical homes (PCMHs), received $20.8 million in HCIA funding to implement a patient-centered medical neighborhood (PCMN) program. The PCMN is a model of care founded on the principles of the PCMH. For its HCIA intervention, TransforMED—which closed in 2015—focused almost exclusively on implementing health information technology (IT), providing software and technical assistance to 90 primary care practices nominated by 15 health systems recruited by TransforMED to participate in the PCMN program. Practices were recruited based on several criteria, including Medicare fee-for-service (FFS) and Medicaid patient volume, active use of a practice management system and an electronic health record (EHR) for at least one year, collection of patients' email addresses and/or efforts to engage patients in a patient portal, and leadership and staff motivation to engage in the PCMN program and work toward practice transformation. TransforMED referred to each health system and the practices recruited by the health system to participate in the PCMN program as a *community.* Practices joined the PCMN program in January 2013 and HCIA-funded program services ended in June 2015.

TransforMED's goal for its HCIA program was to reduce the cost of health care for Medicaid and Medicare FFS patients by 4 percent (or $49.5 million) by the end of the award (Table II.1). However, this target was reduced to $44.5 million when it became clear that timely data feeds regarding pharmacy costs were not available to support financial analysis and impact modeling. TransforMED expected to achieve this goal through two interrelated program components: (1) providing population health management and cost-reporting software to practices so they could more effectively use data to improve clinical processes (for example, producing reports that support identifying high-risk patients who might benefit from preventive services or care coordination); and (2) technical assistance to practices and participating health systems on how to use the software, and to promote practice transformation around PCMH principles. TransforMED expected that these program components would lead to reductions in patients' need for acute care—such as inpatient admissions and emergency department (ED)

visits—and improve coordination of care across providers within the PCMN community. The reductions in acute care and improvements in coordination of care were expected, in turn, to reduce total Medicare spending. (Section III.A.3 describes the awardee's theory of action in detail.)

Table I.1. Summary of TransforMED's PCR program and our evaluation for estimating its impacts on patients' outcomes

Program description	
Award amount	$20,750,000
Award start date	June 2012
Implementation date	January 1, 2013[a]
Award end date	June 30, 2015
Awardee description	TransforMED was a national learning and dissemination contractor that guided transformation efforts in primary care practices across the country. TransforMED was a nonprofit subsidiary of the American Academy of Family Physicians, which closed TransforMED in 2015.
Intervention overview	TransforMED used its HCIA funding to implement a patient-centered medical neighborhood (PCMN) program, a model of care founded on the principles of the patient-centered medical home (PCMH). The goal of the TransforMED program was to promote primary care transformation, coordination, and integration across provider organizations, thereby decreasing overall health care costs and improving patients' health and experiences with care. The program recruited 15 health systems and 90 primary care practices to implement the PCMN program.
Intervention components	1. **Health Information technology (IT).** The TransforMED program focused almost exclusively on health IT, providing software to practices so they could more effectively use data to improve clinical processes, such as by monitoring utilization and spending across their patients, identifying patients who would benefit from preventive services, and sending automated emails encouraging patients to schedule recommended follow-up visits. Specifically, TransforMED implemented three IT tools at participating practices: a. **Population management systems.** TransforMED worked with Phytel, a health care technology company, to implement two types of population management software in participating practices: Phytel Insight™, a software data organization program, and Phytel Coordinate™, an automated care management program. TransforMED expected that the combination of these data organization and automated care management capabilities would enable practices to facilitate improvements to specific quality indicators and patient populations. b. **Cost management reporting.** Practices implemented Cobalt Talon cost management reporting software to support the analysis of Medicare fee-for-service (FFS) claims and to generate dashboard reports on utilization and cost of care at the community and practice levels. c. **Data analytics.** The third tool, data analytics, was designed to integrate the population management and the cost management data to target patients whose care could be improved and whose cost of care could be reduced through improved coordination of care across providers with the PCMN. Data analytics was added in the third year of the program. 2. **Technical assistance to health systems and practices.** TransforMED provided regular on-site and virtual support to practices and health systems to learn how to implement population management systems and cost management reporting functions and on PCMH concepts to guide practices in the use of new forms of patients' data to improve care. TransforMED support included biannual communitywide learning collaboratives, monthly conference calls, quarterly community leadership meetings, and cross-community learning and PCMN collaboration.

Table I.1 *(continued)*

Target population	All patients treated at the 90 participating practices
Target impacts on patients' outcomes	• Reduce the cost of health care for Medicaid and Medicare FFS patients by 4 percent (or $49.5 million) • Improve condition-specific quality measures by 15 percent • Improve patients' experiences by 25 percent
Workforce development	The program funded no new clinical positions. The program provided training for existing practice staff. One health coach was identified from existing staff in each practice to attend a training conducted by the Iowa Chronic Care Consortium. Health coaches learned about motivational interviewing, health coaching, population health, and risk-stratification. Health coaches served as experts for population management in each practice and helped practices implement PCMH workflows that targeted specific quality indicators and patient populations. Three staff were identified as super users (the main point of contact for implementation of the cost management reporting) from existing staff in each community to attend a two-day training conducted by Cobalt Talon in which super users learned how to generate reports from the Cobalt Talon system and discussed health IT, clinical integration, and PCMH and PCMN concepts. Super users served as experts on cost management reporting within their PCMNs. TransforMED hosted two follow-up telephone calls with super users to discuss their experiences and challenges using the cost-management reports generated from Cobalt Talon.
Location	Multistate (urban, rural and suburban areas): Alabama, Connecticut, Florida, Georgia, Indiana, Kansas, Kentucky, Maryland, Massachusetts, Michigan, Mississippi, Nebraska, North Carolina, Oklahoma, and West Virginia
Impact evaluation	
Core design	Difference-in-differences with matched comparison group
Treatment group	Medicare FFS beneficiaries who we attributed to treatment practices, using an algorithm similar to that used by CMMI for the Comprehensive Primary Care Initiative
Comparison group	Medicare FFS beneficiaries we attributed to matched comparison practices, using the same rules we used for the treatment group
Intervention component(s) included in impact evaluation	**All components described above.** All major components of the practice transformation initiative could potentially affect the outcomes of all beneficiaries in the treatment group.
Extent to which the evaluation's treatment group reflects the awardee's full target population (for component[s] evaluated)	**Medium**. The awardee's target population was all patients of participating practices, whereas the evaluation's full treatment group was restricted to participating practices' Medicare FFS population. This treatment group did not include Medicare Advantage beneficiaries, Medicaid beneficiaries and those patients who had private coverage or were uninsured, all of whom might be affected by some or all of the intervention components. We included 87 of the 90 participating practices in the treatment group. Three participating practices were dropped because we were unable to attribute Medicare beneficiaries for several program quarters.
Study outcomes, by domain	1. **Quality-of-care processes.** Preventive care for diabetes, lipid testing for patients with IVD, and 14-day follow-up to hospitalization 2. **Service use.** All-cause inpatient admissions and outpatient ED visits 3. **Spending.** Medicare Part A and B spending

Sources: Review of TransforMED reports, including its original application, operational plan, and 15 quarterly narrative reports to CMS.

[a] TransforMED began implementing the PCMN program in November 2012 after receiving CMMI's approval of its operational plan, but participating health systems could not begin recruiting practices and implementing the PCMN program until January 2013, when they received approval of their operational plans.

CMMI = Center for Medicare & Medicaid Innovation; ED = emergency department; FFS = fee-for-service; HCIA = Health Care Innovation Award; IVD = ischemic vascular disease; PCR = primary care redesign.

B. Overview of impact evaluation

To estimate program impacts on patients' outcomes, we compared outcomes for Medicare FFS beneficiaries served by 87 of the 90 practices participating in TransforMED's HCIA intervention (treatment practices) with outcomes for Medicare FFS beneficiaries served by 286 matched comparison practices, adjusting for any differences in outcomes between these two groups of practices before the intervention began. We summarize our impact evaluation design in the bottom panel of Table II.1.

We selected the 286 comparison practices for the evaluation from the pool of all nontreatment practices in the same market areas where the TransforMED intervention was implemented. As described in Section V.A.3, we selected practices that were similar to the 87 treatment practices in factors that can influence patients' outcomes, especially those that TransforMED used when determining practices to recruit for the intervention.

We estimated impacts on outcomes, as measured in Medicare FFS claims data, which we grouped into three domains: (1) quality-of-care processes, (2) service use, and (3) spending. We did not assess impacts in a fourth domain—quality-of-care outcomes—because the TransforMED intervention did not expect to affect any of the outcomes we could observe in that domain using claims data. Across the HCIA awardees implementing primary care redesign (PCR) programs, we designed our impact evaluations to identify promising interventions or intervention components—consistent with evaluation goals from the Center for Medicare & Medicaid Innovation (CMMI) to find programs that could be scaled or retested as part of a future model test. Before conducting the analysis, we specified a series of primary tests, describing the evidence we would need to conclude that the program was effective; TransforMED and CMMI reviewed these tests. Each test specified a population, outcome, period, expected direction of effect, and threshold that we count as substantively important. The purpose of these primary tests was to focus the impact evaluation on hypotheses that would provide the most robust evidence about program effectiveness. We used the results from the primary tests and robustness checks (secondary tests) to draw conclusions about program impacts in each of the three evaluation domains. Because we sought to identify promise, rather than only those programs with unequivocally demonstrated success, we conducted one-sided statistical tests (that is, testing only for program benefits) and used a threshold for statistical significance of 0.10, which is not as strict as the conventional standard of 0.05.

Our impact evaluation design aligns well with several aspects of TransforMED's PCMN program, as the evaluation should reflect the effects of both program components. However, the evaluation is limited in that it captures only a portion of the full target population. Whereas all patients treated at participating practices were eligible to receive PCMN-related services, due to limitations in data availability, the evaluation's treatment group was restricted to participating practices' Medicare FFS population. Furthermore, we did not include three participating practices in the treatment group because we were unable to attribute Medicare beneficiaries for several program quarters; see Section V.A.2 for detail on the treatment group.

III. PROGRAM IMPLEMENTATION

This section provides a detailed description of TransforMED's HCIA-funded program, highlighting how it evolved over time and its theory of action. It also assesses the evidence on the extent to which the program was implemented as planned based on measures of enrollment, service delivery, staffing, training, and timeliness of program implementation. We then summarize the facilitators and barriers associated with implementation effectiveness.

We based our evaluation of TransforMED's program implementation on reviews of the awardee's quarterly reports to CMMI and self-monitoring program metrics, telephone discussions and follow-up communications with program administrators, and in-depth interviews with health system leadership and practice staff conducted during site visits in April 2014 and March through April 2015. We did not verify the quality of the performance data reported by the awardee in its self-measurement and monitoring reports.

A. Program design and adaptation

1. Target population and patient identification, recruitment, and enrollment

TransforMED's HCIA-funded program targeted all patients treated at participating practices. However, the program's *primary* target population was Medicare and Medicaid beneficiaries treated at participating practices, in the sense that TransforMED projected its cost savings only for this population. In this section, we describe TransforMED's target population and how some aspects of TransforMED's program focused on rising- or high-risk beneficiaries at rising or high risk of hospitalization and other costly care.

Identification of health systems and primary care practices. Before implementing the PCMN program, TransforMED had collaborated with VHA (known as Vizient since 2015), a national network of nonprofit health care organizations, on implementing primary care practice transformation initiatives, such as the PCMH. TransforMED worked with the previously named VHA to identify 15 health systems for participation in the PCMN program. TransforMED required the 15 health systems to nominate multiple family medicine or internal medicine practices for participation in the PCMN program based on the following criteria. The practices must have (1) a patient volume of at least 350 attributed Medicare FFS and Medicaid patients per provider; (2) a practice management system and an EHR that practice staff had actively used for at least one year, and the ability to accommodate new software requirements for exporting data; (3) strong, motivated leadership and staff willing to actively engage in quality improvement; (4) a collection of patients' email addresses so that practice staff could communicate with patients outside of practice visits, and/or a patient portal that patients were encouraged to use; (5) documentation of a patient's preferred primary care provider (PCP) in the practice management system; and (6) strong commitment to participating in the PCMN and the potential to be recognized as a leader within the community. Health systems were to nominate both system-affiliated and independent practices. TransforMED worked with VHA and the population health management software vendor to select the nominated practices. The incentives for health system and practice participation in the program included the availability of population health management software, cost reporting software, and technical assistance with implementing

PCMH principles for the award period. Health systems and practices did not receive a financial incentive for participating in the PCMN program.

Target population of patients. TransforMED implemented the PCMN program at the practice level; it encouraged practices to implement program processes for all patients in the practice, regardless of payer. Therefore, all patients had the potential to benefit from the program. For example, the population-management software organized data on all patients who visited a participating practice, regardless of insurance status. One of the program's health IT innovations was more restrictive: the cost-management reporting tool applied only to Medicare FFS beneficiaries.

Identification, recruitment, and enrollment of patients. All patients treated by a PCP in a participating practice were passively enrolled in the program; no formal enrollment procedures existed. Depending on the PCMH concepts implemented by each practice, these patients received team-based care, risk-stratified care management, and/or planned care for chronic conditions. Some aspects of TransforMED's program focused on improving care for patients at high risk of hospitalization and other costly care. Practices could use a variety of methods to identify high-risk patients, including quality indicators; cost and utilization metrics; anecdotal patient information; internal automated algorithms; or the Milliman Advanced Risk Adjusters model, which calculated risk scores based on Medicare FFS claims.

2. Intervention components

TransforMED's program had two components, both focused almost exclusively on health IT. Specifically, in the first component, TransforMED's program provided and assisted practices and health systems in implementing three health IT tools: population health management software; cost management reporting software; and data analytics reports to integrate the population health and cost management data. The second component was technical assistance: TransforMED provided regular on-site and virtual support to practices and health systems, including learning collaboratives, site visits, conference calls, community leadership meetings, and cross-community learning and collaboration. The program intended that the three health IT tools would become part of routine service delivery in the participating practices after the intervention ended, but the technical assistance would not.

a. Component 1

Population health management systems. TransforMED worked with Phytel, a health care technology company, to implement two types of population health management software in participating practices: Phytel Insight™ and Phytel Coordinate™. Phytel Insight organized clinical data by patient and population characteristics and quality indicators. Phytel Coordinate automated care management processes within practices by providing care teams with the following capabilities: (1) patient attribution, which involved assigning patients to primary care providers who were responsible for coordinating their care needs; (2) risk-stratification, which involved assessing a patient's health risk status and categorizing the patient based on his or her care needs; and (3) patient outreach, which involved targeting communications to patients based on their individual care needs, as measured by quality indicators. TransforMED expected that the

combination of these data organization and automated care management capabilities would enable practices to target improvements to specific quality indicators and patient populations.

TransforMED did not specify required changes to patient care protocols, such as stratifying and prioritizing specific patient populations with care needs or providing care management services to high-risk patients, although TransforMED expected that practices would implement new patient care protocols as a result of having access to the tools they received through the program. TransforMED intended practices to use the population health management systems to fully incorporate data automation to eliminate manual tasks, enabling care team members to spend less time searching for information and more time focusing on patients' care. Health coaches were intended to lead the work in this area; clinicians had no specified role, but each practice could determine how involved clinicians were in the program.

Cost management reporting. TransforMED provided practices with Cobalt Talon cost management reporting software. The cost management reporting software gave practices the capability to generate dashboard reports on utilization and cost of care at the practice and health system levels, supporting the analysis of Medicare FFS claims and identification of utilization and cost issues at both levels of delivery.

Data analytics. TransforMED designed data analytics reporting to integrate the population health management and the cost management data, which would identify patients whose care could be improved and whose cost of care could be reduced through improved coordination of care across providers within the PCMN. For example, data analytics gave practices the ability to identify patients seeking care from multiple specialists and knowledge of those patients' outcomes of specialist care. TransforMED provided health systems and practices with patient profile reports that gave practices information on all the services received by a patient and a risk score based on cost and utilization. TransforMED also provided practices with the Cave Grouper tool, which gave practices the ability to compare physicians' efficiency (based on cost and utilization metrics) against a national peer group, and produce efficiency scores for physicians by certain diagnoses and procedures. The Cave Grouper tool also gave individual providers, practices, and health systems information on patient referral patterns and providers' efficiency related to specific disease conditions or specialties. Data analytics activities were introduced in the second half of 2014 after most practices had implemented population health management systems and cost management reporting functions. However, these activities were discontinued because of the three- to six-month lag in the claims-based cost management data, which limited the utility of the patient profile reports.

b. Component 2

Technical assistance to health systems and practices. TransforMED provided regular on-site and virtual support to practices and health systems to provide guidance on implementing population health management systems and cost management reporting functions, and on PCMH concepts related to the use of new forms of patients' data to improve care. TransforMED support included biannual communitywide learning collaboratives, monthly conference calls, quarterly community leadership meetings, and cross-community learning and PCMN collaboration to promote practice transformation.

3. Theory of action

Based on extensive review of TransforMED's program activities and goals, we developed a theory of action to depict the mechanisms through which program administrators expected the program to improve the outcomes we selected for the impact evaluation (Table II.1 lists these outcomes). TransforMED expected that the HCIA-funded program would improve outcomes through two pathways.

Primary pathway to improved outcomes. Participating primary care practices generated new forms of patients' data to improve quality-of-care processes, manage inappropriate service use (such as inappropriate ED visits), and reduce Medicare spending. This pathway included the following planned mechanisms:

1. **Population health management led to systematically identifying patients in need of preventive services.** Identifying patients in need of preventive services enabled practices to target communications to patients based on their individual care needs, which would increase the extent to which a practice's patients received recommended clinical care. This could include patients receiving recommended preventive care for diabetes and lipid testing for ischemic vascular disease (IVD).
2. **Risk-stratifying patients led to improved care management support for patients with chronic conditions who were at high risk for acute care use.** This care management support could include activities such as increased outreach to high-risk patients and helping them with medication adherence and self-management of their conditions, including improved diet and exercise regimens. Increased care management support could ensure that patients received recommended preventive care for diabetes and lipid testing for IVD. Increased care management support could also reduce the frequency of acute exacerbations and, therefore, the need for acute care (outpatient ED visits and inpatient admissions). Because acute care drives overall Medicare spending, reductions in acute exacerbations should reduce total Medicare spending.
3. **Cost management reporting led to a focus on developing care coordination processes to reduce high rates of ED use and inpatient admissions.** These care coordination processes could include primary care office staff contacting patients following an inpatient admission to ensure they received ambulatory follow-up care within 14 days of discharge. These processes could also include primary care office staff contacting patients following an ED visit to ensure they received necessary ambulatory follow-up care and educate patients regarding the appropriate use of the ED. These care coordination processes could reduce inpatient admissions and ED visits.

Secondary pathway to improved outcomes. Participating practices used data analytics, which integrated population health management and cost management data. This pathway included the following planning mechanism:

1. **Data analytics supported practices in identifying potentially inappropriate practice patterns (such as imaging for back pain) and patients whose care could be improved and whose cost of care could be reduced through improved coordination of care within**

the PCMN. In turn, PCPs had information to identify high-performing specialists (those who provide care in line with recommended practice patterns) and improve the coordination of their patients' care and/or modify referral patterns to reduce redundant or unnecessary services. Patients were more likely to receive cost-effective care in appropriate settings, all in the larger effort of providing better care to patients and reducing total Medicare spending.

Text box III.1. Example from TransforMED's quarterly reports illustrating the program's theory of action

"… [One] health system [in STATE] is identifying 100 patients with the highest utilization of the ED to determine care management/coordination support needs. From the top 100 ED utilizers for the system, they contacted their top ER [emergency room] patient by working through the care coordinators in the inpatient setting. They all gathered together in the patient's room in her most recent visit and introduced themselves and expressed their concern with her health. Prior to this, they had not been able to reach the patient by phone. After this visit, the patient is now calling the care coordinator and the practice for education, guidance, medication reconciliation, etc. The patient has not been back in the ER for the past three weeks. She had visited the ER over 43 times in their [the health system's] prior reporting period."

Source: TransforMED's 6th quarterly report to the Centers for Medicare & Medicaid Services.

4. Program staff and workforce development

Table III.1 provides key details about existing staff with new roles for the HCIA-funded program. It is important to note that TransforMED hired new staff for the program, but these staff primarily provided technical assistance and were not directly involved in the processes in TransforMED's theory of action. The participating health systems and practices did not receive funding to hire staff under the award. However, at least one care manager was hired in each community during the funding period; these new hires were not funded by HCIA.

Table III.1. Key details about intervention staff and workforce development

Program component	Staff members	Staff/team responsibilities	Adaptations?
Population management systems	One existing staff member in each practice (for example, nurse care manager or comparable position)	Served as the main point of contact for using the population management information generated by Phytel; however, the role had no specific requirements; participating staff were referred to as clinical health coaches	No significant adaptations identified
Cost management reporting	Three existing clinical and/or administrative staff in each community	Served as the main point of contact for implementation of the cost management reporting generated by Cobalt Talon; participating staff were referred to as super users	Relied on standardized reports developed by Cobalt Talon; it took too much staff time for super users to learn to build and run their own customized reports

Sources: Interviews from second site visit, April 2015; document review, March 2016.

B. Implementation effectiveness

In this section, we examine the evidence on implementation effectiveness—that is, we analyze measures of the intervention delivered and, when possible, compare those measures with the services the awardee intended to deliver. We assess the evidence on implementation effectiveness in five areas: (1) program enrollment, (2) service delivery, (3) staffing, (4) training, and (5) implementation timeliness. To conduct this analysis, we used data from interviews with program administrators and frontline staff, and self-reported metrics included in TransforMED's self-monitoring and measurement reports to CMMI.

1. Program enrollment

TransforMED initially recruited 15 health systems and 90 primary care practices. All patients, regardless of payer, were passively enrolled in the program. TransforMED projected the total number of these potentially affected patients who were also Medicare or Medicaid beneficiaries; we call these patients indirect program participants. TransforMED projected 872,647 indirect participants in each quarter from July 2013 through July 2014, and 1,154,011 indirect participants in each quarter from July 2014 through June 2015. Projections are not available for the period from January to June 2013, a period corresponding to the first two quarters of our evaluation intervention period. According to TransforMED reports, the number of indirect participants who actually participated in the program varied by quarter and ranged from an estimated 845,980 (July through September 2014) to 1,058,405 (April through June 2014). Overall, TransforMED came very close to meeting its target number of indirect participants each quarter.

2. Service-related measures

The service metrics TransforMED reported were not direct outcomes of health IT implementation or technical assistance, but instead focused on patient–provider contact measures (such as number of visits and telephone follow-ups) and clinical process measures (such as number of screenings). A number of these measures were retired when communities achieved scores of 90 to 100 percent within their patient panels. Other measures that were targeted and later retired included availability of same-day appointments and extended office hours. Practices were guided, not required, to implement PCMH concepts reflected in these service metrics. Because of this, we did not collect qualitative data to aid interpretation of the PCMH-focused metrics.

TransforMED used a variety of strategies and activities to facilitate the implementation of the PCMN program in the 15 communities; some strategies were structured (for example, communitywide learning collaboratives), whereas others occurred as needed (for example, conference calls with practices).We do not have quantitative information on the delivery of technical assistance activities intended to support PCMN implementation. However, the site visit interview data we collected from two communities indicated that different communities had different perceptions of the benefit of the technical assistance TransforMED provided.

3. Staffing measures

As mentioned in Section III.A.4, TransforMED hired new staff to provide technical assistance; these positions did not directly support mechanisms described in TransforMED's theory of action. These positions included project managers, facilitators, one trainer, one project data analyst, one program director, one project control manager, and a part-time administrative support staff member. TransforMED met approximately 78 percent of its target for staffing these positions.

4. HCIA-funded training

TransforMED provided training to practice staff, specifically to health coaches and super users (participating staff who served as the main point of contact for using the population management information were referred to as health coaches while participating staff who served as the main point of contact for implementation of the cost management reporting were referred to as super users), to help practices implement the health IT component of the program. Each practice selected an existing staff member to receive training as a health coach. Health coaches attended an in-person training in 2013–2014 and virtual trainings during the rest of the award period. The Iowa Chronic Care Consortium conducted these trainings. Health coaches learned about motivational interviewing, evidence-based health coaching, population health and risk-stratification, and coaching using the Myers-Briggs Type Indicator® (a questionnaire used to characterize individuals' personalities and work styles). Each community also selected three existing staff members to be super users of cost management reports from Cobalt Talon. In 2013–2014, TransforMED partnered with Cobalt Talon to provide a two-day training for super users to learn how to generate reports from that reporting system. The training included discussions on health IT, clinical integration, and PCMH and PCMN concepts. TransforMED hosted two follow-up telephone calls with super users to discuss their experiences and challenges using the Cobalt Talon reports. In addition, TransforMED trained all practices on using quality improvement processes (in the form of plan-do-study-act cycles) to make quality-of-care process improvements that aligned with PCMH concepts.

To assess perspectives of HCIA-funded staff who received these trainings, we administered the HCIA Primary Care Redesign Trainee Survey from January to March 2015 (three to six months before the end of the HCIA funding). We sent the survey to the 211 people TransforMED identified as practice staff at participating practices and health systems who had received HCIA-funded training. This group included administrators and directors, care coordinators and care managers, data analysts, nurses, and medical assistants, among others. The overall response rate to the survey was 62 percent.

In regard to training received, 44 percent of respondents reported receiving health coach training, 38 percent reported receiving super user training, and 64 percent reported receiving training on population health management systems. A small percentage of respondents (8 percent) reported receiving training they classified as other.

As shown in Table III.2, of the 122 respondents eligible for the survey (that is, they responded yes to the screener questions that they worked at a participating practice or health

system and were involved in the PCMN program), most respondents (82 percent) reported that the training they received was good or excellent. Many respondents reported that the training had improved their ability to provide care in a way that aligned with PCMH concepts, including explaining information about patients' care to patients and their families in lay terms (36 percent), relaying relevant information to the care team (39 percent), working with a diverse set of patients (37 percent), accessing the care patients need (36 percent), helping patients take control of their own care (38 percent), and using data to evaluate staff performance to improve the services they provide to patients (42 percent). Fewer than a quarter (24 percent) of trainees reported that the training improved their ability to help patients access nonmedical services, which is not surprising as the TransforMED program did not have a strong focus on linking patients to nonmedical services. Overall, most trainees (55 percent) thought the training had a positive effect on their patient-centeredness.

Table III.2. TransforMED staff perceptions of the effects of training on the care they provided to patients, from the trainee survey

Survey question		Percentage of 122[a] respondents (and number) who rated the training
Please indicate how you would rate all of the training you received.	1. Excellent	30% (31)
	2. Good	52% (55)
	3. Fair	17% (18)
	4. Poor	<11
Survey question		**Percentage of 122[a] respondents (and number) who reported the training had a positive effect on this dimension of their care**
Please indicate the impact you believe the training you received for the TransforMED program has had on the following aspects of care you provide to patients enrolled at your practice site	1. Quality of care	50% (59)
	2. Ability to respond in a timely way to patients' needs	35% (41)
	3. Efficiency/cost-effectiveness of care	38% (45)
	4. Patient-centeredness	55% (64)
	5. Equity	45% (53)
Please indicate whether the training you received has had a positive or negative effect on your ability to …	1. Explain information about patients' care to patients and their families in lay terms	36% (42)
	2. Relay relevant information to the care team	39% (46)
	3. Work with diverse set of patients	37% (43)
	4. Access the care they need	36% (42)
	5. Help patients access nonmedical services	24% (28)
	6. Help patients take control of their own care	38% (45)
	7. Use data to evaluate my performance to improve the services I provide to patients	42% (49)

Note: Questions with less than 11 responses are suppressed because the numerator is less than 11.

[a] The denominator includes all 122 people found eligible for the survey (that is, they responded yes to the screener questions that they [1] worked at a participating practice or health system and [2] were involved in the PCMN program).

PCMN = patient-centered medical neighborhood.

5. Program timeline

TransforMED began implementing the PCMN program in November 2012 after receiving CMMI's approval of its revised operational plan. This was four months later than initially planned due to a period waiting for CMMI to approve the revised operational plan. (CMMI requested several revisions and a budget modification.) During the four-month waiting period, TransforMED continuously updated the project timeline and milestones to ensure that it could quickly start implementation upon approval of the operational plan. Participating health systems could not begin recruiting practices and implementing the PCMN program until they received approval of their operational plans, which occurred in January 2013.

C. Summary of facilitators of and barriers to implementation

Several factors facilitated implementation of TransforMED's HCIA-funded program, and others hindered it. We described those factors in detail in the second annual report (Keith et al. 2015). Here, we summarize key facilitators and barriers, along with any new information since the second annual report that supports those facilitators or barriers (Table III.3).

One characteristic of the TransforMED program that facilitated PCMN implementation in the communities and participating practices was the perceived advantage of the availability of cost and quality data. Respondents in both communities at which we conducted site visits felt they benefited from the availability of cost and quality data provided by the program and used these data to monitor quality improvements.

Three important barriers to implementation included challenges (1) implementing the population health management system, (2) implementing cost management reporting, and (3) dedicating resources to the PCMN program. The participating practices experienced difficulty implementing the population health management software. For this software to effectively organize data, practices had to ensure that the software mapped to their EHR systems correctly. During our site visits in both 2014 and 2015, some respondents described initial challenges mapping the software to their EHRs because of how they had previously entered data and the location of the data in the EHR system. The participating practices also experienced several difficulties using cost management reporting as intended. Due to the technical demands on using the cost management reporting software to generate reports, super users did not customize cost management reporting to guide practice-specific cost management activities. Communities were able to generate standardized reports developed only by the cost management reporting vendor, Cobalt Talon, as opposed to customized reports developed by the practices to examine unique utilization and cost issues within each practice. Therefore, health systems and practices did not use cost management reporting to change the way they delivered or monitored their patients' care. In addition, in the communities that included independent practices, the availability of the cost data at the practice level led to conflicts about the use of cost management reporting, due to financial competition between the health system and the independent practices. In regard to dedicating resources to TransforMED's program, participating practices and health systems did not receive funds to support PCMN implementation and respondents across both communities acknowledged this as a challenge to implementation. As a practice manager said during our 2015 site visit, "Let's be clear. There was no money. The money makes a big difference. It doesn't

have to be a huge amount. I don't think we are asking to pay for a physician's time. Just something that makes it feel like you can support your physician to attend this meeting. That is important to support this work."

Table III.3. Summary of key facilitators of and barriers to the implementation of TransforMED's program

Item	Description based on findings in the second annual report	New data (if applicable) that help to support our listing of this item as a facilitator or barrier
	Facilitators	
Perceived relative advantage of the availability of cost and quality data (program characteristic)	During our site visits, respondents reported that they benefited from the availability of cost and quality data provided by the program and used these data to monitor quality improvements and improve population management process, such as identifying and following up on care gaps.	No new data
	Barriers	
Challenge implementing population management system (internal factor)	Practices experienced difficulty implementing the population management system software. For the software to pull data electronically, practices had to ensure that the software mapped to their electronic health record (EHR) systems correctly. During our 2014 and 2015 site visits, respondents described challenges mapping the Phytel software to their EHRs because of how data had previously been entered and where the data were located in the EHR. In both communities, respondents at practices that did not correctly map the software to their EHRs used inaccurate data on care gaps to notify patients they believed were due for services but in reality were not. As a result, a number of patients expressed frustration with the practices.	Based on responses to the trainee survey, trainees were almost equally divided in reporting on whether there was adequate health IT to help them perform their job duties. Half of the respondents said it was not a barrier but 47 percent said it was either a minor or a major barrier. As health IT is one of the major program components, one would expect that the number of trainees reporting no barrier would be significantly higher.
Challenge dedicating resources to program (implementation process)	Practices and health systems did not receive funds to support PCMN implementation, and respondents across both communities acknowledged this as a challenge to implementation. Another expense not covered under the award was physicians' time, either to spend more time with patients or to attend PCMN-related meetings.	No new data

Source: Keith et al. 2015.

Note: We reviewed four domains associated with implementation experience: (1) program characteristics, (2) implementation process, (3) internal factors, and (4) external environment (Keith et al. 2015). Implementation research suggests that barriers and facilitators within these domains are important determinants of implementation effectiveness.

IT = information technology; PCMN = patient-centered medical neighborhood.

D. Conclusions about the extent to which the program, as implemented, reflected the core design

Despite not implementing some aspects of the program to the full extent expected (such as the cost management reporting functions), TransforMED implemented its HCIA-funded program largely as planned, and well enough to be a reasonable test of the PCMN program's core design. As noted previously, TransforMED successfully recruited 15 health systems and 90 practices without any major delays. TransforMED also met its goal of recruiting practices with sizeable Medicare and Medicaid patient populations, as evidenced by information on the payment source of indirect program participants (Section III.B.1).

In addition, TransforMED implemented the two health IT systems in participating practices. Of 90 participating practices, 78 implemented the population health management systems software. Practices' scores on patient contact and process measures—many of which TransforMED retired due to extremely high values—were consistent with successful implementation of the population health management systems. Although there were some initial implementation challenges, the population health management systems gave practices the ability to run reports from their EHRs to make quality improvements related to PCMH processes. In regard to Cobalt Talon, we do not have evidence on the number of practices that used cost management reporting; however, qualitative evidence from TransforMED self-monitoring reports suggests all 15 communities had the ability to use cost management base reporting functions. However, several difficulties prevented practices from using cost reporting as intended, including technical challenges using Cobalt Talon to generate reports and conflicts related to using cost management reports due to financial competition between the convening health system and the nonsystem practices.

We also concluded that TransforMED implemented its HCIA-funded program largely as planned because it nearly reached its goals for providing training to practices and communities to learn how to implement population health management systems and cost management reporting functions. These goals included the number of health coaches trained on population health and the number of super users trained on Cobalt Talon. Throughout the term of the award, TransforMED also provided technical assistance to communities and practices to promote practice transformation.

IV. CLINICIANS' PERCEPTIONS OF PROGRAM EFFECTS ON THE CARE THEY PROVIDE TO PATIENTS

This section describes the available evidence on the extent to which TransforMED's program had its intended effects on changing clinicians' behavior as a way to achieve desired impacts on patients' outcomes. As described in Section III.A.3, according to the program's theory of action, the primary pathway did not specify how clinicians should change the ways in which they treated or interacted with patients, but gave them information and tools to improve their existing care processes. Practice managers and allied health professionals, such as care managers, were the primary implementers of the program, so that most clinicians, especially physicians, did not have a formal role in the program. However, in the secondary pathway, the use of physicians' efficiency scores via Cave Grouper indicated that clinicians were intended to

play a role in data analytics (for example, using the data to target patients whose care could be improved and whose cost of care could be reduced through improved coordination of care across providers within the PCMN community). However, TransforMED did not explicitly draw a connection between the use of data analytics and clinicians' involvement in the program.

A. Clinician survey

Survey methods. We administered the clinician survey in two rounds (fall 2014 and summer 2015). We sent the survey to PCPs (physicians, nurses, nurse practitioners, and physician assistants) working in the 90 participating practices at the time of each survey. A total of 319 and 110 clinicians participating in TransforMED's HCIA program responded to the survey during the first and second rounds, respectively (a response rate of 63 percent in round 1 and 51 percent in round 2). As noted, because only the data analytics tool (with its physician efficiency measures) was intended to change the way that all clinicians provided care, and because this tool was introduced during the third year of the program, we describe here only the round 2 survey results.

Survey results. Most surveyed clinicians (64 percent or 70 clinicians) reported being somewhat or very familiar with the HCIA program (23 percent reported being very familiar). As shown in Table IV.1, among these 70 clinicians, fewer than half reported that the program improved care on all dimensions. Most of these 70 clinicians reported that there was either no impact or it was too soon to tell in regard to their ability to respond in a timely way to patients' needs (51 percent), efficiency (51 percent), safety (54 percent), equity (66 percent), and information available for clinical decision making (57 percent).

Table IV.1. Clinicians' perceptions of the effects of the program on the care they provided to patients, from the clinician surveys (round 2)

	Percentage (and number) of clinicians reporting that the HCIA had the following effect on the care they provided to patients enrolled in their practice in the past year	
	Second round of survey (summer 2015) N = 70	
Dimension of care	Positive impact	No impact or too soon to tell
Quality	47% (33)	43% (30)
Ability to respond in a timely way to patient needs	36% (25)	51% (36)
Efficiency	31% (22)	51% (36)
Safety	40% (28)	54% (38)
Patient-centeredness	47% (33)	43% (30)
Equity	27% (19)	66% (46)
Information available for clinical decision making	37% (26)	57% (40)

Source: Clinician Survey Round 2 (field period May – July 2015).

Note: The numbers and percentages are limited to primary care providers who reported that they were at least somewhat familiar with the HCIA program. Most clinicians surveyed, 64 percent (70), reported being at least somewhat familiar with the program.

HCIA = Health Care Innovation Award.

B. Conclusions about intermediate program effects on clinicians' behavior

Based on available information, the HCIA-funded program appears not to have had its intended effects in the secondary pathway of the theory of action—that is, in changing the care that most primary care practices provided through the use of data analytics software. Although most clinicians (64 percent) reported familiarity with the program, only 23 percent of them were very familiar with the program. This might indicate that clinicians knew about the formal program, but most of them were not involved with it on a regular basis. The intended use of Cave Grouper's physician efficiency scores to compare physicians' efficiency (based on cost and utilization metrics) against a national peer group indicates that clinicians would have to be involved in the program on a regular basis to improve their cost and utilization scores for certain diagnoses and procedures. In addition, fewer than half of the clinicians who reported familiarity with the HCIA program also reported that they believed the program improved care.

V. PROGRAM IMPACTS ON PATIENTS' OUTCOMES

In this section of the report, we draw conclusions based on available evidence about the impacts of TransforMED's HCIA program on patients' outcomes in three domains: quality-of-care processes, service use, and spending. We first describe the methods for estimating impacts (Section V.A.) and then the characteristics of the 87 treatment practices at the start of the intervention (Section V.B). We next demonstrate that the treatment practices were similar at the start of the intervention to the practices we selected as a comparison group, which is important for limiting potential bias in impact estimates (Section V.C). Finally, in Section V.D, we describe the quantitative impact estimates, their plausibility given implementation findings, and our conclusions about program impacts in each domain. The findings in this report update the impact results from the second annual report for TransforMED (Keith et al. 2015), extending the outcome period by 6 months and adding new outcomes.

A. Methods

1. Overview

We estimated program impacts as the difference in outcomes for Medicare FFS patients attributed to the 87 treatment practices and those served by 286 matched comparison practices, adjusting for any differences between these groups before TransforMED's HCIA intervention began. We prespecified primary tests, describing the evidence we would need to conclude that the program was effective, and the awardee and CMMI reviewed these. Each test specified a population, outcome, time period, expected direction of effect, and threshold that we counted as substantively important. The purpose of these primary tests was to focus the impact evaluation on hypotheses that would provide the most robust evidence about program effectiveness. We used the results from the primary and secondary tests (robustness checks) to draw conclusions about program impacts in each of the three evaluation domains. The remaining subsections describe each component of the impact evaluation in more detail.

2. Treatment group definition

The treatment group consisted of Medicare FFS beneficiaries attributed to 87 treatment practices in four baseline quarters before the intervention began (January 1, 2012, to December

31, 2012) and 10 intervention quarters (January 1, 2013, to June 30, 2015). We could include only 87 of the 90 participating practices in the treatment group. We dropped three participating practices from the analysis because we were unable to attribute Medicare beneficiaries for several program quarters.

We constructed the treatment group in three steps.

1. First, we attributed beneficiaries to practices using an algorithm similar to that used by CMMI for the Comprehensive Primary Care (CPC) Initiative. Specifically, in each baseline and intervention month, we attributed beneficiaries to the primary care practice whose providers (physicians, nurse practitioners, or physician assistants) provided the plurality of primary care services in the past 24 months. When there were ties, we attributed the beneficiary to the practice he or she visited most recently. This attribution method required identifiers for the practice site or the providers who worked in the treatment practices (and when), as well as identifiers for providers in other practices in the comparison regions who could compete for beneficiaries (when determining which practice provided the plurality of primary care services). TransforMED provided identifiers for the treatment providers. We obtained identifiers for the comparison group from SK&A (described in Section V.A.3).
2. Second, in each baseline and intervention period, we assigned each beneficiary to the first treatment practice he or she was attributed to in that period, and continued to assign him or her to that practice for all quarters in the period. This assignment rule, which is distinct from the attribution method, ensures that, during the intervention period, beneficiaries did not exit the treatment group solely because the intervention succeeded in reducing their service use (including visits at treatment practices). The definition for the baseline period corresponds to that of the intervention period so that, across the two periods, interpretation of the population changes over time should be comparable.
3. Third, we applied additional restrictions to refine the analysis sample in each quarter. We included a beneficiary assigned to a treatment practice in a quarter in the analysis sample for that quarter if he or she (1) had observable outcomes for at least one day in the quarter; and (2) lived in the state or state(s), or prespecified surrounding states, of the practice to which the beneficiary was attributed, for at least one day of the quarter. For this sample, outcomes were observable for beneficiaries who were enrolled in Medicare FFS (Part A and B), were alive, and had Medicare as their primary payer.

Definition of high-risk subgroup. In addition to the full treatment sample, because some aspects of TransforMED's intervention (including care management services) focused on improving care for beneficiaries at rising or high risk of hospitalization and other expensive care, we also defined a high-risk subgroup of the treatment group each quarter. For each baseline quarter, this subgroup consisted of the beneficiaries with a January 2012 Hierarchical Condition Category (HCC) score in the top quartile among all beneficiaries who ever visited a treatment practice during the baseline period, by market area. In each intervention quarter, the high-risk subgroup consisted of beneficiaries whose HCC scores were in the top quarter, by market area, among all observable treatment group members at the start of the intervention period. The HCC score is a continuous variable that predicts a beneficiary's Medicare spending in the following

year relative to the national average, with 1.0 indicating that the predicted spending is at the national average and 2.0 indicating that it is twice that average.

3. Comparison group definition

The comparison group consisted of Medicare FFS beneficiaries assigned to 286 matched comparison practices during each quarter of the baseline and intervention periods. We selected comparison practices that were similar during the baseline period to the treatment practices in factors that can influence patients' outcomes, especially those that TransforMED used when determining practices to recruit for the intervention. This section describes how we constructed the matched comparison group; Section V.C shows the balance we achieved between the two groups on the matching variables.

We identified the 286 comparison practices in six steps:

1. To obtain a pool of comparison practices from which to draw our sample, we purchased data on providers practicing in nontreatment practices in the same 15 states that implemented the TransforMED intervention from SK&A, an outside health care data vendor that maintains and verifies lists of providers who work in practices throughout the country. For federally qualified health centers (FQHCs) and rural health centers (RHCs), we obtained CMS Certification Numbers from the Integrated Data Repository for all such health centers in the five states in which FQHCs and RHCs participated in the TransforMED program.
2. We identified market areas from which to draw potential comparison practices. We chose the entire state for 3 of the 15 TransforMED program sites (Kansas, Mississippi, and Nebraska). For each of the remaining 12 sites, we selected a within-state region or, for one site that had treatment practices in two states (Kentucky and Indiana), a region that included a portion of both states. In all cases, we balanced the need for a large pool of comparison practices to ensure a sufficient sample of well-matched comparison practices against the desire to restrict the pool to potential comparison practices located in areas similar to those of treatment practices, ensuring face validity of our approach.
3. We developed matching variables, defined at the start of the intervention period (January 1, 2013), for all treatment and potential comparison practices. These variables included characteristics of all Medicare FFS beneficiaries assigned to the practices (for example, mean HCC score and utilization in the baseline period); characteristics of high-risk beneficiaries assigned to the practices; characteristics of the geographic location of the practices; and, for nonhealth centers, characteristics of the practices overall (for example, the number of providers in the practice or whether a hospital or health system owned the practice). We did not include measures of quality-of-care processes in the matching because, when we completed matching (spring 2015), these measures were not yet available.
4. We narrowed the pool to 7,376 potential comparison practices by excluding those practices that (1) were participating in one of the three other federal primary care initiatives that were operating in the 15 TransforMED market areas (the Multi-Payer Advanced Primary Care Practice [MAPCP] Demonstration, the Comprehensive Primary Care [CPC] Initiative, and the Federally Qualified Health Center [FQHC] Demonstration). The exception to this is in Michigan, where both of the treatment practices are participating in MAPCP; we did not

exclude practices participating in the MAPCP initiative from the potential comparison pool in Michigan; (2) were owned by one of the 15 participating health systems; (3) were recruited by TransforMED during a second phase of its HCIA program (not included in this report because we did not evaluate its impacts) to expand the PCMN program's reach by 18 to 22 additional practices within each community, see Keith et al. (2015) for detail on this component; (4) had an average of fewer than 25 assigned Medicare FFS beneficiaries during the four baseline quarters; and (5) had a practice size of 100 or more total providers.

5. We used propensity-score methods to select comparison practices (from the pool of 7,376) that were most similar to the 87 treatment practices on the matching variables. The propensity score for a given practice is the predicted probability, based on all matching variables, that the practice is part of the treatment group (Stuart 2010). The score collapses information from all of the matching variables into a single number for each practice that we used to assess how similar practices were to one another. We matched each treatment practice to one or more comparison practices with similar propensity scores, with the aim of generating a comparison group that was similar, on average, to the treatment group on the matching variables. The approach, however, did not ensure that each comparison practice matched exactly to its treatment practice on all matching variables.

   We ran two separate propensity-score matching models—matching health centers separately from nonhealth centers because the variables available for matching these two groups differed slightly. Within each propensity-score model for matching, we further required that a treatment practice could match only to a comparison practice located in the same market area. We required each treatment practice to match to at least one, but no more than five, comparison practices and that the ratio of comparison to treatment practices be at least 3:1. This matching ratio increased the statistical certainty in the impact estimates (relative to a 1:1 overall matching ratio), because it created a more stable comparison group against which the treatment group's experiences could be compared.

6. After completing the matching, we reviewed the list of selected comparison practices and removed any that seemed qualitatively unlike the target practices for the HCIA intervention—that is, Indian Health Services and walk-in clinics—as well as four practices that appeared to have closed or merged with other practices during the intervention period.

After completing the matching, we assigned Medicare FFS beneficiaries to the 286 comparison practices in each intervention quarter using the same rules we used for the treatment group (Section V.A.2). We also defined a high-risk subgroup of the comparison group using the same rules as for the treatment group. That is, a beneficiary was in the high-risk group in the intervention quarter if his or her HCC score at the start of the intervention period was in the top third among all observable Medicare beneficiaries assigned to the treatment practices at the start of the intervention period.

4. Construction of outcomes and covariates

We used Medicare claims from January 1, 2009, to June 30, 2015, for beneficiaries assigned to the treatment and comparison practices to construct two types of variables: (1) outcomes, defined for each person in each baseline or intervention quarter; and (2) covariates, which describe a beneficiary's characteristics at the start of the baseline and intervention periods and

are used in the regression models for estimating impacts to adjust for beneficiaries' characteristics before the period began. We used covariates defined at the start of each period, without updating them each quarter, to avoid controlling in each intervention quarter for previous quarters' program effects, as this would bias the effect estimates away from detecting true impacts. For example, the intervention could result in greater contact with the health system and earlier diagnoses of diseases and conditions, which could affect both health-related characteristics and outcomes. If we adjusted for changes in health-related status during the intervention period, we might adjust away part of the impact of the intervention. Appendix 1 provides details on the methods we used to construct both outcome and covariate variables.

Outcomes. For each person, we calculated six outcomes that we grouped into three domains:

1. Domain: Quality-of-care processes
    a. Diabetes quality-of-care composite (binary variable for each beneficiary); calculated as whether a beneficiary with diabetes had had all four recommended tests—lipid profile, hemoglobin A1c test, dilated eye exam, and nephropathy screening—during the previous 12 months
    b. IVD lipid profile (binary variable for each beneficiary); calculated as whether a beneficiary with IVD had a complete lipid profile during the previous 12 months
    c. Ambulatory-care follow-up visit within 14 days of a hospital discharge (binary variable for each beneficiary); calculated as whether all of an individual's discharges in a quarter were followed by an ambulatory visit with a primary care or specialist physician within 14 days of the discharge
2. Domain: Service use
    a. All-cause inpatient admissions (number/quarter)
    b. Outpatient ED visit rate (number/quarter); outpatient ED visits are defined as ED visits or observational stays that do not end in a hospital admission
3. Domain: Spending
    a. Total Medicare Part A and B spending ($/month)

Three of these outcomes—the three in the service use and spending domains—are outcomes that CMMI has specified as core for the evaluations of all HCIA programs. One additional core outcome—the number of unplanned inpatient readmissions within 30 days—was assessed in our quarterly reporting to CMMI, but is not included here because the awardee did not explicitly expect to affect this outcome.

All outcomes are quarter-specific—meaning that we calculated them for each baseline and intervention quarter separately—except for the two quality-of-care process measures for IVD and diabetes. Because these two measures assess whether a beneficiary received recommended preventive care services over a year-long period, we calculated these measures over full years rather than quarters. For example, over the baseline year (that is, the period corresponding to the

four baseline quarters), over the first year of the intervention period (corresponding to intervention quarters 1 through 4), and over the second year of the intervention period (corresponding to intervention quarters 5 through 8). We avoided calculating these measures for overlapping periods, meaning that no measurement year included services provided in another measurement year.

Finally, we defined all outcomes except for the measures of three quality-of-care processes for all treatment and comparison group members. We calculated the measure of 14-day follow-up after discharge among only those beneficiaries with at least one hospital discharge in the relevant quarter. We calculated the diabetes composite measure among beneficiaries ages 18 to 75 with diabetes at the beginning of the period (baseline or intervention period), and calculated the measure of lipid screening among beneficiaries ages 18 or older with IVD at the beginning of the period.

Covariates. The covariates included (1) 18 indicators for whether a beneficiary had each of the following chronic conditions: Alzheimer's and related dementia, asthma, atrial fibrillation, bipolar disorder, cancer, chronic kidney disease, chronic obstructive pulmonary disease, depression, diabetes, heart failure, hip fracture, hyperlipidemia, hypertension, ischemic heart disease, osteoporosis, rheumatoid arthritis, schizophrenia, and stroke; (2) HCC score; (3) demographics (age, gender, and race or ethnicity); (5) whether the beneficiary is dually eligible for Medicare and Medicaid; (6) whether the beneficiary is a member of the high-risk subgroup, and (7) original reason for Medicare entitlement (old age, disability, or end-stage renal disease). We defined all covariates as of the start of the relevant period (baseline or intervention).

5. Regression model

We used a regression model to implement the difference-in-differences design for estimating impacts. For each outcome, the model estimated the relationship between the outcome and a series of predictor variables, assuming that each of the predictor variables had a linear (additive) relationship with the outcome. The predictor variables included the beneficiary-level covariates (defined in Section V.A.4); whether the beneficiary was assigned to a treatment or a comparison practice; an indicator for each practice (which accounted for differences between practices in their patients' outcomes at baseline); indicators for each post-intervention quarter (or, for the diabetes and IVD measures, for the final post-intervention quarter of the year-long measurement period); and an interaction of a beneficiary's treatment status with each post-intervention quarter (or, for the diabetes and IVD measures, the final post-intervention quarter of the year-long measurement period).

The estimated relationship between the interaction term and the outcome in a given quarter was the impact estimate for that quarter (or, for the diabetes and IVD measures, for the year ending with that quarter). It measured the average difference between outcomes for beneficiaries assigned to the treatment and comparison practices during that period, subtracting out any differences between these groups during the four baseline quarters. By providing separate impact estimates for each intervention quarter (or year, for the diabetes and IVD measures), the model enabled the program's impacts to change the longer the practices were enrolled in the program. We could also test impacts over discrete sets of quarters or years, which was needed to

implement the primary tests discussed in the next section. Finally, the model quantified the uncertainty in the impact estimates, allowing for statistical tests that determined whether observed differences in outcomes between the treatment and comparison groups were likely due to chance. The model used robust standard errors to account for clustering of outcomes across quarters for the same beneficiary and a dummy variable for each practices (fixed effects) to account for clustering of outcomes for beneficiaries assigned to the same practice. Appendix 2 provides details on the regression methods, including descriptions of the analytic weights each beneficiary received in the model.

6. Primary tests

Table V.1 shows the primary tests for TransforMED, by domain. Each test specified a population, outcome, time period, expected direction of effect, and threshold that we counted as substantively important. The purpose of these primary tests was to focus the impact evaluation on hypotheses that would provide the most robust evidence about program effectiveness (see Appendix 3 for detail and a description of how we selected each test). We provided both the awardee and CMMI an opportunity to comment on the primary tests.

Our rationale for selecting these primary tests follows:

- **Outcomes.** TransforMED's central goal was to decrease total Medicare and Medicaid spending by 4 percent by Year 3 of the program. For this reason, we chose to analyze impacts on Medicare Part A and B spending. (We do not have Medicaid data for this evaluation.) In addition, reductions in hospitalizations and ED visits were identified as primary drivers that would enable these spending reductions. Therefore, we selected primary tests examining hospitalizations and ED visits. Finally, we included three quality-of-care process measures that, based on TransforMED's theory of action, we thought the program could improve: (1) a composite measure for whether a beneficiary with diabetes received all of four recommended processes of care during the year, (2) receipt of a complete lipid profile for people with IVD, and (3) receipt of a follow-up ambulatory care visit with a primary care or specialist provider within 14 days of hospital discharge.
- **Time period.** TransforMED expected to have measurable impacts on spending by the third year of the program; few impacts were expected in the first two years of practice participation. Given this, we chose to analyze impacts for spending during the final two quarters of the program's operation (that is, intervention quarters 9 and 10), as these fell during the third year of the program, following two complete years of program operation.

  We chose to analyze impacts for hospitalizations, ED visits, and the three quality-of-care process measures over an earlier and longer period because reductions in these outcomes were expected to occur earlier, as practices began using the cost and population management data to better manage their patients' care. For the service use outcomes, we analyzed impacts in the final year of the practices' participation (that is, intervention quarters 7 through 10). For the three quality-of-care process measures, our primary tests covered the second year of program operation (that is, the 12 months corresponding to quarters 5 through 8) for the IVD and diabetes measures, and the second and third year of program operations (quarters 5 through 10) for receipt of a follow-up ambulatory care visit.

Table V.1. Specification of the primary tests for TransforMED

Domain (number of tests in the domain)[a]	Outcome (units)	Time period for impacts (controlling for baseline differences)[b, c]	Population	Substantive threshold (expected direction of effect)[d]
Quality-of-care process (3)	Received all four recommended diabetes processes of care in the year (binary [yes or no]/beneficiary/year)	The one-year period from January 2014 through December 2014	Medicare FFS beneficiaries with diabetes and ages 18 to 75 attributed to treatment practices	15.0% (+)
	Received complete lipid profile in the year (binary [yes or no]/beneficiary/year)	The one-year period from January 2014 through December 2014	Medicare FFS beneficiaries with IVD, ages 18 or older, and attributed to treatment practices	15.0% (+)
	All inpatient admissions within a quarter were followed by an ambulatory care visit with a primary care or specialist provider within 14 days (binary [yes or no]/beneficiary/year)	Average over intervention quarters 5 through 10	Medicare FFS beneficiaries with at least one hospital stay in the quarter attributed to treatment practices	15.0% (+)
Quality-of-care outcomes (0)[e]	n.a.—Awardee does not explicitly plan to affect quality-of-care outcomes	n.a.	n.a.	n.a.
Service use (4)	All-cause inpatient admissions (#/beneficiary/quarter)	Average over intervention quarters 7 through 10	All Medicare FFS beneficiaries attributed to treatment practices	5.0% (-)
	Outpatient ED visit rate (#/beneficiary/quarter)	Average over intervention quarters 7 through 10	All Medicare FFS beneficiaries attributed to treatment practices	5.0% (-)
	All-cause inpatient admissions (#/beneficiary/quarter)	Average over intervention quarters 7 through 10	Medicare FFS high-risk beneficiaries attributed to treatment practices	15.0% (-)
	Outpatient ED visit rate (#/beneficiary/quarter)	Average over intervention quarters 7 through 10	Medicare FFS high-risk beneficiaries attributed to treatment practices	15.0% (-)
Spending (2)	Medicare Part A and B spending ($/beneficiary/month)	Average over intervention quarters 9 through 10	All Medicare FFS beneficiaries attributed to treatment practices	3.0% (-)
	Medicare Part A and B spending ($/beneficiary/month)	Average over intervention quarters 9 through 10	Medicare FFS high-risk beneficiaries attributed to treatment practices	15.0% (-)

Note: High-risk beneficiaries are defined as beneficiaries with a Hierarchical Condition Category score in the top quarter among all beneficiaries seen by treatment practices at the start of the period (baseline or intervention), by market area.

Table V.1 *(continued)*

[a] We adjusted the *p*-values from the primary test results for the multiple comparisons made within each domain, but not across domains.

[b] The regression models for estimating program impacts control for differences in outcomes between the pre-intervention treatment and comparison groups.

[c] For all but the diabetes and IVD quality-of-care process measures, we will take the average across the relevant quarterly impact estimates (one for each intervention quarter from 5 through 10, 7 through 10, or 9 through 10, respectively). For the diabetes and IVD measures, which are defined annually, we will take the impact estimates for the 12-month period from January through December 2014. This period corresponds to the second year of program operation.

[d] The counterfactual is the outcome the treatment group would have had in the absence of the HCIA-funded intervention.

[e] The quality-of-care outcome measures we can evaluate are (1) 30-day unplanned hospital readmissions and (2) inpatient admissions for ambulatory care-sensitive conditions.

ED = emergency department; FFS = fee-for-service; HCIA = Health Care Innovation Award; IVD = ischemic vascular disease.

n.a. = not applicable.

- **Population.** For hospitalizations, ED visits, and Medicare Part A and B spending, we chose two separate primary test populations: (1) all Medicare FFS beneficiaries attributed to the treatment practices and (2) a high-risk subset of Medicare beneficiaries. The former was the most inclusive definition possible and corresponded with how TransforMED defined its target population; the intervention sought to affect the care of patients of all risk levels. Although the program did not explicitly target a specific population for services, TransforMED's impacts on the acute care and spending outcomes of interest should have concentrated among high-risk beneficiaries, both because there were more opportunities to reduce acute care for this high-risk population and because beneficiaries in this group were more likely to receive intensive interventions, such as care management services.

  For the diabetes and IVD quality-of-care process measures, we limited the population to beneficiaries ages 18 to 75 with diabetes or those ages 18 or older with IVD, respectively. For the 14-day follow-up measure, we limited the sample in each quarter to those who had at least one index hospitalization during the quarter for which we could observe whether the person had a 14-day follow-up visit. We did not define these outcomes separately for the high-risk subgroup because there was no indication from TransforMED's theory of action that the program would improve these differently from those of the full population.

- **Direction (sign) of the impact estimate.** For the quality-of-care process measures, we expected the impact estimate to be positive, signaling an increase in the percentage of people receiving recommended care. For all other outcomes, we expected the impact estimates to be negative, indicating a reduction in service use or overall spending.

- **Substantive thresholds.** Some impact estimates could be large enough to be substantively interesting (to CMMI and other stakeholders) even if they were not statistically significant and, for this reason, we have specified thresholds for what we call substantive importance. For the full population, the 3 percent threshold we chose for substantive importance on spending was 75 percent of TransforMED's expected effect on this outcome in the third year of practice participation. (We used 75 percent recognizing that TransforMED could still be considered successful if it came close to, but did not fully achieve, its anticipated effects.) The awardee did not specify anticipated impacts on the intermediate outcomes of hospitalizations, ED visits, or the three quality-of-care process measures or among subpopulations, so all of our other thresholds were instead taken from the literature (Peikes et al. 2011; Rosenthal et al. 2016). Thresholds for the service use and spending outcomes were based on the assumption that a successful primary care intervention could cause a reduction in spending or service use of 5 percent among a general population and 15 percent among a high-risk population; the 15 percent threshold for the quality-of-care process measures was likewise extrapolated from the literature.

7. Secondary tests

We also conducted secondary tests to help corroborate the findings from the primary tests. This was important because some of the differences observed between the treatment and comparison groups in the primary test results could have been due to the non-experimental design of our study or random fluctuations in the data. We will have greater confidence in the

primary results if they are generally consistent with the expected broader pattern of results. Specifically, we estimated the program's impacts on hospitalizations and spending for the full population during four additional intervention quarters—that is, the first 12 months of program operation (intervention quarters 1 through 4). Because we and TransforMED expected program impacts to increase over time, the following pattern would be highly consistent with an effective program—largest impacts in the last quarters of program operations (that is, the time period for the primary tests), and smaller impacts during intervention quarters 1 through 4. In contrast, if we found larger differences in outcomes (favorable or unfavorable) in the first year of the program than in the last quarters, this could suggest a limitation in the comparison group, not true program impacts.

8. Synthesizing evidence to draw conclusions

Within each domain, we drew one of five conclusions about program effectiveness, based on the primary test results, the results of secondary tests, and the plausibility of those findings given the implementation evidence:

1. Statistically significant favorable effect (the highest level of evidence)
2. Substantively important (but not statistically significant) favorable effect
3. Substantively important (but not statistically significant) unfavorable effect
4. No substantively large effect
5. Indeterminate effect

We cannot conclude that a program had a statistically significant unfavorable effect because, in consultation with CMMI, we decided to use one-sided statistical tests (which do not test for evidence of unfavorable effects). We used one-sided tests to increase the probability that, if a program truly did have impacts, we would be able to detect them.

Appendix 3 describes our decision rules for each of the five possible conclusions. In short, we concluded that a program had a statistically significant favorable effect in a domain if (1) at least one primary test result in the domain was favorable and statistically significant, after adjusting the statistical tests to account for multiple tests (if applicable) within a domain; or (2) the average impact estimate across all primary tests in the domain was favorable and statistically significant. In both cases, we also had to determine that the primary test results were plausible given the results of the secondary tests and implementation evidence. We concluded that a program had a substantively important favorable effect if the average impact estimate in the domain was substantively important but not statistically significant, and if the result was plausible given the secondary tests and implementation evidence. In contrast, if the average impact estimate was unfavorable (opposite the hypothesized direction), larger than the substantive threshold, and unfavorable effects were plausible given the other evidence, we concluded the program had a substantively important unfavorable effect. If the tests in a domain did not meet any of these criteria, we drew one of two conclusions. First, if the tests for at least one outcome in the domain (or all outcomes in the domain together) had sufficient statistical power to detect an impact of the size of the substantive threshold with at least 75 percent

probability, we concluded there was not a substantively large effect because we are reasonably confident that we would have detected such an effect had there been one. Alternatively, if the power was not sufficient (less than 75 percent) to detect this type of impact, we concluded the impact in the domain was indeterminate. Indeterminate means either that the program truly did not have effects that were substantively large or that it did, but our statistical tests were not able to detect them.

B. Characteristics of the treatment group at baseline

This section describes the characteristics of the treatment group at the start of the intervention (January 1, 2013). We also show this information in the second column of Table V.2. That table serves a second purpose—to show the equivalence of the treatment and comparison practices at the start of the intervention—which we describe in Section V.C. For benchmarking purposes, the last column shows the values of relevant variables for the national Medicare FFS population, when available.

Characteristics of the practices overall. Our analysis includes 87 treatment practices at the start of the intervention, 10 of which are FQHCs or RHCs. The intervention was implemented in 15 states, with Georgia and Indiana sharing the most practices (8 each) and Michigan the fewest (2 practices). Underscoring the variation in the areas where TransforMED implemented the intervention, although the vast majority of practices were located in an urban zip code (79 percent), 12 percent were located in areas designated as health professional shortage areas for primary care.

Additional practice characteristics were available only for the 77 nonhealth centers. A hospital or health system owned most treatment practices (84 percent) and almost all treatment practices had providers receiving payment from the Centers for Medicare & Medicaid Services (CMS) for using EHRs in a meaningful way (94 percent). This latter proportion was consistent with TransforMED's target population, as one of the program's eligibility criteria was an EHR system that had been actively used among practice staff for at least a year. Treatment practices, on average, had 6.6 total providers and most providers in these practices had primary care as their specialty (90 percent).

Characteristics of the practices' Medicare FFS beneficiaries. The Medicare FFS beneficiaries assigned to the treatment practices during the baseline period (January 1, 2012, through December 31, 2012) were, overall, similar to the nationwide FFS population. The HCC risk score for the treatment group was close with the national average (1.1 versus 1.0). Participants in the treatment practices also had hospital admission rates and total Medicare spending that were close to the national averages. The mean outpatient ED visit rate (138/1,000 beneficiaries/quarter) was higher than the national average of 105. The high-risk beneficiaries in the treatment group had substantially greater health care needs during the baseline period than the full treatment group. For example, the mean HCC risk score in this group was more than twice the mean for all treatment group members (2.3 versus 1.1), consistent with how the group was defined.

Table V.2. Characteristics of treatment and comparison practices when the intervention began (January 1, 2013)

Characteristic of practice	Treatment practices (N = 87)	Matched comparison group (N = 286)	Absolute difference[a]	Standardized difference[b]	Medicare FFS national average
Exact match variables[c]					
Health center (%)	11.5	11.5	0.00	0.00	n.a.
Market area (%)					
Alabama	6.9	6.9	0.00	0.00	n.a.
Connecticut	6.9	6.9	0.00	0.00	n.a.
Florida	5.8	5.8	0.00	0.00	n.a.
Georgia	9.2	9.2	0.00	0.00	n.a.
Indiana	9.2	9.2	0.00	0.00	n.a.
Kansas	6.9	6.9	0.00	0.00	n.a.
Kentucky/Indiana	4.6	4.6	0.00	0.00	n.a.
Maryland	8.1	8.1	0.00	0.00	n.a.
Massachusetts	8.1	8.1	0.00	0.00	n.a.
Michigan	2.3	2.3	0.00	0.00	n.a.
Mississippi	8.1	8.1	0.00	0.00	n.a.
Nebraska	6.9	6.9	0.00	0.00	n.a.
North Carolina	6.9	6.9	0.00	0.00	n.a.
Oklahoma	5.8	5.8	0.00	0.00	n.a.
West Virginia	4.6	4.6	0.00	0.00	n.a.
Propensity-matched variables[d]					
Characteristics of a practice's location(s)					
Located in an urban zip code (%)	79.3	76.5	2.85	0.07	80.7[e]
Medicare Advantage penetration rate (2011) (%)	17.5	17.9	-0.47	-0.05	NA
Located in a health professionals shortage area (primary care) (2011) (%)	11.5	15.7	-4.16	-0.12	NA
Characteristics of all patients attributed to practices during the baseline year (January 1, 2012–December 31, 2012)					
Number of beneficiaries	957.5	1,036.0	-78.5	-0.11	n.a.
HCC risk score	1.12	1.12	0.00	0.01	1.0
All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	79.46	80.47	-1.01	-0.05	74[f]
Outpatient ED visit rate (#/1,000 beneficiaries/quarter)	137.89	133.33	4.56	0.08	105[g]
Medicare Part A and B spending ($/beneficiary/month)	845	845	0	0	860[h]
30-day unplanned hospital readmissions (#/beneficiary/quarter)	11.02	10.96	0.06	0.01	NA
Age as original reason for Medicare entitlement (%)	74.7	75.0	-0.3	-0.02	83.3[i]
Disability as original reason for Medicare entitlement (%)	25.1	24.9	0.3	-0.02	16.7[i]
ESRD as original reason for Medicare entitlement (%)	0.1	0.1	0.0	0.03	0.1[i]

Table V.2 *(continued)*

Characteristic of practice	Treatment practices (N = 87)	Matched comparison group (N = 286)	Absolute difference[a]	Standardized difference[b]	Medicare FFS national average
Percentage of the practice's patients dually eligible for Medicaid	19.8	19.3	0.5	0.03	21.7[j]
Age (years)					71[k]
Younger than 50 (%)	6.9	6.3	0.7	0.12	16.7[i]
50–64 (%)	10.9	10.8	0.2	0.02	
65–74 (%)	43.0	42.7	0.2	0.03	45.5[i]
75–84 (%)	27.5	28.1	-0.6	-0.09	25.4[i]
85 or older (%)	11.7	12.1	-0.4	-0.07	12.4[i]
Female (%)	59.4	58.7	0.7	0.12	54.7[i]
Characteristics of high-risk patients attributed to practices during the baseline year (January 1, 2012–December 31, 2012)					
Number of high-risk beneficiaries	225.5	244.4	-19.0	-0.10	n.a.
HCC risk score	2.32	2.35	-0.03	-0.13	1.0
All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	172.61	174.44	-1.82	-0.04	74[f]
Outpatient ED visit rate (#/1,000 beneficiaries/quarter)	235.29	230.83	4.46	0.05	105[h]
Medicare Part A and B spending ($/beneficiary/month)	1,738	1,737	1	0.00	860[h]
Characteristics of the practices (nonhealth centers only)					
Meaningful use of EHR (%)	93.5	90.5	3.05	0.10	n.a.
Ownership: owned by hospital or health system (%)	84.4	80.0	4.45	0.10	n.a.
Number of clinicians at practice [l]	6.6	7.5	-0.88	-0.14	n.a.
Has 1 clinician (%)	2.6	3.8	-1.23	-0.06	n.a.
Has 2 or 3 clinicians (%)	18.2	21.0	-2.86	-0.07	n.a.
Has 4 or 5 clinicians (%)	18.2	17.0	1.19	0.03	n.a.
Has 6 to 14 clinicians (%)	52.0	50.2	1.71	0.03	n.a.
Has 15 or more clinicians (%)	9.1	7.9	1.19	0.05	n.a.
Percentage of practices' clinicians with primary care specialty	89.8	89.2	0.55	0.03	n.a.
Variables not included in matching[m]					
Characteristics of all patients attributed to practices during the baseline year (January 1, 2012–December 31, 2012)					
Diabetes processes of care					
FFS beneficiaries meeting inclusion criteria (%)	14.4	14.1	0.3	0.06	NA
Received all recommended diabetes care (%)	39.2	37.2	2.0	0.14	NA
Lipid testing for those with IVD					
FFS beneficiaries meeting inclusion criteria (%)	25.4	26.5	-1.1	-0.18	NA
Received lipid test (%)	75.4	74.8	0.6	0.05	NA
Ambulatory care visit within 14 days of discharge					
FFS beneficiaries meeting inclusion criteria (%)	6.2	6.3	-0.1	-0.10	NA
Received visits after all discharges (%)	58.1	56.8	1.2	0.13	NA

Table V.2 *(continued)*

Characteristic of practice	Treatment practices (N = 87)	Matched comparison group (N = 286)	Absolute difference[a]	Standardized difference[b]	Medicare FFS national average
Characteristics of high-risk patients attributed to practices during the baseline year (January 1, 2012–December 31, 2012)					
Diabetes processes of care					
FFS beneficiaries meeting inclusion criteria (%)	21.4	20.6	0.8	0.10	NA
Received all recommended diabetes care (%)	38.8	36.6	2.2	0.13	NA
Lipid testing for those with IVD					
FFS beneficiaries meeting inclusion criteria (%)	50.0	51.6	-1.6	-0.18	NA
Received lipid test (%)	71.6	70.3	1.3	0.09	NA
Ambulatory care visit within 14 days of discharge					
FFS beneficiaries meeting inclusion criteria (%)	12.8	13.0	-0.2	-0.06	NA
Received visits after all discharges (%)	60.6	58.0	2.6	0.21	NA

Sources: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at CMS. Zip code household income data merged from the American Community Survey ZIP Code Characteristics.

Notes: The comparison group means were weighted based on the number of matched comparisons per treatment beneficiary. For example, if four comparison practices were matched to one treatment practice, each of the four comparison practices had a matching weight of 0.25.

High-risk beneficiaries were defined as beneficiaries with an HCC score in the top quartile among all beneficiaries seen by treatment practices during the period (baseline or intervention), by market area.

Absolute differences might not be exact due to rounding.

[a] The absolute difference is the difference in means between the treatment and matched comparison groups.

[b] The standardized difference is the difference in means between the matched treatment and comparison groups divided by the standard deviation of the variable. The standard deviation is calculated among the pooled treatment and matched comparison groups.

[c] Variables for which we required treatment and comparison members to match on exactly. For example, a treatment practice that was a health center could be matched only to a comparison practice that was a health center, and each treatment practice could match only to comparison practices in the same market area.

[d] Variables that we matched on through a propensity score, which captures the relationship between a practice's characteristics and its likelihood of being in the treatment group.

[e] U.S. Census Bureau, 2010 Census, urban and rural areas.

[f] Health Indicators Warehouse (2014b).

[g] Gerhardt et al. (2014).

[h] Boards of Trustees (2013).

[i] Chronic Conditions Warehouse (2014a, Table A.1).

[j] Health Indicators Warehouse (2014c).

[k] Health Indicators Warehouse (2014a).

[l] Clinicians include physicians, nurse practitioners, and physician assistants.

[m] These baseline process of care measures were not available at the time we conducted matching.

CMS = Centers for Medicare & Medicaid Services; ED = emergency department; EHR = electronic health record; ESRD = end-stage renal disease; FFS = fee-for-service; HCC = Hierarchical Condition Category.
NA = not available.

n.a. = not applicable.

C. Equivalence of the treatment and comparison groups at the start of the intervention

Demonstrating that the treatment and comparison groups were similar at the start of the intervention is critical for the evaluation design. This similarity increases the credibility of a key assumption underlying difference-in-differences models—that the change over time in outcomes for the comparison group is the same change that would have happened for the treatment group, had the treatment practices not received the intervention.

Table V.2 shows that the 87 treatment practices and the 286 selected comparison practices were similar at the start of the intervention on variables used in matching. By construction, there were no differences between the two groups on the market area in which practices were located. The treatment and matched comparison group beneficiaries differed somewhat on the variables we matched through propensity scores, but the standardized differences across the propensity-score matching variables were within our target of 0.25 standardized differences, and nearly all were actually within 0.15 standardized differences (the 0.25 target is an industry standard; see Institute of Education Sciences 2014).

D. Beneficiaries' outcomes and intervention impacts

In this section, we first present sample sizes and mean outcomes, by quarter, for the treatment and comparison groups. These mean outcomes provide context for interpreting the difference-in-differences estimates that follow; however, the differences in mean outcomes are not regression-adjusted and not impact estimates by themselves. Next, we present the results of the primary tests, by domain. Then, we present the results of the secondary tests (robustness checks) and assess whether the primary test results are plausible given the secondary test results and the implementation evidence. We end with conclusions about program impacts in each domain.

1. Sample sizes

The sample sizes for impact estimation differ depending on the outcome. We present sample sizes by domain.

Quality-of-care processes (Table V.3)

- The **diabetes preventive care composite measure** was defined among Medicare FFS beneficiaries with diabetes ages 18 to 75. The sample size for the treatment group and the weighted comparison group ranged from 12,119 to 14,362 across the baseline year and each of the two intervention years for which the outcome was measured. This population accounted for 12 to 15 percent of the total Medicare FFS sample in both the treatment and comparison groups.

Table V.3. Sample sizes and unadjusted means for Medicare FFS beneficiaries in the treatment and comparison groups for TransforMED, by quarter (quality-of-care processes domain)

		Number of Medicare FFS beneficiaries (practices)			Mean outcomes		
Period	Quarter	T	C (not weighted)	C (weighted)	T	C	Difference (%)
Among those with diabetes and ages 18 to 75 , received all four recommended diabetes processes of care in the year (binary [yes or no]/beneficiary/year)							
Baseline	B1–B4[a]	12,119 (87)	40,621 (286)	12,728	40.1	36.1	4.1 (11.3%)
Intervention	I1–I4[a]	13,472 (87)	42,353 (285)	14,362	43.8	39.2	4.6 (11.7%)
	I5–I8[a]	12,218 (87)	38,085 (285)	12,771	45.1	39.4	5.6 (14.3%)
Among those with ischemic vascular disease and ages 18 or older, received complete lipid profile in the year (binary [yes or no]/beneficiary/year)							
Baseline	B1–B4[a]	22,238 (87)	76,920 (286)	23,863	75.8	75.1	0.7 (0.9%)
Intervention	I1–I4[a]	24,698 (87)	78,092 (286)	27,385	77.4	76.2	1.2 (1.6%)
	I5–I8[a]	23,196 (87)	72,840 (286)	25,639	76.5	74.1	2.4 (3.2%)
Among those with at least one inpatient admission in the quarter, all inpatient admissions in the quarter were followed by an ambulatory care visit with a primary care or specialist provider within 14 days of discharge (binary [yes or no]/beneficiary/year)							
Baseline	B1	5,127 (87)	17,614 (285)	5,281	58.0	56.3	1.8 (3.1%)
	B2	5,028 (87)	16,998 (283)	4,964	58.8	56.7	2.1 (3.7%)
	B2	5,132 (87)	17,179 (283)	5,145	58.9	57.8	1.1 (2.0%)
	B4	5,417 (87)	18,012 (284)	5,388	57.0	54.4	2.5 (4.7%)
Intervention	I1	5,429 (87)	17,821 (284)	5,672	59.6	56.2	3.4 (6.0%)
	I2	5,412 (87)	17,183 (282)	5,438	60.9	58.2	2.7 (4.7%)
	I3	5,308 (87)	17,096 (285)	5,852	61.4	58.8	2.7 (4.5%)
	I4	5,466 (87)	17,177 (284)	5,639	58.9	59.3	-0.3 (-0.5%)
	I5	5,709 (87)	17,849 (285)	5,787	61.4	56.2	5.2 (9.2%)
	I6	5,664 (87)	17,716 (286)	6,672	61.9	61.4	0.5 (0.8%)
	I7	5,493 (87)	17,578 (285)	6,120	63.1	60.3	2.8 (4.7%)
	I8	5,789 (87)	18,280 (286)	6,353	61.2	58.0	3.2 (5.5%)
	I9	6,144 (87)	19,328 (284)	7,035	62.0	50.1	11.8 (23.6%)
	I10	6,076 (87)	18,539 (283)	7,171	62.3	61.4	0.9 (1.4%)

Table V.3 *(continued)*

Sources: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at the Centers for Medicare & Medicaid Services.

Notes: The baseline quarters are measured relative to when the baseline period began on January 1, 2012. For example, the first baseline quarter (B1) ran from January 1, 2012, to March 31, 2012. The intervention quarters are measured relative to the start of the intervention period on January 1, 2013. For example, the first intervention quarter (I1) ran from January 1, 2013, to March 31, 2013. In each period (baseline or intervention), the treatment group in each quarter included all beneficiaries attributed to a treatment practice by the start of the quarter and enrolled in FFS Medicare. In each period, the comparison group in each quarter included all beneficiaries attributed to a comparison practice by the start of the quarter and who met the other sample criteria. See text for details.

The outcome means were weighted, such that (1) each treatment beneficiary gets a weight of 1; and (2) each comparison beneficiary received a weight that was the product of two weights: (a) a matching weight, equal to the reciprocal of the total number of comparison practices matched to the same treatment practice as the beneficiary's assigned practice; and (b) a practice size weight, which equaled the average number of beneficiaries assigned to the matched treatment practice during the four baseline quarters divided by the average number of beneficiaries assigned to the beneficiary's comparison practice over those quarters. The difference between the treatment and comparison groups in a quarter was calculated by subtracting the mean outcome for the comparison group from the mean outcome for the treatment group. The percentage difference equals that difference divided by the mean outcome for the comparison group.

[a] The quality-of-care process measures for diabetes and ischemic vascular disease were calculated over year-long periods, corresponding to the baseline and intervention quarters shown in the table.

- The **lipid profile measure for people with IVD** was defined among Medicare FFS beneficiaries with IVD ages 18 or older. The sample size for the treatment group and the weighted comparison group ranged from 22,238 to 27,385 across the baseline year and each of the two intervention years for which the outcome was measured. This population accounted for about 22 to 28 percent of the total Medicare FFS sample in the treatment and comparison groups. This percentage was higher than for the diabetes measure because (1) IVD (which is a broad disease category) was more common than diabetes among the treatment and comparison beneficiaries, and (2) the diabetes measure excluded beneficiaries older than 75, unlike the IVD measure.
- The **14-day follow-up measure** was defined among Medicare FFS beneficiaries who had at least one hospital stay in the quarter. The sample size for the treatment group and the weighted comparison group ranged from 4,964 to 7,171 across the baseline and intervention quarters. This population accounted for 6 to 7 percent of the total Medicare FFS sample in the treatment and comparison groups.

Service use and spending (all beneficiaries). The sample sizes for all outcomes in these two domains were the same. In the first baseline quarter (B1), the treatment group included 79,042 beneficiaries assigned to the 87 participating practices and the comparison group includes 262,501 beneficiaries (77,655 weighted beneficiaries) assigned to the 286 comparison practices (Table V.4a). The sample sizes increased modestly during the four baseline quarters (by about 10 percent from the first to the last baseline quarter). This net increase indicated that sample addition (due to beneficiaries being newly attributed to the treatment or comparison practices) exceeded sample attrition (due to beneficiaries dying, switching from FFS Medicare to managed care, or moving out of state). The sample sizes dropped modestly from the last baseline quarter to the first intervention quarter (I1), reflecting that the sample definition (Section V.A.2) retained

sample members in successive baseline and intervention quarters, even if they were no longer attributed to the treatment or comparison practice, but not between the baseline and intervention periods. The sample increased modestly during the intervention period, again reflecting greater sample addition than attrition over time. The net sample increase during the intervention period was slightly larger for the treatment group (14.7 percent from the first intervention to the last [10th] intervention quarter) than the comparison group (11.7 percent over the same time period).

Service use and spending (high-risk beneficiaries). As with the full beneficiary sample, sample sizes were the same for all outcomes in these two domains. In the first baseline quarter (B1), the treatment group included 19,870 high-risk beneficiaries assigned to the 87 participating practices and the comparison group included 67,676 high-risk beneficiaries (19,346 weighted beneficiaries) assigned to the 286 comparison practices (Table V.4b). The sample sizes modestly decreased during the during the four baseline quarters (by about 3 percent from the first to the last baseline quarter) for both the treatment and comparison group. This small decrease indicates that sample attrition slightly exceeded sample addition for the high-risk sample. The sample sizes increased modestly from the last baseline quarter to the first intervention quarter, and then decreased again during the intervention period, reflecting greater sample attrition than addition.

2. Mean outcomes for the treatment and comparison groups, by domain and quarter

Quality-of-care processes. During the baseline year, 40.1 percent of treatment and 36.1 percent of comparison beneficiaries with diabetes, ages 18 to 75, received all four recommended processes of care. These rates increased slightly to 45.1 percent in the second program year for the treatment group and to 39.4 percent for the comparison group.

During the baseline year, 75.8 and 75.1 percent of the treatment and comparison beneficiaries, respectively, ages 18 or older with IVD received the recommended lipid test. These rates increased to 77.4 and 76.2 percent, respectively, in the first intervention year, but then fell to 76.5 and 74.1 percent in the second intervention year.

During the baseline period, 57.0 to 58.9 percent of treatment group beneficiaries and 54.4 to 57.8 percent of comparison group beneficiaries with any hospital stay in a quarter had all of those stays followed by an ambulatory care visit within 14 days of discharge. These percentages increased modestly during the intervention period, so that by I10 the value was 62.3 percent among the treatment group and 61.4 percent among the comparison group.

Service use. All-cause inpatient admissions fluctuated over time for both the treatment and comparison groups among the full beneficiary sample. In all but two intervention quarters, treatment group beneficiaries had fewer inpatient admissions, with that difference generally increasing over time. For the high-risk subgroup, treatment group beneficiaries had generally more admissions in the first five intervention quarters than comparison group beneficiaries, but fewer admissions in the last five intervention quarters.

Table V.4a. Sample sizes and unadjusted mean outcomes for all Medicare FFS beneficiaries in the treatment and comparison groups for TransforMED, by quarter (service use and spending domains)

	Number of Medicare FFS beneficiaries (practices)			All-cause inpatient admissions (#/1,000 beneficiaries/quarter)			Outpatient ED visit rate (#/1,000 beneficiaries/quarter)			Medicare Part A and B spending ($/beneficiary/month)		
Q	T	C (no wgt)	C (wgt)	T	C	Diff (%)	T	C	Diff (%)	T	C	Diff (%)
Baseline period (January 1, 2012–December 31, 2012)												
B1	79,042 (87)	262,501 (286)	77,655	83.1	87.1	-4.0 (-4.6%)	131.1	138.5	-7.4 (-5.3%)	$832	$837	$-5 (-0.6%)
B2	82,133 (87)	271,797 (286)	81,573	78.5	78.1	0.4 (0.5%)	133.9	137.7	-3.7 (-2.7%)	$851	$859	$-8 (-0.9%)
B3	84,746 (87)	279,556 (286)	85,023	77.1	76.5	0.7 (0.9%)	140.3	146.8	-6.5 (-4.4%)	$853	$844	$8 (1.0%)
B4	87,299 (87)	286,480 (286)	88,969	81.0	77.8	3.2 (4.1%)	134.5	139.6	-5.1 (-3.7%)	$874	$844	$31 (3.6%)
Intervention period (January 1, 2013–June 30, 2015)												
I1	85,448 (87)	275,408 (286)	90,630	79.6	79.8	-0.3 (-0.4%)	125.0	140.6	-15.5 (-11.0%)	$836	$900	$-64 (-7.1%)
I2	88,740 (87)	283,236 (286)	93,610	78.6	74.8	3.8 (5.0%)	133.3	145.7	-12.4 (-8.5%)	$856	$878	$-22 (-2.5%)
I3	91,457 (87)	289,602 (286)	96,038	74.6	79.1	-4.5 (-5.7%)	134.2	150.7	-16.4 (-10.9%)	$841	$876	$-35 (-4.0%)
I4	93,970 (87)	294,712 (286)	97,701	74.2	70.8	3.4 (4.8%)	128.3	138.2	-9.9 (-7.1%)	$853	$860	$-7 (-0.9%)
I5	93,213 (87)	291,663 (286)	96,280	78.4	78.9	-0.6 (-0.7%)	127.1	134.6	-7.5 (-5.6%)	$852	$894	$-42 (-4.7%)
I6	94,913 (87)	296,705 (286)	97,646	75.6	88.9	-13.2 (-14.9%)	134.3	143.8	-9.5 (-6.6%)	$883	$941	$-58 (-6.2%)
I7	96,300 (87)	301,169 (286)	99,730	72.2	79.8	-7.6 (-9.5%)	138.4	155.7	-17.3 (-11.1%)	$860	$1,030	$-170 (-16.5%)
I8	97,636 (87)	305,674 (286)	101,899	76.7	82.4	-5.8 (-7.0%)	132.2	148.3	-16.1 (-10.8%)	$879	$898	$-19 (-2.1%)
I9	97,080 (87)	304,376 (286)	102,457	80.4	88.6	-8.2 (-9.3%)	133.5	147.3	-13.9 (-9.4%)	$874	$920	$-46 (-5.0%)
I10	97,994 (87)	307,529 (286)	103,819	78.1	86.5	-8.5 (-9.8%)	142.0	151.4	-9.5 (-6.2%)	$926	$939	$-13 (-1.4%)

Table V.4a *(continued)*

Sources: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at the Centers for Medicare & Medicaid Services.

Notes: The baseline quarters are measured relative to when the baseline period began on January 1, 2012. For example, the first baseline quarter (B1) ran from January 1, 2012, to March 31, 2012. The intervention quarters are measured relative to the start of the intervention period on January 1, 2013. For example, the first intervention quarter (I1) ran from January 1, 2013, to March 31, 2013.

In each period (baseline or intervention), the treatment group in each quarter included all beneficiaries attributed to a treatment practice by the start of the quarter and enrolled in FFS Medicare. In each period, the comparison group in each quarter included all beneficiaries attributed to a comparison practice by the start of the quarter and who met the other sample criteria. See text for details.

The outcome means were weighted, such that (1) each treatment beneficiary received a weight of 1; and (2) each comparison beneficiary received a weight that was the product of two weights: (a) a matching weight, equal to the reciprocal of the total number of comparison practices matched to the same treatment practice as the beneficiary's assigned practice; and (b) a practice-size weight, which equaled the average number of beneficiaries assigned to the matched treatment practice during the four baseline quarters divided by the average number of beneficiaries assigned to the beneficiary's comparison practice over those quarters. The difference between the treatment and comparison groups in a quarter was calculated by subtracting the mean outcome for the comparison group from the mean outcome for the treatment group. The percentage difference equals that difference divided by the mean outcome for the comparison group.

B = baseline; C = comparison; Diff = difference; ED = emergency department; FFS = fee-for-service; I = intervention; Q = quarter; T = treatment; wgt = weight.

Table V.4b. Sample sizes and unadjusted mean outcomes for Medicare FFS high-risk beneficiaries in the treatment and comparison groups for TransforMED, by quarter (service use and spending domains)

	Number of Medicare FFS beneficiaries (practices)			All-cause inpatient admissions (#/1,000 beneficiaries/quarter)			Outpatient ED visit rate (#/1,000 beneficiaries/quarter)			Medicare Part A and B spending ($/beneficiary/month)		
Q	T	C (no wgt)	C (wgt)	T	C	Diff (%)	T	C	Diff (%)	T	C	Diff (%)
Baseline period (January 1, 2012–December 31, 2012)												
B1	19,870 (87)	67,676 (285)	19,346	189.0	203.8	-14.8 (-7.3%)	234.3	241.4	-7.0 (-2.9%)	$1,816	$1,869	$-53 (-2.8%)
B2	19,746 (87)	66,979 (286)	19,549	167.5	169.2	-1.6 (-1.0%)	236.4	235.9	0.6 (0.2%)	$1,712	$1,768	$-56 (-3.2%)
B3	19,544 (87)	66,261 (286)	19,470	161.3	162.4	-1.1 (-0.7%)	243.5	253.5	-10.0 (-3.9%)	$1,678	$1,694	$-16 (-0.9%)
B4	19,308 (87)	65,390 (286)	20,103	174.7	160.9	13.8 (8.6%)	236.9	244.8	-7.9 (-3.2%)	$1,731	$1,667	$64 (3.8%)
Intervention period (January 1, 2013–June 30, 2015)												
I1	21,673 (87)	69,854 (286)	24,462	185.3	170.4	14.9 (8.7%)	227.5	242.5	-15.1 (-6.2%)	$1,874	$2,010	$-136 (-6.8%)
I2	21,426 (87)	68,625 (286)	24,574	175.3	152.4	23.0 (15.1%)	240.9	254.4	-13.5 (-5.3%)	$1,795	$1,779	$16 (0.9%)
I3	21,219 (87)	67,464 (286)	24,392	164.1	161.0	3.1 (1.9%)	243.9	254.8	-10.9 (-4.3%)	$1,729	$1,699	$31 (1.8%)
I4	20,996 (87)	66,113 (286)	23,926	167.4	144.0	23.4 (16.2%)	233.0	229.0	4.0 (1.7%)	$1,748	$1,685	$63 (3.8%)
I5	20,175 (87)	63,369 (286)	22,918	173.0	169.0	4.1 (2.4%)	225.9	220.8	5.1 (2.3%)	$1,731	$1,913	$-182 (-9.5%)
I6	19,759 (87)	61,808 (286)	22,112	168.1	209.2	-41.1 (-19.6%)	237.6	239.9	-2.3 (-1.0%)	$1,805	$1,975	$-170 (-8.6%)
I7	19,287 (87)	60,460 (286)	21,581	160.0	175.5	-15.5 (-8.8%)	250.2	263.1	-12.9 (-4.9%)	$1,747	$2,298	$-550 (-24.0%)
I8	18,814 (87)	59,230 (286)	21,126	174.9	178.2	-3.3 (-1.8%)	234.6	255.7	-21.0 (-8.2%)	$1,789	$1,754	$35 (2.0%)
I9	18,096 (87)	57,210 (286)	20,822	181.4	193.3	-12.0 (-6.2%)	243.6	238.7	4.9 (2.0%)	$1,809	$1,833	$-25 (-1.3%)
I10	17,511 (87)	55,391 (286)	20,124	173.3	198.9	-25.6 (-12.9%)	256.7	233.6	23.1 (9.9%)	$1,868	$1,874	$-6 (-0.3%)

Table V.4b *(continued)*

Sources: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at the Centers for Medicare & Medicaid Services.

Notes: The baseline quarters are measured relative to when the baseline period began on January 1, 2012. For example, the first baseline quarter (B1) rans from January 1, 2012, to March 31, 2012. The intervention quarters are measured relative to the start of the intervention period on January 1, 2013. For example, the first intervention quarter (I1) rans from January 1, 2013, to March 31, 2013.

In each period (baseline or intervention), the treatment group in each quarter included all beneficiaries attributed to a treatment practice by the start of the quarter and enrolled in FFS Medicare. In each period, the comparison group in each quarter included all beneficiaries attributed to a comparison practice by the start of the quarter and who met the other sample criteria. See text for details.

The outcome means were weighted, such that (1) each treatment beneficiary gets a weight of 1; and (2) each comparison beneficiary gets a weight that is the product of two weights: (a) a matching weight, equal to the reciprocal of the total number of comparison practices matched to the same treatment practice as the beneficiary's assigned practice; and (b) a practice-size weight, which equals the average number of beneficiaries assigned to the matched treatment practice during the four baseline quarters divided by the average number of beneficiaries assigned to the beneficiary's comparison practice over those quarters. The difference between the treatment and comparison groups in a quarter is calculated by subtracting the mean outcome for the comparison group from the mean outcome for the treatment group. The percentage difference equals that difference divided by the mean outcome for the comparison group.

B = baseline; C = comparison; Diff = difference; ED = emergency department; FFS = fee-for-service; I = intervention; Q = quarter; T = treatment; wgt = weight.

Similarly, outpatient ED visit rates fluctuated over time for the full beneficiary sample; however, in all baseline and intervention quarters, treatment group beneficiaries had fewer outpatient ED visits than did comparison group beneficiaries. Differences between the two groups were larger across intervention quarters (5.6 to 11.1 percent) than in baseline quarters (2.7 to 5.3 percent). For the high-risk population, this pattern of lower ED visit rates for the treatment group held for some, but not all intervention quarters. In a number of intervention quarters, high-risk treatment group beneficiaries had higher ED visit rates than high-risk comparison group beneficiaries.

Spending. For the full beneficiary sample, total spending in the treatment group averaged about $869 per beneficiary per month in the intervention quarters, quite similar to the baseline average of $853. In each quarter, this was 1 to 17 percent lower than total spending in the comparison group, which ranged from a low of $860 in I4 to a high of 1,030 in I7. For the high-risk beneficiary sample, Medicare Part A and B spending in the treatment group averaged $1,790 per month over the intervention quarters. Spending in any particular intervention quarter was quite consistent around the average, showing no discernable pattern. For the comparison group, average spending among high-risk beneficiaries was more variable across intervention quarters, ranging from a high of $2,298 in I7 to a low of $1,685 in I4. As a result, the percentage difference between the treatment and comparison group also varied considerably; however, average high-risk spending in the treatment group was lower than that in the comparison group in all but one of the final six intervention quarters.

3. Results for primary tests, by domain

Overview. For two of the study domains—quality-of-care processes and spending—the regression-adjusted differences between the treatment and comparison groups were small (Table V.5). None of these differences were statistically significant or larger than the substantive thresholds in either a favorable or unfavorable direction. In contrast, in the service use domain, we found statistically significant favorable differences.

Quality-of-care processes. The likelihood of receiving recommended processes of care for diabetes or IVD was 1.2 and 1.9 percent higher, respectively, for the treatment group (a favorable estimate) than the estimated counterfactual. (Our estimated counterfactual—the outcome the treatment group members would have had in the absence of the HCIA intervention—is the treatment group mean minus the difference-in-differences estimate.) We do not consider these point estimates to be substantively large because both were smaller than the substantive threshold for these outcomes of 15 percent. Further, these estimates were not statistically significant, with *p*-values of 0.48 and 0.28, respectively.

The likelihood of receiving an ambulatory care visit within 14 days of hospital discharge was 1.3 percent higher in the treatment group than its estimated counterfactual, a (favorable) difference that was neither substantively large nor statistically significant.

Table V.5. Results of primary tests for TransforMED

Primary test definition					Statistical power to detect an effect that is[a]		Results			
Domain (# of tests in domain)	Outcome (units)	Time period for impacts	Population	Substantive threshold (expected direction of effect)[b]	Size of the substantive threshold	Twice the substantive threshold[c]	Treatment group mean	Regression-adjusted difference between the treatment and estimated counterfactual[b] (standard error)	Percentage difference[d]	*p*-value[e]
Quality of care process (3)	Received all four recommended diabetes processes of care in the year (binary [yes or no]/beneficiary/year)	The one-year period January through December 2014	Medicare FFS beneficiaries ages 18 to 75 with diabetes assigned to treatment practices	15.0% (+)	> 99.9%	> 99.9%	45.1	0.5 (1.3)	1.2%	0.48
	Received complete lipid profile in the year (binary [yes or no]/beneficiary/year)	The one-year period January through December 2014	Medicare FFS beneficiaries ages 18 or older with ischemic vascular disease assigned to treatment practices	15.0% (+)	> 99.9%	> 99.9%	76.5	1.4 (1.2)	1.9%	0.28
	All inpatient admissions within a quarter were followed by an ambulatory care visit with a primary care or specialist provider within 14 days (binary [yes or no]/beneficiary/year)	Intervention quarters 5–10	Medicare FFS beneficiaries with at least one hospital stay in the quarter assigned to treatment practices	15.0% (+)	> 99.9%	> 99.9%	62.0	0.8 (1.0)	1.3%	0.41
	Combined (%)	Varies by test	Varies by test	15.0% (+)	> 99.9%	> 99.9%	n.a.	n.a.	1.5%	0.12
Service use (4)	All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	Intervention quarters 7–10	All observable Medicare FFS beneficiaries attributed to treatment practices	5.0% (-)	62.0%	97.1%	76.8	-5.8** (2.6)	-7.1%	0.04
	Outpatient ED visits (#/1,000 beneficiaries /quarter)	Intervention quarters 7–10	All observable Medicare FFS beneficiaries attributed to treatment practices	5.0% (-)	82.0%	99.9%	136.5	-8.2** (3.3)	-5.7%	0.02

Table V.5 *(continued)*

Primary test definition					Statistical power to detect an effect that is[a]		Results			
Domain (# of tests in domain)	Outcome (units)	Time period for impacts	Domain (# of tests in domain)	Outcome (units)	Time period for impacts	Domain (# of tests in domain)	Outcome (units)	Time period for impacts	Domain (# of tests in domain)	Outcome (units)
	All-cause inpatient admissions (#/1,000 beneficiaries/ quarter)	Intervention quarters 7–10	All observable high-risk Medicare FFS beneficiaries attributed to treatment practices	15.0% (-)	89.5%	> 99.9%	172.4	-16.8 (11.2)	-8.9%	0.17
	Outpatient ED visits (#/1,000 beneficiaries/ quarter)	Intervention quarters 7–10	All observable high-risk Medicare FFS beneficiaries attributed to treatment practices	15.0% (-)	94.1%	> 99.9%	246.3	-0.7 (13.0)	-0.3%	0.50
	Combined (%)	Intervention quarters 7–10	Varies by test	10.0% (-)	98.4%	> 99.9%	n.a.	n.a.	-5.5%**	0.03
Spending (2)	Medicare Part A and B spending ($/beneficiary/month)	Intervention quarters 9–10	All observable Medicare FFS beneficiaries attributed to treatment practices	3.0% (-)	50.6%	90.5%	$900	-$10 ($21.0)	-1.1%	0.41
	Medicare Part A and B spending ($/beneficiary/month)	Intervention quarters 9–10	All observable high-risk Medicare FFS beneficiaries attributed to treatment practices	15.0% (-)	95.6%	> 99.9%	$1,838	$34 ($90.5)	1.9%	0.56
	Combined (%)	Intervention quarters 9–10	Varies by test	9.0% (-)	89.9%	> 99.9%	n.a.	n.a.	0.4%	0.55

Sources: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at Centers for Medicare & Medicaid Services.

Notes: The results for each outcome are based on a difference-in-differences regression model. For each intervention quarter, the model calculated the regression-adjusted difference between outcomes for the treatment and comparison groups in that quarter, subtracting out any differences between the treatment and comparison groups during the baseline period.

High-risk beneficiaries are defined as beneficiaries with a Hierarchical Condition Category score in the top quarter among all beneficiaries seen by treatment practices during the period (baseline or intervention), by market area.

Table V.5 *(continued)*

[a] The power calculation is based on actual standard errors from analysis. For example, in the first row of the service use domain, a 5 percent effect on all-cause inpatient admissions for all Medicare FFS beneficiaries (from the counterfactual of 76.8 + 5.8) would be a change of 4.1 percent. Given the standard error of 2.6 percent from the regression model, we would be able to detect a statistically significant result 62.0 percent of the time if the impact was truly 4.1 percent, assuming a one-sided statistical test at the $p = 0.10$ significance level.

[b] The substantive threshold is the impact as a percentage of the counterfactual. The counterfactual is the outcome the treatment group would have had in the absence of the HCIA-funded intervention. Our estimate of the counterfactual is the treatment group mean minus the regression-adjusted difference-in-differences estimate.

[c] We show statistical power to detect a very large effect (twice the size of the substantive threshold) because this provided additional information about the likelihood that we would find effects if the program was indeed effective. If power to detect effects is less than 75 percent even for a very large effect, then the evaluation is extremely poorly powered for that outcome.

[d] Percentage difference is calculated as the regression-adjusted difference between the treatment and comparison group, divided by the adjusted comparison group mean.

[e] *p*-values test the null hypothesis that the regression-adjusted difference-in-differences estimate is less than or equal to zero for outcomes in the quality-of-care processes domain, or greater than or equal to zero in all other domains (a one-sided test). Because it is a one-sided test, as the difference-in-differences estimate approaches infinity in an unfavorable direction (negative for quality-of-care process measures and positive for all other measures), the *p*-value approaches 1, whereas it would approach 0 in a two-sided test. We adjusted the *p*-values for the multiple (three) comparisons made within the quality-of-care processes domain, and (separately) for the four comparisons made within the service use domain.

*/**/*** Significantly different from zero at the .10/.05/.01 levels, one-tailed test, respectively. No difference-in-differences estimates were significantly different from zero at the .10 or .01 level.

ED = emergency department; FFS = fee-for-service; HCIA = Health Care Innovation Award.

The combined estimate across the three measures in the quality-of-care processes domain was 1.5 percent, a favorable point estimate that was not substantively large. The statistical power to detect substantively large effects was good (more than 99 percent) for all three quality-of-care process measures individually and, in addition, combined across the measures.

Service use. For the full beneficiary sample, treatment group admission and outpatient ED visit rates were 7.1 and 5.7 percent lower, respectively, than the estimated counterfactuals, both of which were statistically significant, substantively large (relative to the threshold of 5 percent), and favorable. For the high-risk subgroup, the treatment group's admission rate was 8.9 percent lower and the outpatient ED visit rate was 0.3 percent lower, than the estimated counterfactuals. However, neither of these differences were statistically significant after adjusting for multiple statistical tests in the domain, nor where they substantially large (relative to the subgroup threshold of 15 percent).

When combining results across the four outcomes in this domain, the outcomes for the treatment group were statistically significantly lower (5.5 percent) relative to the estimated counterfactual, though not substantively larger than the 10 percent threshold for the domain. Power to detect effects of the size of the substantive thresholds was marginal for the admission rates for the full sample (62 percent) but good for the remaining outcomes (82 to 94 percent) as well as for the four outcomes combined (98 percent).

Spending. For the full beneficiary sample, the treatment group averaged $900 per beneficiary per month in Part A and B spending during the I9 and I10, a value 1.1 percent (or $10) lower than the estimated counterfactual. This difference was smaller than the substantive threshold of 3 percent. Statistical power to detect an effect the size of the substantive threshold was marginal (51 percent). For the high-risk subgroup, the treatment group's per beneficiary per month Part A and B spending was 1.9 percent higher than the estimated counterfactual. Neither of these differences was statistically significant. After combining results across the two tests in this domain, the outcome for the treatment group was almost identical (0.4 percent higher) to the estimated counterfactual.

Aggregate estimates for CMMI's core measures. The estimates presented for the CMMI core outcomes—that is, for all-cause inpatient admissions, outpatient ED visits, and Medicare Part A and B spending—have so far been expressed per 1,000 beneficiaries per quarter, or, for spending, per beneficiary per month. Table V.6 translates these rates or per-beneficiary-month estimates into estimates of aggregate impacts among the full Medicare FFS population during the primary test periods. (We do not report aggregate impacts separately for the high-risk group because the full population includes them.) We calculated these aggregate impacts by multiplying the point estimates by the average number of Medicare beneficiaries in the treatment group and by the number of quarters or months during the primary test period for the relevant outcome. Although the point estimates were small for most of these measures, the aggregate estimates for the full beneficiary population were fairly large because they were scaled to the full Medicare population of slightly fewer than 100,000 beneficiaries and to the full length of the primary test period. For example, the results in Table V.5 show the intervention was associated with a decrease in Medicare Part A and B spending of $10 per beneficiary per month, or 1.1

percent relative to the estimated counterfactual. However, across roughly 100,000 beneficiaries and six months, this small spending decrease per beneficiary per month translated into an aggregate estimated savings of the program of roughly $5.8 million. This large aggregate estimate for spending in particular should be interpreted with caution because the estimate (unlike those for ED visits or inpatient admissions) was not statistically significant. (The *p*-values for these aggregate estimates were the same as for the main results shown in Table V.5.)

Table V.6. Results for primary tests for CMMI's core outcomes expressed as aggregate effects for all Medicare FFS beneficiaries in the treatment group for TransforMED

Outcome (units)	Aggregate impact estimate during the primary test period[a]	*p*-value
All-cause inpatient admissions (#)	-2,275	0.04
Outpatient ED visits (#)	-3,186	0.02
Medicare Part A and B spending ($)	-$5,761,747	0.41

Sources: Authors' calculation, based on analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at the Centers for Medicare & Medicaid Services.

Note: To estimate the aggregate impact during the primary test period we (1) multiplied the per beneficiary per quarter (or month) estimate from Table V.5 by the average number of Medicare FFS beneficiaries in the treatment group during the primary test quarters, then (2) scaled the estimate to the full primary test period by multiplying the resulting product by the number of quarters (or months). The *p*-values are taken from Table V.5, and are therefore one-sided (testing that the program improved outcomes) and adjusted for multiple comparisons conducted within each outcome domain.

[a] The primary test period for inpatient admissions and ED visits covered July 1, 2014, through June 30, 2015 (intervention quarters 7 through 10), and for Medicare Part A and B spending covered January 1, 2015, through June 30, 2015 (intervention quarters 9 and 10).

CMMI = Center for Medicare & Medicaid Innovation; ED = emergency department; FFS = fee-for-service.

4. Results for secondary tests

As shown in Table V.7, the differences in inpatient admissions (full and high-risk samples) and spending for the treatment group and its estimated counterfactual were generally small and not statistically significant during the first intervention year (January 1 to December 31, 2013). These results helped to support the credibility of the comparison group because we did not see significant differences (favorable or unfavorable) during the first year of practice participation, a period during which we and TransforMED did not expect to see large program effects. This increased confidence in the comparison group, in turn, gave us greater confidence in the primary test results and, eventually, the conclusions of the impact evaluation.

Table V.7. Results of secondary tests for TransforMED

Secondary test definition				Results			
Domain	Outcome (units)	Time period for impacts	Population	Treatment group mean	Regression-adjusted difference between treatment and comparison groups (standard error)	Percentage difference[a]	*p*-value[b]
Service use	All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	Intervention quarters 1–4	All observable Medicare FFS beneficiaries attributed to treatment practices	76.7	2.2 (2.4)	3.0%	0.82
	All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	Intervention quarters 1–4	All observable high-risk Medicare FFS beneficiaries attributed to treatment practices	173.0	10.0 (8.9)	6.1%	0.87
Spending	Medicare Part A and B spending ($/beneficiary/month)	Intervention quarters 1–4	All observable Medicare FFS beneficiaries attributed to treatment practices	$846	-$12 (21.9)	-1.4%	0.29
	Medicare Part A and B spending ($/beneficiary/month)	Intervention quarters 1–4	All observable high-risk Medicare FFS beneficiaries attributed to treatment practices	$1,786	$4 (85.9)	0.2%	0.52

Sources: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at the Centers for Medicare & Medicaid Services.

Notes: The results for each outcome are based on a difference-in-differences regression model, as described in the text.

High-risk beneficiaries are defined as beneficiaries with a Hierarchical Condition Category score in the top quarter among all beneficiaries seen by treatment practices during the period (baseline or intervention), by market area.

[a] Percentage difference is calculated as the regression-adjusted difference between the treatment and comparison groups, divided by the adjusted comparison group mean.

[b] The *p*-values from the secondary test results were *not* adjusted for multiple comparisons within the domain.

FFS = fee-for-service.

5. Consistency of quantitative estimates with implementation findings

The impact estimates in the primary tests were plausible given the implementation findings. The primary test results showed favorable effects on service use during the last year of the program that were statistically significant. The implementation evidence showed the components of the program were generally implemented as planned with no major delays in recruiting systems and practices. Overall, 77 of the 90 participating practices completed the implementation process; 13 practices did not implement Phytel but implemented the cost management reporting software. TransforMED reported that it reached its program process and service goals; practices met goals related to patient contact measures and process measures (such as number of screenings and number of care plans), with several of these measures retired by the end of the program as practices delivered relevant services to 90 to 100 percent of their patient panels.

The lack of effects on spending and quality-of-care processes, given the sizable reduction in service use, is surprising but not implausible. We discuss possible explanations for the lack of impacts on spending in the next section. However, the impact findings overall are consistent with the implementation evidence. Despite meeting most implementation goals, TransforMED faced a few key implementation barriers, as described in Section III.C, that might have prevented practices from realizing the expected impacts on savings. In addition, practices and health systems had wide latitude in the type of process improvements they focused on during the intervention period. At least for some practices, improvements in quality-of-care processes that led to the observed reductions in service use might have come through pathways other than those examined in this evaluation. Some of those pathways might not be examinable in claims.

6. Conclusions about program impacts, by domain

Based on all evidence currently available, we have drawn the following conclusions about program impacts during the primary test periods. Table V.8 summarizes these conclusions and their support:

- **The program had statistically significant favorable effects on service use.** The primary test results showed a statistically significant favorable estimate for service use outcomes, driven by estimates for the full beneficiary sample. Large impact estimates for hospitalizations and ED visits among the full population of Medicare FFS beneficiaries drove the impact on service use. The null finding for secondary tests during a period during which we and the awardee did not expect to see large program effects supported these results. This conclusion also aligned with the implementation findings that the program was implemented reasonably well.
- **The program did not have a substantively large impact on quality-of-care processes or spending.** For all outcomes in these domains, the primary test results were neither substantively large nor statistically significant. The statistical power to detect effects in these domains was good (more than 75 percent). Specifically, in the quality-of-care processes domain, power was greater than 99 percent for each of the measures in the domain. In the spending domain, power was very good (90 percent) for the combined impact estimate across the two samples in the domain. The fact that we did not observe any declines in spending—which the awardee anticipated would follow from reductions in service use—is

somewhat counterintuitive. However, the time period used to assess impacts on spending was shorter and did not fully overlap with the time period used in the primary tests for service use. In addition, savings due to the reduction in inpatient visits might have been partially offset by increases in outpatient spending due to greater use of primary care as a result of the intervention.

Table V.8. Conclusions about the impacts of TransforMED's HCIA program on patients' outcomes, by domain

		Evidence supporting conclusion		
Domain	Conclusion	Primary test result(s) that supported conclusion	Primary test result(s) plausible given secondary tests?	Primary test result(s) plausible given implementation evidence?
Quality-of-care processes	No substantively large effect	No substantively large or statistically significant effects; evaluation was well powered to detect effects if they existed	Yes	Yes
Service use	Statistically significant favorable effect	The estimates for both all-cause inpatient admissions and outpatient ED visits (among the full Medicare FFS population) were favorable and statistically significant after adjusting for four tests in domain; the combined effect estimate in the domain was also statistically significant and favorable	Yes	Yes
Spending	No substantively large effect	No substantively large or statistically significant effects; evaluation was well powered to detect effects if they existed	Yes	Yes

Sources: Tables V.5 and V.7.

ED = emergency department; FFS = fee-for-service; HCIA = Health Care Innovation Award.

VI. DISCUSSION AND CONCLUSIONS

TransforMED used its $20.8 million HCIA to implement a PCMN program. The PCMN program focused almost exclusively on implementing health IT, with TransforMED providing software and technical assistance to 90 primary care practices recruited by 15 participating health systems in 15 states. The program had two components: (1) providing population health management and cost reporting software to practices so they could more effectively use data to improve clinical processes (for example, by monitoring utilization and spending across their patients, identifying patients who would benefit from preventive services, and sending automated emails encouraging patients to schedule recommended follow-up visits); and (2) technical assistance to practices and participating health systems on how to use the software and to promote practice transformation. Through these two components, TransforMED aimed to reduce the cost of health care for Medicaid and Medicare FFS patients by 4 percent (or $49.5 million) by the end of the award. The organization planned to achieve this aim by reducing patients' need

for acute care–such as inpatient admissions and ED visits–and improving coordination of care across providers within the PCMN community.

The results from our impact evaluation suggest TransforMED partially met these goals during the original three-year award period. Outcomes for Medicare FFS patients served by the 87 treatment practices were statistically better than those for Medicare patients served by 286 matched comparison practices in the service use domain. However, the impact estimates indicate that the intervention did not improve patients' outcomes in the two other evaluation domains; there was no evidence of statistically significant or substantively large favorable effects in either the quality-of-care processes or spending domains. The evaluation was well powered to detect substantively large impacts in all three evaluation domains.

The TransforMED program was generally implemented as planned, providing support that the favorable impacts observed on service use were due to the HCIA-funded program. Several indicators capture the successful implementation:

- TransforMED met its recruiting targets: 15 health systems and 90 practices participated in the HCIA-funded program; there were no major delays in recruiting systems and practices.
- TransforMED equipped most of the participating practices with population health management systems (78 of 90) and all health systems with cost management reporting functions.
- TransforMED provided training to practices and health systems to learn how to implement population health management systems and cost management reporting functions.
- Throughout the award, TransforMED provided technical assistance to practices and health systems to promote practice transformation.

Further, evidence from the assessment of the experience of the HCIA-funded training supports the conclusion that the program engaged practice staff at a high level. Trainee survey respondents generally reported that the training they received through the program improved their ability to provide care in a way that aligned with PCMH concepts.

However, there were two key implementation barriers for two tools practices were expected to use to identify opportunities to improve patients' care and reduce spending. These barriers might have prevented these tools from being used to the intended extent, and might have contributed to the lack of effects on spending. First, several difficulties prevented practices from using cost management reporting functions as intended. Due to the technical demands of using the cost management software to generate reports, practice staff (super users) did not customize cost management reporting to guide cost management activities. Communities were able to generate only standardized reports developed by the software vendor, Cobalt Talon, as opposed to customized reports developed by the practices that focused on unique utilization and cost issues within each community. Therefore, health systems and practices did not use cost management reporting to change the way they delivered or monitored their patients' care. In addition, in the communities that included independent practices, the transparency of the cost data at the practice level led to conflicts about the use of cost management reporting, due to

financial competition between the participating health system and the independent practices. The incentives for participation in the program were the availability of population health management software, cost reporting software, and technical assistance with implementing PCMH principles. Although these tools promote the coordination of care within a medical neighborhood, health systems and practices did not receive financial incentives to coordinate with other providers to make the program a success.

In addition, there were challenges using the data analytics reports, which TransforMED started to provide in the second half of 2014. These reports were discontinued because of the three- to six-month lag in the claims-based cost management data. This lag limited the utility of patient profile reports, which gave participating practices information on all services received by a patient and a risk score based on cost and utilization metrics. The profile reports were expected to help practices identify patients whose care could be improved and whose cost of care could be reduced through improved coordination of care across providers within the PCMN.

Overall, these findings suggest that providing practices with population health management and cost-management reporting software—along with technical assistance for how to use them—can complement practices' own PCMH transformation efforts and add meaningfully to their impacts on service use. These favorable findings likely would not replicate, however, in settings where providers lack incentives to use the IT systems and technical assistance in the same way that providers do when participating in broader PCMH efforts to transform primary care delivery, nor in practices that lack the advanced infrastructure required by TransforMED for participation in the initiative. These findings also suggest that similar practice transformation efforts might find achieving favorable impacts on spending a more difficult challenge than improving service use, because of the potential for financial competition among providers in a community. Without new or reallocated incentives for coordinating care within a community, components that were not part of the PCMN program, providers within a community might not work together to reduce overall spending.

REFERENCES

Boards of Trustees, Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds. "2013 Annual Report of the Boards of Trustees of the Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds." Table V.D1. Washington, DC: Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds, 2013. Available at http://downloads.cms.gov/files/TR2013.pdf. Accessed August 13, 2014.

Chronic Conditions Data Warehouse. "Table A.1. Medicare Beneficiary Counts for 2003–2012." Baltimore, MD: Centers for Medicare & Medicaid Services, 2014a. Available at https://www.ccwdata.org/cs/groups/public/documents/document/ccw_website_table_a1.pdf. Accessed November 19, 2014.

Keith, Rosalind, Sean Orzol, Mynti Hossain, Boyd Gilman, Catherine DesRoches, Sandi Nelson, Laura Blue, Keith Kranker, Kate Stewart, Frank Yoon, and Lorenzo Moreno. "Findings for TransforMED." In Moreno, Lorenzo, Boyd Gilman, Greg Peterson, Catherine DesRoches, Sheila Hoag, Linda Barterian, Laura Blue, Katherine Bradley, Emily Ehrlich, Kristin Geonnotti, Lauren Hula, Keith Kranker, Rumin Sarwar, Rachel Shapiro, KeriAnn Wells, Joseph Zickafoose, Sandi Nelson, Frank Yoon with the Implementation Team, Impact Team, Data Processing Team, Surveys Team, and Production Coordination and Editorial Team. "Evaluation of the Health Care Innovation Awards (HCIAs): Primary Care Redesign Programs. Second Annual Report, Volumes I and II." Princeton, NJ: Mathematica Policy Research, December 11, 2015.

Gerhardt, Geoffrey, Alshadye Yemane, Keri Apostle, Allison Oelschlaeger, Eric Rollins, and Niall Brennan. "Evaluating Whether Changes in Utilization of Hospital Outpatient Services Contributed to Lower Medicare Readmission Rate." *Medicare & Medicaid Research Review,* vol. 4, no. 1, 2014, pp. E1–E13.

Health Indicators Warehouse. "Average Age of Medicare Beneficiaries (mean)." Hyattsville, MD: National Center for Health Statistics, HIW, 2014a. Available at http://www.healthindicators.gov/Indicators/Average-age-of-Medicare-beneficiaries-mean_308/Profile/ClassicData. Accessed November 19, 2014.

Health Indicators Warehouse. "Hospital Inpatient Medicare Admissions (per 1,000 beneficiaries)." Hyattsville, MD: National Center for Health Statistics, HIW, 2014b. Available at http://www.healthindicators.gov/Indicators/Hospital-inpatient-Medicare-admissions-per-1000-beneficiaries_2001/Profile/ClassicData. Accessed August 13, 2014.

Health Indicators Warehouse. "Medicare Beneficiaries Who Are Also Eligible for Medicaid (percent)." Hyattsville, MD: National Center for Health Statistics, HIW, 2014c. Available at http://www.healthindicators.gov/Indicators/Medicare-beneficiaries-eligible-for-Medicaid-percent_317/Profile/ClassicData. Accessed August 4, 2015.

Institute of Education Sciences. "What Works Clearinghouse: Procedures and Standards Handbook, Version 3.0." Washington, DC: U.S. Department of Education, IES, 2014. Available at http://ies.ed.gov/ncee/wwc/DocumentSum.aspx?sid=19. Accessed September 15, 2014.

Peikes, Deborah, Stacy Dale, Eric Lundquist, Janice Genevro, and David Myers. "Building the Evidence Base for the Medical Home: What Sample and Sample Size Do Studies Need? White Paper." AHRQ Publication No.11-0100-EF. Rockville, MD: Agency for Healthcare Research and Quality, October 2011.

Rosenthal, Meredith, Shehnaz Alidina, Mark Friedberg, Sarah Singer, Diana Eastman, Zhonghe Li, and Eric Schneider. "A Difference-in-Differences Analysis of Changes in Quality, Utilization, and Cost Following the Colorado Multi-Payer Patient-Centered Medical Home Pilot." *Journal of General Internal Medicine,* 2016, vol. 31, no. 3, March 2016, pp. 289–296.

Stuart, Elizabeth A. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science,* vol. 25, no. 1, 2010, pp. 1–21.

U.S. Census Bureau. "2010 Census Urban Area Facts." Washington, DC: U.S. Census Bureau, 2013. Available at https://www.census.gov/geo/reference/ua/uafacts.html. Accessed July 15, 2016.

CHAPTER 10

WYOMING INSTITUTE OF POPULATION HEALTH

Andrea Wysocki, KeriAnn Wells, Greg Peterson, Boyd Gilman, Laura Blue, Keith Kranker, Kate Stewart, Sheila Hoag, and Lorenzo Moreno

This page has been left blank for double-sided copying.

WYOMING INSTITUTE OF POPULATION HEALTH

CHAPTER SUMMARY

Introduction. The Wyoming Institute of Population Health (WIPH) used its $14.2 million Health Care Innovation Award (HCIA) to implement a five-component program designed to transform rural care delivery in Wyoming, including through a patient-centered medical home (PCMH) program at 20 primary care practices serving about 130,000 patients. WIPH intended to facilitate these practices' achievement of National Committee for Quality Assurance (NCQA) PCMH recognition. WIPH's goals were that the PCMH program component—collectively with the four other distinct components—would reduce emergency department (ED) visits by 10 percent, reduce hospital admissions by 5 percent, and reduce total spending by 5 percent.

Objectives. Our objectives in this report are to (1) describe the design and implementation of WIPH's HCIA-funded intervention, primarily focusing on the PCMH component; (2) assess impacts of the PCMH component on patients' outcomes and Medicare Part A and B spending during the award period; and (3) use both implementation and impact findings to identify possible explanations for the observed impacts.

Methods. We reviewed WIPH's program documents and self-monitoring metrics, conducted interviews with WIPH leadership and program staff, and surveyed participating clinicians. To estimate impacts, we compared outcomes for Medicare fee-for-service (FFS) patients served by the 20 participating practices with outcomes for Medicare FFS patients served by 75 matched comparison practices.

Program design and implementation. The intervention had five components: (1) transformation of primary care practices into PCMHs, (2) hospital transition assistance for participants 65 or older with one of 10 qualifying conditions, (3) telehealth videoconferencing technology in hospitals and doctors' offices, (4) community-based access to free medications, and (5) the Virtual Pharmacy Program; this report focuses on the PCMH component because it had the largest target population and was the only component for which we could conduct a robust analysis. WIPH hired TransforMED—a consulting service owned by the American Academy of Family Physicians—to facilitate 20 practices' transformation into PCMHs. TransforMED held quarterly learning collaboratives, conducted site visits and telephone calls with practices, helped practices develop customized transformation plans, and reviewed practices' PCMH application documents before submission to NCQA. Participating practices were independent from WIPH and opted into the initiative.

Intermediate program effects. WIPH provided very little direct HCIA funding to participating practices because the awardee thought providing funds to practices would make them less likely to sustain the PCMH approach to care after the HCIA funding ended. As a result, practices' efforts were largely self-directed and self-funded, and their implementation experience varied. Furthermore, practices provided little data to WIPH to support our implementation evaluation. WIPH's program design required clinicians to implement new activities such as team huddles, evidence-based clinical guidelines, patient chronic condition management education, and care coordination with other providers. Most of the participating

practices received support from TransforMED, and half achieved Level 2 or Level 3 NCQA PCMH recognition by the end of the award.

Clinicians' perceptions of the intervention's effects on the care they provide. Most clinicians thought that the HCIA initiative had a positive impact on the degree to which care was patient-centered. On most other dimensions of primary care, the majority of clinicians reported that the program had no effect or that it was too early to tell. Clinicians generally supported the PCMH model of care, but many found the transformation process overwhelming given limited staff capacity, electronic health record challenges, and competing priorities.

Impacts on patients' outcomes. We are unable to draw conclusions about program impacts on patients' outcomes. The impact estimates indicate substantively large and unfavorable differences between the treatment and comparison groups for inpatient admissions, ED visits, and Medicare spending—the three target outcomes. However, secondary tests (robustness checks) and implementation findings do not support these results. The statistical power to detect effects was low for all statistical tests. Robustness checks suggest that, even though the treatment and comparison practices were well matched on observable characteristics before the intervention began, unobserved differences between the groups might have confounded the results. In addition, from the implementation findings, we would not expect to find such large unfavorable impact estimates for the outcomes examined, even if practices experienced problems.

Conclusion. WIPH contracted with TransforMED to facilitate the PCMH component, but provided little direct funding to the transforming practices. Practices' approaches to and engagement in the transformation process varied, and clinicians had mixed opinions about program effectiveness. We cannot draw conclusions about impacts on patients' outcomes. When evaluating similar programs in the future, the Center for Medicare & Medicaid Innovation and other stakeholders could consider program design changes that would improve the ability to assess impacts on patients' outcomes. Specifically, randomization could be considered, or if this is not possible, a program could be designed that is more focused to allow for valid comparison group selection.

I. INTRODUCTION

This report presents findings from the evaluation of the Health Care Innovation Award (HCIA) received by the Wyoming Institute of Population Health (WIPH), with a focus on the program's impacts on patients' outcomes. Section II provides an overview of WIPH's HCIA-funded intervention and the design of the impact evaluation. Section III describes the design and implementation of the intervention, including how the program could be expected to affect evaluation outcomes. In Section IV, we discuss intermediate effects of the intervention on practice organization and providers' behavior: specifically, the section discusses evidence on the extent to which planned changes in providers' behavior occurred, providers' perceptions of the intervention's effectiveness, and practices' success in achieving patient-centered medical home (PCMH) status. Section V describes our methods for, and results from, attempting to estimate program impacts on patients' outcomes in three domains: quality-of-care outcomes, service use, and spending. As we will describe, we are unable to draw conclusions about program impacts due to concerns about likely biases in the impact estimates. However, we still present the data for transparency and so that readers can judge the evidence for themselves. In Section VI, we discuss ways that the Centers for Medicare & Medicaid Services (CMS) or other stakeholders could modify the program design of future interventions similar to WIPH's to increase the chances of drawing reliable impact conclusions.

II. OVERVIEW OF WIPH'S HCIA-FUNDED INTERVENTION AND THE IMPACT EVALUATION

A. WIPH's HCIA-funded intervention

WIPH, a division of Cheyenne Regional Medical Center, received a $14.2 million HCIA to transform rural care delivery through the creation of medical neighborhoods across Wyoming. The Wyoming Medical Neighborhoods program included five components: (1) transformation of primary care practices into PCMHs, (2) hospital transition assistance for participants 65 or older with one of 10 qualifying conditions, (3) telehealth videoconferencing technology in hospitals and doctors' offices, (4) community-based access to free medications, and (5) the Virtual Pharmacy Program (Table II.1). The findings in this report focus on the PCMH component, for which WIPH partnered with TransforMED—a consulting service owned by the American Academy of Family Physicians—to facilitate 20 practices' transformation into National Committee for Quality Assurance (NCQA) recognized PCMHs. Participating practices were dispersed across Wyoming, with one practice in Nebraska.

WIPH's goals for the five components collectively were to reduce (1) emergency department (ED) visits by 10 percent, (2) hospital admissions by 5 percent, and (3) total spending by 5 percent by the end of the award in June 2015. WIPH also aimed to reduce preventable adverse drug events and to improve access to primary care and prescription medication. The goal of the PCMH component was to have all participating practices achieve NCQA PCMH recognition by the end of the award.

Table II.1. Summary of WIPH PCR program and our evaluation for estimating its impacts on patients' outcomes

Program description	
Award amount	$14,246,153
Award start date	June 2012
Implementation date	January 1, 2013
Award end date	June 30, 2015
Awardee description	WIPH is a division of Cheyenne Regional Medical Center dedicated to helping Wyoming communities and providers take a more proactive approach to patient care and population management.
Intervention overview	The intervention aimed to leverage existing strategic partnerships to transform rural care delivery by creating medical neighborhoods across Wyoming.
Intervention components	1. **PCMH practice transformation.** WIPH hired TransforMED to facilitate primary care clinics' transformation into PCMHs, which function as the center of the medical neighborhood. 2. **Wyoming Rural Care Transitions (WyRCT) program.** Care transitions nurses provided hospital transition assistance to participants 65 or older with at least one qualifying condition. Hospital-based nurses managed transitions for participants discharged from 14 participating acute care settings. WIPH also piloted a similar program called Transition across Community Teams (TACT) that embedded nurses in primary care settings rather than hospitals. 3. **Telehealth.** WIPH provided infrastructure for provider connectivity to facilitate care coordination and increase access to care. 4. **Medication Donation Program.** WIPH partnered with the Wyoming Department of Health to increase access to medications for eligible uninsured and underinsured low-income patients. 5. **Virtual Pharmacy.** Participating pharmacists provided participants with medication therapy management service at local pharmacies and communicated information about participants' medication use and adherence to providers. This component ended about two years into the award in July 2014, before the overall award end date.
Target population	Target populations varied by component. 1. The PCMH component targeted all patients served at participating practices. 2. The WyRCT component targeted patients ages 65 and older being discharged from the hospital with any of the following conditions: congestive heart failure, chronic obstructive pulmonary disease, coronary artery disease, diabetes, stroke, medical/surgical back disorder, hip fracture, peripheral vascular disease, cardiac arrhythmia, and pulmonary embolism. 3. The telehealth component did not specify a target population, but most consultations were for behavioral health. Clinicians also used telehealth for trainings and provider-to-provider consultations. 4. The Medication Donation Program component targeted patients with incomes up to 200 percent of the FPL, patients with no prescription coverage, patients on the Wyoming Prescription Drug Assistance Program who required three or more prescriptions per month, and Medicare beneficiaries struggling with the Part D coverage gap. 5. The Virtual Pharmacy component targeted Medicaid patients ages 18 to 65 with one of the following conditions: Depression/bipolar disorder, pain, asthma, cardiovascular disease, gastroesophageal reflux disorders/ulcers, and diabetes
Target impacts on patient outcomes	• Reduce ED visits by 10 percent • Reduce hospital admissions by 5 percent • Reduce total spending by 5 percent • Improve clinical outcomes, patients' engagement, and satisfaction (amount unspecified) • Reduce preventable adverse drug events (amount unspecified) • Improve access to primary care and prescription medication (amount unspecified)

Table II.1 *(continued)*

Workforce development	WIPH did not compensate intervention staff participating in the PCMH, telehealth, or Medication Donation Program components. However, WIPH used HCIA funding for WyRCT and TACT nurses' salaries for the duration of the award. WIPH also provided pharmacists participating in the Virtual Pharmacy component a capitated payment for each patient served.
Location	Urban and rural areas in Wyoming, plus one location in Nebraska
Impact evaluation	
Core design	Difference-in-differences with matched comparison group
Treatment group	Medicare FFS beneficiaries whom we attributed to the 20 treatment practices in Wyoming and Nebraska participating in the PCMH component
Comparison group	Medicare FFS beneficiaries whom we attributed to 75 matched comparison practices in Montana
Intervention component(s) included in impact evaluation	**PCMH component.** The impact evaluation captures the effect of the PCMH intervention component. However, beneficiaries attributed to practices in the PCMH intervention could also have been exposed to the (1) WyRCT component if they were hospitalized at one of WyRCT's 14 participating hospitals across Wyoming, met the eligibility criteria, and enrolled in the transitional care program; (2) telehealth component if they received care from one of the PCMH providers using telehealth; or (3) Medication Donation Program if they received care from a participating provider and were low-income, had no prescription coverage, or were enrolled in a Medicare Part D Prescription Drug Plan and struggling with the drug plan coverage gap.
Extent to which the treatment group reflects WIPH's target population (for the component(s) evaluated)	**Medium:** WIPH's target population included all patients seen by treatment practices and the evaluation's treatment group includes only Medicare FFS beneficiaries.
Study outcomes, by domain	1. **Quality of care-outcomes:** Inpatient admissions for ambulatory care-sensitive conditions 2. **Service use**: All-cause inpatient admissions and outpatient ED visits 3. **Spending:** Medicare Part A and B spending

Source: Review of WIPH reports, including its original application, operational plan, and 12 quarterly narrative reports and a final progress report to the Centers for Medicare & Medicaid Services.

ED = emergency department; FFS = fee-for-service; FPL = federal poverty level; HCIA = Health Care Innovation Award; PCMH = patient-centered medical home; PCR = primary care redesign; WIPH = Wyoming Institute for Population Health.

B. Overview of impact evaluation

To estimate program impacts on patients' outcomes, we compared outcomes for Medicare fee-for-service (FFS) beneficiaries served by the 20 practices participating in the HCIA PCMH intervention (treatment practices) with outcomes for beneficiaries served by 75 matched comparison practices, adjusting for observed differences in outcomes between these two groups before the intervention began. The bottom panel of Table II.1 summarizes our impact evaluation design. We estimated program impacts for the PCMH component because that component was central to the medical neighborhood, had the largest target population, and was the component for which we could create the strongest evaluation design. We were unable to estimate program impacts for the Wyoming Rural Care Transitions (WyRCT) component due to challenges identifying an appropriate comparison group. We did not estimate program impacts for the telehealth, Medication Donation Program, or Virtual Pharmacy program components because we either lacked identifiers for the participating providers, lacked claims for the majority of patients who benefited from the programs, or were unable to replicate the enrollment criteria. These factors made it difficult to construct a meaningful comparison group.

We selected the 75 comparison practices for the PCMH evaluation from a pool of practices in neighboring Montana because the WIPH PCMH program operated throughout Wyoming and those practices that chose not to participate could differ systematically from those that did. We used propensity-score matching to select practices that were similar to the 20 treatment practices before the intervention began on observable factors that can influence patients' outcomes.

We estimated impacts on outcomes, as measured in Medicare FFS claims data, which we grouped into three domains: (1) quality-of-care outcomes, (2) service use, and (3) spending. Across the HCIA awardees in primary care redesign (PCR), we designed our impact evaluations to identify promising interventions or intervention components—consistent with evaluation goals from the Center for Medicare & Medicaid Innovation (CMMI) to find programs that could be scaled or retested as part of a future study. Before conducting analyses, we specified a series of primary tests, describing the evidence we would need to conclude that the program was effective, and WIPH and CMMI reviewed these tests. Each test specified a population, outcome, period, expected direction of effect, and threshold that we count as substantively important. The purpose of these primary tests was to focus the impact evaluation on hypotheses that would provide the most robust evidence about program effectiveness. Because we sought to identify promise, rather than only those programs with unequivocally demonstrated success, we conducted one-sided statistical tests (that is, testing only for program benefits) and used a threshold for statistical significance of 0.10, which is not as strict as the conventional standard of 0.05. We used the results from the primary tests and robustness checks, in combination with the implementation findings, to determine whether we could draw conclusions about program impacts in each of the three evaluation domains. After applying standard decision rules, which we describe in Section V.A.8, we determined that it was not possible to draw impact conclusions for WIPH for any domain. We still present the full set of results for transparency and to enable readers to judge the evidence for themselves.

WIPH's target population for the PCMH program component included all patients seen by treatment practices, but the treatment group for our impact evaluation included only Medicare FFS beneficiaries. These beneficiaries could also have been exposed to the (1) WyRCT component if they were hospitalized at one of WyRCT's 14 participating hospitals across Wyoming, met the eligibility criteria, and enrolled in the transitional care program; (2) telehealth component if they received care from one of the PCMH providers using telehealth; or (3) Medication Donation Program if they received care from a participating provider and were low-income, had no prescription coverage, or were enrolled in a Medicare Part D Prescription Drug Plan and struggling with the drug plan coverage gap. Thus, even though we designed our impact evaluation to capture the marginal impact of the PCMH program component, our estimates might, in some cases, also capture effects of other program components. Our impact evaluation is not a test of the PCMH model; rather, it is a test of WIPH's PCMH intervention component that included practice facilitation from TransforMED.

III. PROGRAM IMPLEMENTATION

This section first provides a detailed description of WIPH's HCIA-funded intervention, highlighting intervention components, PCMH practice recruitment and target population, the PCMH theory of action, and workforce development. Second, it assesses the evidence on the

extent to which the intervention was implemented as planned based on services provided and timeliness. This section addresses only implementation of HCIA-funded services, delivered by WIPH. We describe the intermediate effects of these services on practice organization and providers' behavior—including success achieving PCMH recognition from NCQA—in Section IV.

We based our evaluation of WIPH's program implementation on a review of its quarterly reports to CMMI and self-monitoring program metrics, telephone discussions and follow-up communications with program administrators, and information collected during site visit interviews with frontline and administrative staff at selected practices conducted in April 2014 and April 2015. We did not verify the quality of the performance data reported by WIPH in its self-measurement and monitoring reports.

A. Program design and adaptation

1. Intervention components

WIPH's intervention had five components, which were designed to collectively serve as medical neighborhoods that coordinated care among PCMHs, specialists, pharmacists, hospitals, community organizations, and public health agencies to provide comprehensive, patient-centered care. We describe the five components next.

1. **Transform primary care practices into NCQA recognized PCMHs.** WIPH hired TransforMED to facilitate transformation of 20 primary care practices into NCQA-recognized PCMHs that would serve as the centers of medical neighborhoods throughout Wyoming. However, WIPH directed less than $1 million of its total $14.2 million HCIA funds to TransforMED for practice facilitation services and to participating practices for NCQA application fees. Otherwise, practices received no direct HCIA funding. WIPH was concerned that funding practices directly would make it difficult for them to sustain the intervention after the HCIA funding ended, and chose instead to allocate the majority of funds invested in the PCMH component to TransforMED.
2. **WyRCT program.** WyRCT trained hospital-based nurses to manage transitions for patients discharged from 14 acute care settings. At two practices in Cheyenne, WIPH piloted a similar outpatient-based health coaching program, Transition Across Community Teams (TACT). TACT nurses offered similar services as WyRCT nurses, including post-discharge home visits and medication reconciliation, with the goal of preventing hospitalizations. WIPH funded WyRCT and TACT nurses' salaries and benefits for the duration of the award, which accounted for more than half of HCIA funds. (WIPH allocated remaining HCIA funds to the other three components and to administrative expenses, including compensation for project leaders and administrators, marketing and evaluation consultants, and indirect costs.)
3. **Telehealth.** The telehealth component provided infrastructure for providers' connectivity to facilitate care coordination and increase access to care. WIPH installed desktop and mobile video conferencing technology at practices and hospitals across Wyoming, including some practices participating in the PCMH component. Most patients' telehealth consultations

were for mental and behavioral health, bariatrics, rheumatology, endocrinology, and oncology.

4. **The Medication Donation Program.** WIPH partnered with the Wyoming Department of Health to lead the Wyoming Medication Donation Program to increase access to medication for eligible uninsured and underinsured low-income patients. The Department of Health solicited donations from nursing homes, assisted living facilities, detention centers, and other public and health care organizations. Medications were then distributed to participating providers who offered them to eligible patients.
5. **Virtual Pharmacy.** The School of Pharmacy at the University of Wyoming led the Virtual Pharmacy program. Participating pharmacists provided participants with (in-person) medication therapy management service at local pharmacies and (virtually—that is, by fax or email) sent providers information about participants' medication use and adherence. Due to various challenges, patient enrollment in Virtual Pharmacy remained very low. CMMI did not renew funding for Virtual Pharmacy for the third award year, and WIPH closed the program in July 2014.

Together, these five components aimed to improve access to and quality of health care for patients across Wyoming. We focus the rest of the report and our impact evaluation on the PCMH component.

2. Practice recruitment and target population

In this section, we describe how WIPH selected practices to participate in the PCMH component of the HCIA intervention and describe the PCMH component's target population.

Identification of practices for participation. WIPH leveraged existing partnerships to reach out to Wyoming primary care practices, most of which were not affiliated with WIPH. The Wyoming Integrated Care Network (WyICN), a system of hospitals, acted as an early advocate, helping to recruit hospital-based practices. Outreach and word of mouth informed other practices across Wyoming of the HCIA and interested practices opted in to the program. There were no explicit criteria for a primary care practice to qualify for the program. (See Section V.B for characteristics of the practices that opted in.) The PCMH program recruited 20 primary care practices, double its initial target of 10 practices. These 20 practices served about 130,000 patients. By February 2014, the program lost two participating practices that decided not to apply for NCQA PCMH recognition, although it gained two additional practices, bringing the total back to 20 transforming practices. The PCMH settings were diverse, including independent physician practices, hospital-based practices, rural health clinics (RHCs), and federally qualified health centers (FQHCs). Eight participating practices were located in more densely populated regions of Wyoming, including Casper, Cheyenne, Jackson, and Laramie. The remaining practices were located in less-populated cities and rural areas, including Saratoga and Thermopolis, and one in Kimball, Nebraska, near the Wyoming border.

Target population. Because WIPH's PCMH component was a practice transformation initiative, there was no specific target patient population. All patients who received care at the participating practices could benefit from PCMH strategies such as expanded office hours.

However, patients with chronic conditions or frequent hospitalization might have experienced more dramatic changes in the care they received due to PCMH transformation, such as increased coordination with specialists, care management services, or post-hospitalization follow-up care.

3. Theory of action

Based on extensive review of WIPH's program activities and goals, we developed a theory of action to depict the mechanisms through which program administrators expected the PCMH component to improve the outcomes we selected for the impact evaluation (Table II.1 lists these outcomes).

1. **Primary care practices engage in trainings and collaborative meetings to learn the principles of the PCMH model and strategies for practice transformation.** WIPH hired TransforMED to engage participating practices and facilitate practice transformation. TransforMED conducted site visits, led telephone calls, and convened quarterly learning collaboratives to support workforce development and facilitate practice transformation. During site visits, TransforMED helped each practice develop customized plans for transformation. Early learning collaboratives offered foundational information, educating practices generally about the value of the PCMH model. About half way through the award in spring 2014, TransforMED shifted the focus of learning collaboratives to focus more specifically on the NCQA application.
2. **Primary care practices implement changes across the six PCMH standards.** Practices had to demonstrate proficiency in six standards to achieve NCQA PCMH recognition: (1) enhance access and continuity, (2) identify and manage patient populations, (3) plan and manage care, (4) provide self-care support and community resources, (5) track and coordinate care, and (6) measure and improve performance (NCQA 2011; standards were updated in 2014, but practices used the 2011 standards during the award). Each standard included several elements, one of which practices had to pass to achieve recognition. Practices reported implementing these elements in sequential order. Several practices had physician champions who led the transformation within their practices.
3. **Primary care practices work with TransforMED and WIPH to obtain NCQA PCMH recognition.** Staff at participating practices described the process of transformation and completing the application as requiring advanced use of electronic health records (EHRs), especially for population health management, performance measurement, and quality improvement. Completing the NCQA application required practices to upload EHR screen shots demonstrating proficiency across the six standards. TransforMED reviewed practices' applications and provided feedback to practices before they uploaded documents to NCQA.
4. **Implementing practice changes across the six standards results in less fragmented, higher quality care.** On the path to achieving NCQA PCMH recognition, practices offered patients additional access points, such as evening office visits or care managers, hoping to reduce patients' need to visit the ED and increase appropriate-venue care. Clinicians educated patients to self-manage chronic conditions to improve patients' health status. Clinicians also increased coordination with other providers, especially following hospital discharge, to reduce the likelihood of medication errors and other complications. Integrating

performance measurement helped providers identify their most vulnerable patients and improve prescribing and other treatment practices.

5. **The improvements in primary care access, focus on the prevention and management of chronic disease, and adherence to evidence-based care reduce the need for acute care.** WIPH theorized that as patients' access to care increased and as they learned to better self-manage their conditions, their health status would improve, in turn resulting in fewer ED visits and hospitalizations, generating cost savings to Medicare and other payers.

4. Intervention staff and workforce development

Practices participating in the PCMH component were largely independent from WIPH and did not receive HCIA funding to hire or train intervention staff. Therefore, the types of staff working to achieve NCQA PCMH recognition at participating practices varied widely, and WIPH did not report specific staffing information. During site visits to 5 of the 20 practices, we observed that staff leading the transformation included physicians, a nurse practitioner, a practice manager, a quality specialist, and a nurse manager.

Although participating practices did not receive HCIA funds directly, they had access to TransforMED's practice facilitation services. Early in the intervention, TransforMED staff visited transforming practices and helped them develop practice transformation plans (PTPs). Staff at transforming practices also attended TransforMED quarterly learning collaboratives to share information with staff from other transforming practices. The collaboratives initially focused on building foundational awareness about the PCMH approach to care. In response to feedback from practices, in spring 2014 TransforMED began to focus learning collaboratives on the NCQA application process. TransforMED also conducted site visits and conference calls with PCMHs. About halfway through the award period, TransforMED helped practices develop work plans, which focused more on the NCQA PCMH application than had the original PTPs. TransforMED also reviewed practices' application documents and provided feedback, which staff we interviewed at transforming practices said they found very helpful.

B. Implementation effectiveness

In this section, we examine the evidence on implementation effectiveness. We assess the evidence on implementation effectiveness in two areas: (1) TransforMED services and (2) implementation timeliness. To conduct this analysis, we used data from interviews with program administrators and frontline staff, and self-reported metrics included in WIPH's quarterly narrative reports to CMMI. We did not administer a survey to assess the effectiveness of HCIA-funded training for the PCMH component. Instead, we conducted the trainee survey with WyRCT staff, because WyRCT had a formal training curriculum and WIPH included WyRCT staff contact information only in its list of trainees eligible for the survey. Furthermore, because the PCMH component did not directly enroll patients or report the number of patients served, we do not discuss enrollment as a performance metric.

1. TransforMED services

TransforMED led eight quarterly learning collaboratives from March 2013 to December 2014. WIPH reported that representatives from 18 transforming practices attended the first collaborative. Although WIPH did not report on attendance at the subsequent collaboratives, we assume that most transforming practices were also represented at the subsequent collaboratives because WIPH required attendance from physicians and other staff leading the transformation at their practices. WIPH did not report how many site visits or conference calls occurred, but it did report that 20 PTPs were completed by June 2013. Later in the intervention, TransforMED worked with 16 practices to develop NCQA application work plans, which were completed by June 2014. TransforMED also reviewed NCQA documents for at least 12 transforming practices from June 2014 to January 2015, at which point TransforMED's contract with WIPH ended. WIPH hired a new PCMH coordinator to help facilitate transformation after TransforMED's contract ended; we have very little information about this coordinator's work with transforming practices. We also have limited information about which practices participated in learning collaboratives, developed work plans, or received application review services, and about the intensity and content of participating practices' interactions with TransforMED.

2. Program timeline

Table III.1 identifies several major milestones of the PCMH component and the dates those milestones were achieved. WIPH recruited most of the PCMH practices by October 2012. In January 2013, practices began implementation activities; for example, 17 practices completed an initial assessment that TransforMED used to develop PTPs. TransforMED held its first quarterly learning collaborative in March 2013. In February 2014, all 20 participating practices had begun the process of transformation. In response to feedback, TransforMED shifted its practice facilitation from a more general foundational approach to a more specific approach focusing on the NCQA PCMH application. In fall 2014, TransforMED worked with sites to develop work plans for completing the NCQA PCMH application. We discuss practices' success in achieving NCQA PCMH recognition in our discussion of intermediate impacts in Section IV.B.

C. Conclusions about the extent to which the program, as implemented, reflects core design

We do not have sufficient information to determine whether the intervention represented a reasonable test of the PCMH program's core design. WIPH invested less than $1 million in TransforMED's practice facilitation services, but otherwise WIPH offered very little funding to transforming practices and it exercised no official authority or leverage over participating practices. Given these facts, we have minimal data with which to evaluate the implementation of the PCMH component. We know that practices had access to TransforMED's PTPs, learning collaboratives, and NCQA application document review services, and that they found document review services particularly helpful. However, we have incomplete information about the amount of exposure practices had to TransforMED's available services, such as which practices attended learning collaboratives.

Table III.1. WIPH's PCMH component timeline

Milestone	Date completed
Kickoff meeting	September 2012
Seventeen practices recruited	October 2012
Practices began implementation activities	January 2013
Three practices recruited	February–March 2013
First TransforMED quarterly learning collaborative	March 2013
PTPs completed	June 2013
One practice recruited	December 2013
One practice recruited; two practices dropped	February 2014
TransforMED shifted the practice facilitation approach from a general foundational model to a more specific focus on NCQA PCMH application tasks	March 2014
Practice-specific work plans developed	September 2014

Sources: WIPH's quarterly narrative reports and the NCQA PCMH recognition directory.

NCQA = National Committee for Quality Assurance; PCMH = patient-centered medical home; PTP = practice transformation plan; WIPH = Wyoming Institute of Population Health.

IV. INTERMEDIATE PROGRAM EFFECTS ON PRACTICE ORGANIZATION AND CLINICIANS' BEHAVIOR

This section describes the available evidence on the extent to which WIPH's intervention had its intended effects on practice organization and clinicians' behavior, expected to mediate the desired impacts on patients' outcomes. As described in Section III.A.3, the program's theory of action required that clinicians (1) engage in collaborative meetings and other practice facilitation services to learn the principles of the PCMH model and best practices for practice transformation and (2) implement changes in support of the six PCMH standards. The theory of action further required that clinicians and program administrators work with TransforMED to obtain NCQA PCMH recognition.

In this section, first, we use information from site visit interviews to describe practice transformation activities that occurred in clinical settings; we use qualitative data from our site visits because practices' efforts at transformation were largely self-directed and the practices reported a limited amount of data to WIPH. Second, we report transforming practices' success in achieving NCQA PCMH recognition. Third, we use data from two rounds of a survey we administered (the HCIA Primary Care Redesign Clinician Survey) and from site visit interviews to assess changes in providers' behavior. Both survey rounds rely on self-reported responses and reflect clinicians' perceptions of the program, rather than measuring direct program effects on the care clinicians provided. Finally, the last section summarizes the facilitators and barriers associated with implementation effectiveness at the practice level.

A. Services participating practices provided to patients

As noted in Section III.A.3, achieving NCQA PCMH recognition requires practices to demonstrate proficiency in six standards: (1) enhance access and continuity, (2) identify and manage patient populations, (3) plan and manage care, (4) provide self-care support and

community resources, (5) track and coordinate care, and (6) measure and improve performance. Practices must include screen shots of their EHRs and reports in their applications to demonstrate proficiency.

During site visit discussions, staff at practices said they generally implemented standards sequentially. Staff at one practice said it took about six months to write policies, track patient appointments, and write reports necessary to implement the first two standards. Staff at three practices said that team huddles—defined as part of continuity in NCQA's first standard—had been helpful and well received. During huddles, care teams discussed patients scheduled to visit the office—for example, identifying patients who had recently been to the hospital or who were due for a mammogram.

Practices also implemented previsit planning—a requirement for the third standard—often focusing on high-risk patients such as those with congestive heart failure and diabetes. Previsit planning generally involved a nurse or care coordinator calling a patient before a visit to schedule lab tests and review goals before the visit. Staff at two practices noted that previsit planning was challenging to implement due to inadequate staff time.

Related to the fourth standard—that practices provide self-care support and community resources—staff said that the initiative had increased efforts to engage patients, such as educating patients about their conditions and working with them to set goals. A care coordinator at one practice followed up with patients, talked with specialists, and aligned patients with community resources, satisfying the fourth and fifth standards. To demonstrate proficiency in the sixth standard, practices worked with the EHRs to write reports tracking patients and showing quality measure results. For example, practices reported on referrals, next available appointments, patient satisfaction survey results, diabetic control, and patients due for cervical cancer screenings. Staff discussed EHR challenges related to reporting, such as ensuring that data elements were documented and stored correctly. Staff often wrote their own reports rather than relying upon automated reports available in the EHR.

B. Successes in achieving NCQA PCMH recognition

WIPH originally planned to recruit 10 practices, all of which the awardee expected would achieve NCQA PCMH recognition. Instead, it recruited 20 practices that participated in the PCMH program. According to data obtained from WIPH and NCQA's web site (NCQA 2016), 10 of the 20 participating practices ultimately achieved either Level 2 or Level 3 PCMH recognition by the end of the award in June 2015, as shown in Table IV.1. Two clinics notified WIPH in July 2014 that they would not move forward with the NCQA application. One of these clinics cited the facility's transition to a new EHR and the other cited a lack of resources to transform as reasons for withdrawal. The remaining practices experienced delays achieving NCQA PCMH recognition. WIPH's target date for PCMHs to submit their applications for NCQA recognition was January 2015, although only 3 practices met this deadline. Three practices that did not achieve NCQA PCMH recognition by June 2015 still planned to pursue recognition. Given overall trends in health care delivery, it is impossible to determine which practices, if any, might have implemented practice changes and achieved NCQA PCMH recognition in the absence of the HCIA.

Table IV.1. Practices' NCQA PCMH recognition status as of June 2015

Practice	NCQA recognition status	Date recognition earned
Adult and Geriatric Medicine	Will pursue at a later date	-
Big Horn Clinic Basin	Will pursue at a later date	-
Big Horn Family Medicine	Did not achieve	-
Big Horn	NCQA Level 3	April 2015
Carol Fisher, M.D.	NCQA Level 2	June 2015
Cheyenne Plaza Primary Care	NCQA Level 2	April 2015
Community Health Center of Central Wyoming	NCQA Level 2	May 2015
Jackson Whole Family Health	Did not achieve	-
Kimball Health Services	Did not pursue	-
Lander Medical Clinic	NCQA Level 3	April 2015
Memorial Hospital of Converse County	Did not achieve	-
Midway Clinic	NCQA Level 2	June 2015
North Big Horn Hospital Clinic	NCQA Level 3	April 2015
Platte Valley Medical Clinic	NCQA Level 3	January 2015
Red Rock Family Practice	Did not achieve	-
Rendezvous Medical	NCQA Level 3	January 2015
St. John's Family Practice	Will pursue at a later date	-
South Lincoln Medical	Did not achieve	-
University of Wyoming Family Medicine Residency Casper	NCQA Level 3	January 2015
Western Medical Associates	Did not pursue	-

Sources: Correspondence with WIPH and NCQA PCMH recognition directory.

Note: Some practices that did not achieve NCQA recognition by the end of the award might have made progress transforming their practice during the award period, and could still be pursuing NCQA PCMH recognition.

MD = doctor of medicine; NCQA = National Committee for Quality Assurance; PCMH = patient-centered medical home.

C. Clinician survey

Survey methods. We administered a clinician survey in two rounds (fall 2014 and summer 2015). We sent the survey to clinicians—including physicians, physician assistants, and nurse practitioners—working at the 20 practices participating in the PCMH component. In the first round of the survey, 82 of 102 eligible clinicians responded, resulting in a response rate of 80 percent; in the second round, 86 of 143 eligible clinicians responded, resulting in a response rate of 60 percent. There were more eligible clinicians in the second round because (1) the second round included clinicians from a practice not surveyed in the first round and (2) fewer clinicians were deemed ineligible than in the first round. Clinicians were ineligible if their survey responses indicated that they were a resident or fellow or if they reported that they did not have direct contact with patients.

Survey results. Most respondents to the clinician survey reported being somewhat or very familiar with the HCIA program (76 percent in Round 1 and 85 percent in Round 2). As shown in Table IV.2 among clinicians familiar with the program, the program appears to have had limited effects on clinicians' perceptions of several dimensions of care. Specifically, 54 percent (Round 1) and 56 percent (Round 2) of clinicians familiar with the program perceived a positive impact on the degree to which care was patient-centered. However, more than a third (40 percent in Round 1 and 33 percent in Round 2) perceived a negative impact on efficiency. This is consistent with data collected during interviews in which clinicians described EHR-related tasks as particularly burdensome. On other dimensions of care—including safety, quality, equity, ability to respond to patients in a timely way, and the availability of information for clinical decision making—most clinicians familiar with the program perceived either no effect or stated that it was too soon to tell. Although we observed increases from Round 1 to Round 2 in clinicians' perception of positive impacts on quality, equity, and ability to respond to patients in a timely way, fewer than half of clinicians reported that the PCMH component had a positive impact on these dimensions of care in either survey. Discussions during site visits suggested that clinicians and staff felt overwhelmed by the transformation process, given busy schedules, limited staff, and competing initiatives such as Medicare's Physician Quality Reporting System.

Table IV.2. Primary care providers' perceptions of the effects of the program on the care they provided to patients, from the clinician survey (both rounds)

	Percentage (and number) of clinicians reporting that the HCIA had the following effect on the care they provided to patients served by their practices in the past year			
	First round of survey (20 to 22 months after program implementation) N = 63		Second round of survey (28 to 30 months after program implementation) N = 73	
Dimension of care	Positive impact	No impact or too soon to tell	Positive impact	No impact or too soon to tell
Patient-centeredness	54% (34)	40% (25)	56% (41)	41% (30)
Quality	38% (24)	51% (32)	47% (34)	49% (36)
Ability to respond in a timely way to patients' needs	29% (18)	57% (36)	38% (28)	51% (37)
Safety	38% (24)	56% (35)	36% (26)	59% (43)
Information available for clinical decision making	n.a.[a]	n.a.[a]	27% (20)	68% (50)
Equity	22% (14)	71% (45)	26% (19)	70% (51)
Efficiency	21% (13)	40% (25)	21% (15)	47% (34)

Source: HCIA Primary Care Redesign Clinician Survey: Round 1 (field period September 2014 to November 2014), Round 2 (field period May 2015 to July 2015).

Note: The number (and percentage) is limited to clinicians who reported that they were at least somewhat familiar with the HCIA program.

HCIA = Health Care Innovation Award.

[a] The first survey round did not ask this question.

n.a. = not applicable.

D. Summary of facilitators of and barriers to program implementation at the practice level

Several factors facilitated both implementation of WIPH's HCIA-funded intervention and practices' ability to transform their care model based on the intervention; however, other factors hindered implementation and practice transformation. We described those factors in detail in the second annual report (Ehrlich et al. 2015). Here we summarize key facilitators and barriers, along with any new information since the second annual report that supports those facilitators or barriers (Table IV.3).

One factor was particularly important in facilitating the program, one factor both facilitated and hindered the program, and three factors were barriers. First, clinicians and staff cited the availability of same-day appointments, team huddles, previsit planning, and new patient reports as improving their care relative to before the HCIA. Second, TransforMED learning collaboratives, new patient reports showing improved quality metrics, and reduced workloads as physicians allocated certain tasks to nurses and administrative staff all helped increase physicians' buy-in and engagement in the program. However, some physicians were less engaged, citing limited time and competing priorities, and practices lacking a physician champion were not as engaged in the intervention as practices that had champions. Third, technology hindered implementation. Staff at several practices said the PCMH transformation, which required creating new types of patient reports and new EHR processes, was administratively burdensome, but that the availability of qualified staff helped overcome this challenge. Practices lacking dedicated information technology (IT) staff or adapting to a new EHR faced more obstacles applying for and receiving NCQA PCMH recognition. Fourth, insufficient staff capacity hindered program implementation at practices that lacked dedicated IT staff or whose staff had minimal time to dedicate to practice transformation. Finally, WIPH did not provide HCIA funding to participating practices for staff or EHR upgrades, which—had fundingbeen offered—might have mitigated challenges related to technology and staff capacity.

E. Conclusions about intermediate program effects on practice organization and clinicians' behavior

Based on the information available for this evaluation, the HCIA-funded initiative appears to have had limited effects on practice organization and on how clinicians provided care. In some cases, it might even have negatively affected clinicians' care during the award period. Ten practices participating in WIPH's PCMH program achieved NCQA PCMH recognition, some of which reported that TransforMED's services—particularly documentation review—facilitated their ability to achieve this goal. Most clinicians surveyed were aware of the program and most believed the HCIA-funded initiative improved the patient-centeredness of care. However, more than a third of clinicians said that the component had a negative effect on the efficiency of care and a majority observed either no effects or reported that it was too soon to tell whether the initiative improved other dimensions of care. We have few direct metrics to assess the extent of practice transformation. Practices' efforts at transformation were largely self-directed and the practices reported only a small amount of data to WIPH.

Table IV.3. Summary of key facilitators of and barriers to the implementation of WIPH's program and practice transformation

Item	Description based on findings in the second annual report	Additional supporting data not available in the second annual report
	Facilitators (domain)	
Perceived relative advantage (program characteristics)	Perceived relative advantage of PCMH approach to care, including: • Availability of same-day and evening or weekend appointments • Team huddles and previsit planning • Patient reports	
Staff engagement (implementation process)	Increased physician engagement via: • TransforMED's learning collaboratives and assistance with NCQA applications • Reports on quality measures • Reduced workloads	
	Barriers (domain)	
Staff engagement (implementation process)	Lack of physician engagement/physician champion	Most clinician respondents were not very familiar with the HCIA program. In the first round of the clinician survey, 47 percent reported being only somewhat familiar and 19 percent reported being not at all familiar. In the second round, 49 percent were only somewhat familiar and 15 percent were not at all familiar.
Technology (internal factors)	Lack of technical aptitude, especially at practices transitioning to new EHRs or lacking dedicated IT staff	
Capacity (internal factors)	Insufficient staff capacity at many practices to implement required changes, especially changes related to the EHR	
Program resources (implementation process)	Very little direct HCIA funding provided to participating practices	When asked to rate the impact of the level of program funding on the implementation of the HCIA initiative, about half of clinicians surveyed chose "Not applicable," likely because practices did not receive direct HCIA funding. About a third of clinicians in both rounds of the survey chose 3 or higher on a scale in which 1 meant very positive impact and 5 meant very negative impact.

Note: We reviewed four domains associated with implementation experience: (1) program characteristics, (2) implementation process, (3) internal factors, and (4) external environment. Implementation research suggests that barriers and facilitators within these domains are important determinants of implementation effectiveness.

EHR = electronic health record; HCIA = Health Care Innovation Award; IT = information technology; NCQA = National Committee for Quality Assurance; PCMH = patient-centered medical home; WIPH = Wyoming Institute of Population Health.

V. PROGRAM IMPACTS ON PATIENTS' OUTCOMES

This section of the report presents results for the quantitative analysis that aimed to draw conclusions about the impacts of WIPH's HCIA program on patients' outcomes in three domains: quality-of-care outcomes, service use, and spending. We first describe the methods for estimating impacts (Section V.A) and then the characteristics of the 20 HCIA treatment practices at the start of the intervention (Section V.B). We next demonstrate that the treatment practices were similar at the start of the intervention to the practices we selected as a comparison group, which is important for limiting potential bias in impact estimates (Section V.C). Finally, in Section V.D, we describe the quantitative impact estimates, their plausibility given implementation findings, and why we were unable to draw conclusions in any of the study domains.

A. Methods

1. Overview

We estimated program impacts on patients' outcomes as the difference in outcomes for Medicare FFS patients served by the 20 treatment practices and those served by 75 matched comparison practices, adjusting for any observed differences in outcomes between these groups during the year before the intervention began. We prespecified primary tests, describing the evidence we would need to conclude that the program was effective, and WIPH and CMMI reviewed these tests. Each test specified a population, outcome, time period, expected direction of effect, and threshold that we count as substantively important. The purpose of these primary tests was to focus the impact evaluation on hypotheses that would provide the most robust evidence about program effectiveness. Based on the results from the primary and secondary tests (robustness checks), we determined that we were unable to draw conclusions about program impacts for any of the three evaluation domains. The remaining subsections describe each component of the impact evaluation in more detail. The findings in this report update the impact results from the second annual report for WIPH (Ehrlich et al. 2015), extending the outcome period by 6 months and adding to the analysis two treatment group practices that joined the HCIA-funded program after the others.

2. Treatment group definition

Practices joined the PCMH program in waves. For the impact evaluation, we organized practices into two cohorts based on the overall PCMH implementation start date and when practices joined the intervention. We defined the 18 practices that joined by the start of the intervention in January 2013, or joined in February to March 2013, as cohort one practices and set the intervention start date for cohort one practices to January 1, 2013. We defined the 2 practices that joined the intervention in December 2013 to February 2014 as cohort two practices and set the intervention start date for cohort two practices to January 1, 2014. The treatment group consisted of Medicare FFS patients served by the 20[1] treatment practices in 4 baseline

[1] Twenty practices were part of the PCMH intervention, including one practice with two locations and two separate site identifiers that is considered two practices for the purpose of the impact evaluation. We excluded one practice

quarters before the intervention began (January 1, 2012, to December 31, 2012, for cohort one practices and January 1, 2013, to December 31, 2013, for cohort two practices) and 10 intervention quarters for cohort one practices (January 1, 2013, to June 30, 2015) and 6 intervention quarters for cohort two practices (January 1, 2014, to June 30, 2015).

We constructed the treatment group in three steps.

1. We attributed beneficiaries to practices using similar decision rules that CMMI uses for the Comprehensive Primary Care Initiative. Specifically, in each baseline and intervention month, we attributed beneficiaries to the primary care practice whose providers (physicians, nurse practitioners, or physician assistants) provided the plurality of primary care services in the past 24 months. When there was a tie, we attributed the beneficiary to the practice he or she visited most recently. WIPH provided some identifiers for the treatment practices and the providers who worked in them. We obtained data on providers in other practices from SK&A, an outside health care data vendor that maintains and verifies lists of providers who work in practices throughout the country, and we used the SK&A data to supplement the treatment provider data from WIPH.
2. In each baseline and intervention quarter, we *assigned* the beneficiary to the first treatment practice to which he or she was attributed in the period (baseline or intervention), and continued to assign the beneficiary to that practice for all quarters in the period. That is, a beneficiary could be attributed to one practice in one quarter and another practice in the next, based on recent visits; however, we *assigned* each beneficiary to only one practice per period, either baseline or intervention. This rule ensured that, during the intervention period, beneficiaries did not exit the treatment group solely because the intervention succeeded in reducing their service use (including visits at treatment panels). The definition for the baseline period corresponds to that of the intervention period so that, across the two periods, interpretation of the population changes over time should be comparable.
3. We applied additional restrictions to refine the analysis sample in each quarter. A beneficiary assigned to a treatment practice in a quarter was included in the analysis sample for that quarter if he or she (1) had observable outcomes for at least one day in the quarter; and (2) lived in Wyoming, Nebraska, or Montana for at least one day of the quarter. Outcomes were observable for beneficiaries who were enrolled in Medicare FFS (Part A and B), were alive, and had Medicare as their primary payer.

3. Comparison group definition

The comparison group consisted of Medicare FFS beneficiaries whom we assigned to 75 matched comparison practices in each of the baseline and intervention quarters. The comparison practices were similar to the treatment practices during the baseline period on observable factors that can influence patients' outcomes. This section describes how we constructed the matched

that did not submit any identifying information, so although it was part of the intervention, it was not included in the impact evaluation. WIPH did not provide identifiers for the two practices that dropped out of the intervention by February 2014, so these practices were not included in the impact evaluation.

comparison group; Section V.C shows the balance we achieved between the two groups on the matching variables.

We identified the 75 comparison practices in four steps:

1. We used data from SK&A to develop a list of potential comparison practices. We also obtained CMS Certification Numbers from the Integrated Data Repository for FQHCs and RHCs. Because the WIPH PCMH program operated throughout Wyoming and those practices that chose not to participate could differ systematically from those that did, we selected comparison practices from neighboring Montana. Montana was selected as a suitable comparison because it shares many similarities with Wyoming, including similar socioeconomic characteristics, a high proportion of FFS Medicare and Medicaid beneficiaries, frontier state designation relevant for Medicare payment levels, and similar Medicaid income-eligibility limits that affect dual eligibility. Further, WIPH considered Montana a suitable state from which to draw comparison practices.
2. We developed matching variables, defined at the start of the intervention (January 1, 2013, for cohort one practices and January 1, 2014, for cohort two practices), for all treatment and potential comparison practices (N = 342). These variables included characteristics of the practice (for example, the number of primary care providers in the practice and whether a hospital or health system owned the practice); and characteristics of Medicare FFS beneficiaries assigned to the practices (for example, mean Hierarchical Condition Category [HCC] score and utilization in the baseline period). When assigning Medicare beneficiaries to the practices, we used the same attribution and practice assignment logic that we used for the treatment practices, as described previously. Section V.C describes the matching variables and their data sources in detail.
3. We dropped potential comparison practices that were unlike treatment practices because they had (1) NCQA PCMH recognition in the baseline period or (2) an average of fewer than 25 assigned Medicare FFS beneficiaries in the baseline period. We also dropped potential comparison practices that were not appropriate matches for our treatment practices, such as Indian Health Services practices. This resulted in a pool of 329 potential comparison practices.
4. Finally, we used propensity score methods to select comparison practices (from the pool of 329) that were similar to the 20 treatment practices on the matching variables. The propensity score is the predicted probability, based on all of a practice's matching variables, that a given practice was selected for treatment (Stuart 2010). It collapses all of the matching variables into a single number for each practice that can be used to assess how similar practices are to one another. By matching each treatment practice to one or more comparison practices with similar propensity scores, we generated a comparison group that is similar, on average, to the treatment group on the matching variables. The approach, however, does not ensure that each comparison practice matches exactly to its treatment practice on all matching variables. We specified that comparison practices had to match exactly to the treatment practices on two characteristics: whether the practice was a health center (including FQHCs and RHCs) and, for the health centers, whether the practice participated

in the CMS FQHC demonstration program because one of the treatment practices participated in this demonstration program.

We did not match the nonhealth center practices on one key variable we used in other awardee analyses in the HCIA-PCR portfolio—number of assigned beneficiaries. After consultation with CMMI, we chose not to use this characteristic for matching for the nonhealth centers because we did not have comparable data for the treatment and potential comparison practices on the providers working in practices. To determine the providers working in treatment practices, we used National Provider Identifier (NPI) data from WIPH and SK&A. However, for the comparison practices, we had only SK&A data to determine the NPIs of providers working in practices. We know that SK&A data do not contain an exhaustive list of NPIs. Consequently, we might be underidentifying providers in the potential comparison versus treatment groups, which would lead to underassignment of patients to practices. By requiring balance on the measured number of assigned beneficiaries, we could be forcing matches that are, in fact, not similar in patient panel size. Therefore, we decided to use the count of providers from SK&A data for both treatment and potential comparison practices for matching nonhealth centers, and we did not match on the number of assigned patients. Although SK&A might be undercounting providers, that undercount should be similar for both treatment and comparison practices, making the provider count from SK&A a valid matching variable.

We required each treatment practice to match to at least 1, and up to 10, comparison practices so that the total ratio of comparison-to-treatment practices be at least 3:1. This matching ratio increases the statistical certainty in the impact estimates (relative to 1:1 matching) because it creates a more stable comparison group against which to compare the treatment group's experiences.

After completing the matching, we assigned Medicare FFS beneficiaries to the comparison practices in each intervention quarter using the same rules we used for the treatment group (Section V.A.2).

4. Construction of outcomes and covariates

We used Medicare claims from January 1, 2009, to June 30, 2015, for beneficiaries assigned to the treatment and comparison practices to develop two types of variables: (1) outcomes, defined for each person in each baseline or intervention quarter; and (2) covariates, which describe a beneficiary's characteristics at the start of the baseline and intervention periods and are used in the regression models for estimating impacts to adjust for beneficiaries' characteristics before the period began. We used covariates defined at the start of each period, without updating them each quarter, to avoid controlling in each intervention quarter for previous quarters' program effects, as this would bias the effect estimates away from detecting true impacts. Appendix 1 provides details on the methods we used to construct these variables.

Outcomes. For each beneficiary, we calculated four outcomes that we grouped into three domains:

1. Domain: Quality-of-care outcomes

   a. Inpatient admissions (number/quarter) for ambulatory care-sensitive conditions (ACSCs)

2. Domain: Service use
   a. All-cause inpatient admissions (number/quarter)
   b. Outpatient ED visit rate (number/quarter); outpatient ED visits are defined as ED visits or observational stays that do not end in a hospital admission
3. Domain: Spending
   a. Total Medicare Part A and B spending (dollars/month)

Three of these outcomes—all but admissions for ACSCs—are outcomes that CMMI has specified as core for the evaluations of all HCIA programs. The fourth outcome that CMMI has specified as a core outcome, unplanned 30-day hospital readmissions, is not an outcome we assess in our evaluation because WIPH did not explicitly expect to affect readmissions with its PCMH intervention. All outcomes are quarter-specific, meaning that we calculated them for each baseline and intervention quarter separately.

Covariates. The covariates include (1) 18 indicators for whether a beneficiary had each of the following chronic conditions: heart failure, chronic obstructive pulmonary disease, chronic kidney disease, diabetes, Alzheimer's and related dementia, depression, ischemic heart disease, cancer, asthma, hypertension, atrial fibrillation, stroke, hyperlipidemia, hip fracture, osteoporosis, rheumatoid arthritis, bipolar disorder, and schizophrenia; (2) HCC score; (3) demographics (age, gender, and race or ethnicity); and (4) original reason for Medicare entitlement (old age, disability, or end-stage renal disease).

5. Regression model

We used a regression model to implement the difference-in-differences design for estimating impacts. For each outcome, the model estimates the relationship between the outcome and a series of predictor variables, assuming that each of the predictor variables has a linear (additive) relationship with the outcome. The predictor variables include the beneficiary-level covariates (defined in Section V.A.4); whether the beneficiary is assigned to a treatment or a comparison practice; an indicator for each practice (which accounts for differences between practices in their patients' outcomes at baseline); indicators for each post-intervention quarter; and an interaction of a beneficiary's treatment status with each post-intervention quarter.

The estimated relationship between the interaction term and the outcome in a given quarter is the impact estimate for that quarter. It measures the average difference between outcomes for beneficiaries assigned to the treatment and comparison practices during that period, subtracting out any differences between these groups during the four baseline quarters. By providing separate impact estimates for each intervention quarter, the model enables the program's impacts to change the longer the practices are enrolled in the program. We can also test impacts over discrete sets of quarters or years, which is needed to implement the primary tests discussed in the next section. Finally, the model quantifies the uncertainty in the impact estimates, allowing for statistical tests that determine whether observed differences in outcomes between the treatment

and comparison groups are likely due to chance. The model uses robust standard errors to account for clustering of outcomes across quarters for the same beneficiary and a dummy variable for each practice (fixed effects) to account for clustering of outcomes for beneficiaries assigned to the same practice. Appendix 2 provides details on the regression models, including descriptions of the weights each beneficiary receives in the model.

6. Primary tests

Table V.1 shows the primary tests for the WIPH PCMH intervention, by domain. Each test specifies a population, outcome, time period, expected direction of effect, and threshold that we count as substantively important. The purpose of these primary tests is to focus the impact evaluation on hypotheses that will provide the most robust evidence about program effectiveness. (See Appendix 3 for detail and a description of how we selected each test.) We provided both WIPH and CMMI an opportunity to comment on the primary tests.

Our rationale for selecting these primary tests is as follows:

- **Outcomes.** WIPH expected to reduce ED visits, hospitalizations, and spending (three of CMMI's four core outcomes) so our primary tests address these three outcomes. The intervention also expected to improve quality-of-care outcomes, including reducing hospitalizations for ACSCs, so our primary tests also address this outcome.
- **Time period.** WIPH did not specify a time period for intervention impacts. To provide time for the program to be implemented and diffused into practice, we chose to analyze impacts starting one year after the start of the program through the end of the intervention (that is, intervention quarters 5 through 10 [I5 through I10] for cohort one and I5 and I6 for cohort two).
- **Population.** Because WIPH expected to affect all patients served by the treatment practices, the population for our primary tests includes all Medicare FFS beneficiaries assigned to the treatment practices.
- **Direction (sign) of the impact estimate.** For each of the outcome measures, the primary tests are testing for a reduction, relative to the counterfactual—defined as the outcomes that beneficiaries in the treatment group would have had if they had not received the HCIA-funded intervention.

Table V.1. Specification of the primary tests for WIPH

Domain (number of tests in the domain)[a]	Outcome (units)	Time period for impacts (controlling for baseline differences)[b]	Population	Substantive threshold (expected direction of effect)[c,d]
Quality-of-care outcomes (1)	Inpatient admissions for ambulatory care-sensitive conditions (#/beneficiary/quarter)	Average over I5 through I10 for cohort one and I5 and I6 for cohort two	Medicare FFS beneficiaries assigned to treatment practices	5.00% (-)
Service use (2)	All-cause inpatient admissions (#/beneficiary/quarter)	Average over I5 through I10 for cohort one and I5 and I6 for cohort two	Medicare FFS beneficiaries assigned to treatment practices	3.75% (-)
	Outpatient ED visit rate (#/beneficiary/quarter)	Average over I5 through I10 for cohort one and I5 and I6 for cohort two	Medicare FFS beneficiaries assigned to treatment practices	5.00% (-)
Spending (1)	Medicare Part A and B FFS spending ($/beneficiary/month)	Average over I5 through I10 for cohort one and I5 and I6 for cohort two	Medicare FFS beneficiaries assigned to treatment practices	3.75% (-)

[a] We adjust the *p*-values from the primary test results for the multiple comparisons made within each domain, but not across domains.

[b] The regression models for estimating program impacts control for differences in baseline outcomes between the treatment and comparison groups.

[c] For all-cause inpatient admissions and spending, we set the substantive threshold to 75 percent of WIPH's expected effect. For outpatient ED visits and inpatient admissions for ambulatory care-sensitive conditions, we set the substantive threshold based on evidence from the literature (Peikes et al. 2011) about what is feasible among beneficiaries in a patient-centered medical home.

[d] The substantive threshold is expressed as a percentage of the counterfactual. The counterfactual is the outcome the treatment group would have had in the absence of the HCIA-funded intervention.

ED = emergency department; FFS = fee-for-service; I = intervention quarter; WIPH = Wyoming Institute of Population Health.

- **Substantive thresholds.** Some impact estimates could be large enough to be substantively interesting (to CMMI and other stakeholders) even if they are not statistically significant; for this reason, we prespecified thresholds for what we call substantive importance. We express the threshold as a percentage change from the counterfactual. WIPH expected a 10 percent reduction in the ED visit rate, a 5 percent reduction in the all-cause inpatient admission rate, and a 5 percent reduction in total spending. For the all-cause inpatient admission rate and total spending, the substantive thresholds we chose are 75 percent of WIPH's stated goals and are therefore set at 3.75 percent. We chose 75 percent of WIPH's goal recognizing that a program could still be promising even if it did not fully achieve its anticipated effect. We set the substantive threshold for the ED visit rate to be 5 percent, based on evidence from the literature (Peikes et al. 2011). We used this threshold because the literature suggests effects of this size are policy-relevant, even though they are smaller than WIPH's anticipated impact (of 10 percent). Given that WIPH did not explicitly set goals for ACSC admissions, our threshold for this outcome is also based on evidence from the literature (Peikes et al. 2011). The threshold is set to 5 percent.

7. Secondary tests (robustness checks)

We also conducted secondary quantitative tests to help corroborate the findings from the primary tests. This is important because some of the differences observed between the treatment and comparison groups in the primary test results could reflect limitations of the non-experimental impact evaluation design or random fluctuations in the data. We have greater confidence in the primary results if they are generally consistent with the expected broader pattern of results from the secondary tests.

We conducted three sets of secondary tests for WIPH:

1. We estimated the PCMH program component's impacts on the four outcomes during two additional intervention periods: (1) the first 6 months after the practices joined the intervention (I1 and I2) and (2) months 7 to 12 (I3 and I4). Because we expected few or no impacts in the first few months of the program as practices implemented the intervention, the following pattern would be highly consistent with an effective program—few to no measured effects in the first two quarters, growing effects in I3 and I4, and the largest impacts in I5 through I10 (the period for the primary tests). In contrast, large differences in outcomes (favorable or unfavorable) in the first year of the program (that is, I1 to I4) could suggest a limitation in the comparison group, not true program impacts.
2. We reran all of the primary tests, limiting the sample only to Medicare FFS beneficiaries assigned to the treatment and comparison groups by the start of the period, either baseline or intervention. This restriction prevents addition to the intervention sample over time. It is possible that differences in sample addition between the treatment and comparison groups could bias the impact results to some degree if the sample members added over time differ from earlier sample members (for example, they are younger and healthier); this could create differences in mean outcomes between the treatment and comparison groups that are unrelated to the HCIA intervention. We have explored this possibility because, as we described in the second annual report summary for WIPH (Ehrlich et al. 2015), it is possible

that WIPH's four other intervention components influenced the composition of the PCMH treatment group in ways that made PCMH impacts appear to be unfavorable. Specifically, one of the goals of WyRCT was to connect recently hospitalized patients to primary care. This transitional care program could have led, on average, to the assignment of sicker beneficiaries (who had recently been hospitalized) to the treatment practices (relative to beneficiaries assigned to the comparison practices), making it appear that outcomes for the treatment group were worse than those for the comparison group.

3. Last, we examined how many of our matched comparison practices received NCQA PCMH recognition or payments for meaningful use of EHRs during the first year of the intervention to assess whether practices in Montana were on a different trajectory of practice transformation and quality improvement that we could not detect at baseline but that could affect patients' outcomes. If more comparison practices received NCQA PCMH recognition or payments for meaningful use of EHRs during the first year of the intervention, this suggests that there were unobservable differences between the treatment and comparison groups that could affect our results. This provides more evidence about whether the selected comparison group provides a reasonable estimate of the counterfactual.

8. Synthesizing evidence to draw conclusions

Within each domain, we planned to draw one of five conclusions about program effectiveness based on the primary test results, the results of secondary tests, and the plausibility of those findings given the implementation evidence:

1. Statistically significant favorable effect (the highest level of evidence)
2. Substantively important (but not statistically significant) favorable effect
3. Substantively important (but not statistically significant) unfavorable effect
4. No substantively large effect
5. Indeterminate effect

We could not conclude that a program had a statistically significant unfavorable effect because, in consultation with CMMI, we decided to use one-sided statistical tests (which do not test for evidence of unfavorable effects). We used one-sided tests to increase the probability that, if a program truly did have impacts, we would be able to detect them.

Appendix 3 describes our decision rules for each of the five possible conclusions. In short, we concluded that a program had a statistically significant favorable effect in a domain if (1) at least one primary test result in the domain was favorable and statistically significant, after adjusting the statistical tests to account for multiple tests (if applicable) within a domain; or (2) the average impact estimate across all primary tests in the domain was favorable and statistically significant. In both cases, we also had to determine that the primary test results were plausible given the results of the secondary tests and implementation evidence. We concluded that a program had a substantively important favorable effect if the average impact estimate in the domain was substantively important but not statistically significant, and if the result was plausible given the secondary tests and implementation evidence. In contrast, if the average

impact estimate was unfavorable (opposite the hypothesized direction), larger than the substantive threshold, and unfavorable effects were plausible given the other evidence, we concluded the program had a substantively important unfavorable effect. If the tests in a domain did not meet any of these criteria, we instead used the following rules. First, if the tests for at least one outcome in the domain (or all outcomes in the domain together) had sufficient statistical power to detect an impact of the size of the substantive threshold with at least 75 percent probability, we concluded there was not a substantively large effect because we are reasonably confident that we would have detected such an effect had there been one. Second, if the power was not sufficient (less than 75 percent) to detect this type of impact, we concluded the impact in the domain was indeterminate. Indeterminate means either that the program truly did not have effects that were substantively large, or that it did, but our statistical tests were not able to detect them. Finally, if the results for the primary tests in a domain were not plausible given the implementation evidence or the secondary, corroborating tests, we did not draw any conclusions about program impacts in that domain.

B. Characteristics of the treatment group at baseline

This section describes the characteristics of the treatment group at the start of the intervention period (January 1, 2013, for cohort one and January 1, 2014, for cohort two). We also show this information in the second column of Table V.2. (Table V.2 serves a second purpose—to show the equivalence of the treatment and comparison panels at the start of the intervention—which we describe in Section V.C.)

Characteristics of the practices overall. Our analysis includes 20 treatment practices, 7 of which are FQHCs or RHCs. Most treatment practices (75 percent) were located in a zip code considered an urban area or urbanized cluster or in a primary care health shortage area. The 13 nonhealth center practices, on average, consisted of approximately four providers, with 95 percent of these providers having a primary care specialty. A hospital or health system owned a quarter of the nonhealth center practices and almost a third had providers who received payments from CMS for meaningful use of EHRs (30 percent) in the baseline period. None of the practices had any level of NCQA PCMH recognition in the baseline period, consistent with the fact that a key aim of the WIPH PCMH intervention was to facilitate practices becoming NCQA-recognized medical homes.

Characteristics of the practices' Medicare FFS beneficiaries. The characteristics of all Medicare FFS beneficiaries assigned to the treatment practices during the baseline period (January 1, 2012, through December 31, 2012, for cohort one and January 1, 2013, through December 31, 2013, for cohort two) were similar to the nationwide FFS averages on some but not all characteristics. The average HCC risk score for the treatment group (0.97) was slightly lower than the national average (1.00). Beneficiaries in the treatment practices had hospital admission rates close to the national average. Medicare Part A and B spending and the 30-day unplanned readmission rates were lower than the national average, but ED visit rates and inpatient admissions for ACSCs were higher. The higher ED visit rate and ACSC admissions might reflect the fact that the treatment practices served a population in which primary care access is limited, leading to higher ED and inpatient use.

Table V.2. Characteristics of treatment and comparison practices before the intervention start date (January 1, 2013, [cohort one] or January 1, 2014 [cohort two])

Characteristic of practice	Treatment practices (N = 20)	Matched comparison group (N = 75)	Absolute difference[a]	Standardized difference[b]	Medicare FFS national average
Exact match variables[c]					
Characteristics of the practices overall					
Health center (%)	35.0	35.0	0	0	n.a.
Participating in the FQHC demonstration (%)	5.0	5.0	0	0	n.a.
NCQA PCMH recognition (%)[d]	0	0	0	0	n.a.
Cohort one (%)	0.9	0.9	0	0	n.a.
Propensity matched variables[e]					
Characteristics of the practice's location					
Located in an urban zip code (%)[f]	75.0	73.2	1.8	0.04	n.a.
Located in a health professionals shortage area (primary care) (%)	75.0	67.5	7.5	0.16	n.a.
Characteristics of all beneficiaries attributed to practices during the baseline year (January 1 to December 31, 2012, or January 1 to December 31, 2013)					
Number of beneficiaries[g]	607.8	472.3	135.5	0.28	n.a.
HCC risk score	0.97	0.99	-0.02	-0.14	1.0
All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	71.9	71.5	0.5	0.02	74[h]
Outpatient ED visit rate (#/1,000 beneficiaries/quarter)	151.9	163.9	-11.96	-0.15	105[i]
Medicare Part A and B spending ($/beneficiary/month)	765	737	29	0.13	860[j]
30-day unplanned hospital readmission rate (%)	12.4	12.0	0.3	0.06	16.0[k]
Inpatient admissions for ambulatory care-sensitive conditions (#/1,000 beneficiaries/quarter)	14.3	14.1	0.1	0.01	11.8[l]
Disability as original reason for Medicare entitlement (%)	21.2	21.9	-0.8	-0.07	16.7[m]
Dually eligible for Medicare and Medicaid (%)	17.0	17.4	-0.5	-0.04	22[n]
Age (years)	71.7	71.7	-0.1	-0.01	71[o]
Female (%)	56.6	55.7	0.9	0.09	54.7[m]
Characteristics of the practices (nonhealth centers only)[p]					
Providers in practice, according to SK&A (#)	3.7	3.9	-0.2	-0.08	n.a.
Providers in practice with primary care specialty (%)	95.0	95.4	-0.4	-0.03	n.a.
Owned by a hospital or health system (%)	25.0	25.5	-0.5	-0.01	n.a.
Meaningful use of EHRs (%)[q]	30.0	29.7	0.3	0.01	n.a.

Sources: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at CMS. Zip code information (whether an urbanized area or cluster or health professionals shortage area) was merged from the American Community Survey ZIP Code Characteristics and the Area Resource File, respectively. Data on practices with NCQA recognition were merged from the NCQA database. Data on meaningful use of EHRs were merged from CMS data.

Table V.2 *(continued)*

Notes: The comparison group means are weighted based on the number of matched practices per treatment practice. For example, if four comparison practices are matched to one treatment practice, each of the four comparison practices has a matching weight of 0.25.

Absolute differences might not be exact due to rounding.

[a] The absolute difference is the difference in means between the matched treatment and comparison groups.

[b] The standardized difference is the difference in means between the matched treatment and comparison groups divided by the standard deviation of the variable, which is pooled across the matched treatment and matched comparison practices.

[c] Exact match means that a treatment practice could match only to a comparison practice (or practices) that had an identical value for the matching variable. A health center had to be matched to a health center and a nonhealth center had to be matched to a nonhealth center. We also exact matched health centers on whether they participated in the FQHC demonstration program, and exact matched all practices on whether they had achieved NCQA PCMH recognition during the baseline period. Last, we exact matched practices in each cohort only to other practices observed at the same time.

[d] As described in the text, the potential comparison pool was limited to practices that did not have NCQA recognition at the start of the intervention.

[e] We matched practices on these variables through propensity scores.

[f] Zip codes classified as urban for matching included those categorized as urbanized areas (defined as 50,000 or more people) or urban clusters (defined as at least 2,500 and less than 50,000 people).

[g] We did not include the number of attributed beneficiaries in our propensity score model for nonhealth centers, but we did use this as a matching variable for health centers. This measure is reported in the table for all practices (health centers and nonhealth centers) for descriptive purposes even though it was not included in the matching model for nonhealth centers. We chose not to include this variable for matching nonhealth centers because we had differing data sources for the treatment and comparison practices on the number of providers working in these practices. Therefore for nonhealth centers, we matched on the number of providers working in practices, as counted through SK&A data, and not on the number of attributed beneficiaries. Because we explicitly did not match nonhealth centers on this variable, we accepted a standardized difference of 0.28 for this characteristic across all practices, which was above our maximum difference for balance.

[h] Health Indicators Warehouse (2014b).

[i] Gerhardt et al. (2014).

[j] Boards of Trustees (2013).

[k] Centers for Medicare & Medicaid Services (2014).

[l] This is the rate for all individuals ages 65 and older. Truven Health Analytics (2015).

[m] Chronic Conditions Data Warehouse (2016, Table A.1).

[n] Health Indicators Warehouse (2014c).

[o] Health Indicators Warehouse (2014a).

[p] The 20 treatment practices include 13 nonhealth centers. There were 251 nonhealth centers in the unmatched comparison pool and 49 nonhealth centers in the matched comparison group.

[q] Meaningful use of EHRs is calculated as the percentage of practices with at least one provider (NPI) working in the practice who received financial incentives for meaningful use of certified EHRs through Medicare or Medicaid during the baseline period.

CMS = Centers for Medicare & Medicaid Services; ED = emergency department; EHR = electronic health record; FFS = fee-for-service; FQHC = federally qualified health center; HCC = Hierarchical Condition Category; NCQA = National Committee for Quality Assurance; NPI = National Provider Identifier; PCMH = patient-centered medical home.

n.a. = not applicable.

C. Equivalence of treatment and comparison groups at baseline

Demonstrating that the treatment and comparison groups are similar at the start of the intervention is important for the evaluation design. This similarity increases the credibility of a key assumption underlying difference-in-differences models—that the change over time in outcomes for the comparison group is the same change that would have happened for the treatment group, had the treatment group not received the intervention.

Table V.2 shows that the 20 treatment practices and the 75 selected comparison practices were similar at the start of the intervention on most matching variables. By construction, there were no differences between the two groups on whether the practice was a health or a nonhealth center, or whether the practice was participating in the CMS FQHC demonstration (applicable to FQHCs only). There were some differences between the treatment group practices and matched comparison practices on the variables included in the propensity score model, but all but one of the standardized differences across the propensity score matching variables are within our target of 0.25 standardized differences, and most are within 0.15 standardized differences (the 0.25 target is an industry standard; for example, see Institute for Education Sciences [2014]).

The differences for one variable, the number of attributed Medicare FFS beneficiaries, were 0.28 standardized differences. On average, the 13 nonhealth center treatment practices had more attributed Medicare FFS beneficiaries, overall (by 136 beneficiaries). However, as described earlier, we—in consultation with CMMI—decided not to require balance within 0.25 standardized differences on this variable. We decided that it was reasonable to accept the comparison group because (1) we can account for differences in practice size through regression weights in our impact analyses and (2) if any systematic differences in outcomes (that do not vary over time) result from a different number of beneficiaries, the difference-in-differences model would account for them.

D. Beneficiaries' outcomes and intervention impacts

In this section, we first present sample sizes and mean outcomes, by quarter, for the treatment and comparison groups. These mean outcomes provide context for understanding the difference-in-differences estimates that follow; however, the differences in mean outcomes are not regression-adjusted and not impact estimates by themselves. Next, we present the results of the primary tests, by domain. Then, we present the results of the secondary tests (robustness checks) and assess whether the primary test results are plausible given the secondary test results and the implementation evidence. We end with a discussion of why we were unable to draw conclusions in any of the study domains.

1. Sample sizes

In the baseline period, the treatment group ranged from 10,597 (in the first baseline quarter, B1) to 13,670 (in the last baseline quarter, B4) beneficiaries (see Table V.3). The comparison group included 33,542 to 37,527 unweighted beneficiaries during the same period. The sample size for the treatment group dropped from the last baseline quarter to the first intervention quarter from 13,670 to 12,990 beneficiaries (because beneficiaries no longer attributed to the treatment practices were dropped from the sample at that time; see Section V.A.2). The sample

then grew steadily again during the following five intervention quarters to 16,373 beneficiaries for the same reason it grew in the baseline period. Because cohort 2 practices did not have data available for the last four intervention quarters, the sample dropped to 14,300 beneficiaries in I7 but rose steadily again to 15,458 beneficiaries in I10. The comparison group followed the same pattern during the intervention period, with the unweighted sample ranging from 35,155 (I1) to 39,156 (I6) to 35,925 (I10).

2. Mean outcomes for the treatment and comparison groups, by domain and quarter

Table V.3 presents unadjusted mean outcomes for the treatment and comparison groups.

Quality-of-care outcomes. Inpatient admissions for ACSCs were higher for the treatment group than the comparison group for two of the four baseline quarters and across most intervention quarters. For the baseline and intervention periods, ACSC admissions for the treatment group were highest in the first quarter of each period (B1 and I1).

Service use. Inpatient admissions were higher for the treatment group than the comparison group for two of the four baseline quarters and for all intervention quarters. Inpatient admissions fluctuated but generally declined over time for the comparison group. The ED visit rates for the treatment group were lower than for the comparison group in all quarters. The treatment group had a slightly increasing trend over time, whereas the comparison group had more fluctuation.

Spending. Aside from B1 and B4, spending was higher across all quarters for the treatment group than the comparison group. Spending for the treatment group had an increasing trend over time, whereas the comparison group showed some decline toward the beginning of the intervention period but then increased again and remained steady toward the end of the intervention period.

3. Results for primary tests, by domain

Overview. The primary tests suggest substantively large unfavorable effects for the service use and spending domains and indeterminate effects for the quality-of-care outcomes domain (Table V.4).

Table V.3. Unadjusted mean outcomes (quality-of-care outcomes, service use, and spending) for Medicare FFS beneficiaries, by treatment status and quarter

	Number of Medicare FFS beneficiaries (practices)			Inpatient admissions for ambulatory care-sensitive conditions (#/1,000 beneficiaries/quarter)			All-cause inpatient admissions (#/1,000 beneficiaries/quarter)			Outpatient ED visit rate (#/1,000 beneficiaries/quarter)			Medicare Part A and B spending ($/beneficiary/month)		
Q	T	C (no wgt)	C (wgt)	T	C	Diff (%)	T	C	Diff (%)	T	C	Diff (%)	T	C	Diff (%)
Baseline period (January 1 to December 31, 2012, or January 1 to December 31, 2013)															
B1	10,597 (20)	33,542 (75)	11,272	17.2	14.6	2.6 (17.8%)	77.0	82.5	-5.5 (-6.7%)	144.4	171.3	-26.9 (-15.7%)	$763	$821	$-58 (-7.0%)
B2	11,701 (20)	35,126 (75)	11,812	14.8	15.4	-0.7 (-4.3%)	77.1	73.1	4.0 (5.5%)	147.3	198.7	-51.5 (-25.9%)	$801	$753	$47 (6.3%)
B3	12,725 (20)	36,410 (75)	12,446	12.6	17.7	-5.1 (-28.9%)	73.2	79.3	-6.1 (-7.7%)	142.5	191.3	-48.8 (-25.5%)	$777	$738	$39 (5.4%)
B4	13,670 (20)	37,527 (75)	13,163	15.3	14.0	1.3 (9.2%)	74.6	70.4	4.3 (6.1%)	146.3	172.0	-25.7 (-15.0%)	$792	$793	$-1 (-0.1%)
Intervention period (January 1, 2013, to June 30, 2015, or January 1, 2014, to June 30, 2015)															
I1	12,990 (20)	35,155 (75)	12,152	19.6	14.3	5.4 (37.5%)	81.6	77.8	3.8 (4.9%)	132.3	149.1	-16.8 (-11.2%)	$830	$733	$97 (13.2%)
I2	13,873 (20)	36,709 (75)	12,922	13.8	12.3	1.6 (12.8%)	69.5	67.9	1.6 (2.3%)	137.8	155.8	-18.0 (-11.6%)	$796	$704	$92 (13.1%)
I3	14,634 (20)	37,824 (75)	13,480	13.5	10.7	2.8 (26.3%)	73.7	65.9	7.9 (11.9%)	141.8	172.1	-30.2 (-17.6%)	$791	$719	$73 (10.1%)
I4	15,247 (20)	38,637 (75)	13,905	15.2	12.8	2.3 (18.3%)	73.5	62.5	11.0 (17.6%)	137.1	165.9	-28.7 (-17.3%)	$833	$717	$116 (16.1%)
I5	15,757 (20)	38,610 (75)	14,073	13.8	16.1	-2.2 (-14.0%)	72.2	70.1	2.1 (3.0%)	135.8	154.6	-18.9 (-12.2%)	$788	$688	$100 (14.5%)
I6	16,373 (20)	39,156 (75)	14,330	14.1	12.0	2.1 (17.6%)	77.4	68.8	8.6 (12.5%)	140.1	167.4	-27.3 (-16.3%)	$874	$782	$92 (11.8%)
I7	14,300 (18)	35,258 (69)	12,696	13.3	12.4	0.9 (7.1%)	72.7	69.4	3.3 (4.8%)	158.9	185.0	-26.1 (-14.1%)	$858	$763	$96 (12.5%)
I8	14,825 (18)	35,676 (69)	12,969	14.8	11.0	3.8 (34.4%)	71.4	65.3	6.1 (9.3%)	146.4	169.6	-23.2 (-13.7%)	$823	$746	$77 (10.3%)
I9	15,110 (18)	35,533 (69)	12,994	15.4	17.0	-1.6 (-9.5%)	77.2	68.8	8.5 (12.3%)	162.3	165.8	-3.5 (-2.1%)	$884	$747	$137 (18.4%)
I10	15,458 (18)	35,925 (69)	13,196	15.1	13.1	2.1 (15.7%)	77.2	64.7	12.5 (19.3%)	149.0	175.7	-26.7 (-15.2%)	$877	$745	$133 (17.8%)

Sources: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at the Centers for Medicare & Medicaid Services.

Table V.3 *(continued)*

Notes: The baseline quarters are measured relative to the start of the baseline period on January 1, 2012, for cohort one or January 1, 2013, for cohort two. For example, the first baseline quarter (B1) for cohort one runs from January 1 to March 31, 2012. The intervention quarters are measured relative to the start of the intervention period on January 1, 2013, for cohort one or January 1, 2014, for cohort two. For example, the first intervention quarter (I1) runs from January 1 to March 31, 2013, for cohort one. In each period (baseline or intervention), the treatment group each quarter includes all beneficiaries who were assigned to a treatment practice by the start of the quarter and who met other sample criteria—that is, they were alive, enrolled in FFS Medicare, and were living in Wyoming, Nebraska, or Montana. In each period, the comparison group includes all beneficiaries who were assigned to a comparison practice by the start of the quarter and who met the other sample criteria. See text for details.

The outcome means were weighted such that (1) each treatment beneficiary gets a weight of 1; and (2) each comparison beneficiary gets a weight that is the product of two weights: (1) a matching weight, equal to the reciprocal of the total number of comparison practices matched to the same treatment practice as the beneficiary's assigned practice; and (2) a practice size weight, which equals the average number of beneficiaries assigned to the matched treatment practice during the four baseline quarters divided by the average number of beneficiaries assigned to the beneficiary's comparison practice over those quarters. The difference between the treatment and comparison groups in a quarter is calculated by subtracting the mean outcome for the comparison group from the mean outcome for the treatment group. The percentage difference equals that difference divided by the mean outcome for the comparison group.

B = baseline; C = comparison; Diff = difference; ED = emergency department; FFS = fee-for-service; I = intervention; Q = quarter; T = treatment; no wgt = unweighted; wgt = weighted.

Table V.4. Results of primary tests for WIPH

Primary test definition					Statistical power to detect an effect that is[a]		Results			
Domain (# of tests in domain)	Outcome (units)	Time period for impacts	Population	Substantive threshold (expected direction of effect)	Size of the substantive threshold	Twice the substantive threshold[b]	Treatment group mean	Regression-adjusted difference between the treatment and estimated counterfactual[c] (standard error)	Percentage difference[d]	*p*-value[e]
Quality-of-care outcomes (1)	Inpatient admissions for ambulatory care-sensitive conditions (#/1,000 beneficiaries/quarter)	Average over intervention quarters 5–10 for cohort one practices; average over intervention quarters 5 and 6 for cohort two practices	All Medicare FFS beneficiaries assigned to treatment practices	5.00% (-)	18.0%	29.2%	14.4	0.1 (2.0)	0.7%	0.52
Service use (2)	All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	Average over intervention quarters 5–10 for cohort one practices; average over intervention quarters 5 and 6 for cohort two practices	All Medicare FFS beneficiaries assigned to treatment practices	3.75% (-)	23.0%	42.2%	74.7	4.6 (4.9)	6.5%	0.72
	Outpatient ED visits (#/1,000 beneficiaries/quarter)	Average over intervention quarters 5–10 for cohort one practices; average over intervention quarters 5 and 6 for cohort two practices	All Medicare FFS beneficiaries assigned to treatment practices	5.00% (-)	15.5%	22.6%	148.7	16.3 (25.1)	12.3%	0.63
	Combined (%)	Average over intervention quarters 5–10 for cohort one practices; average over intervention quarters 5 and 6 for cohort two practices	All Medicare FFS beneficiaries assigned to treatment practices	4.38% (-)	17.6%	28.0%	n.a.	n.a.	9.4%	0.77
Spending (1)	Medicare Part A and B spending ($/beneficiary/month)	Average over intervention quarters 5–10 for cohort one practices; average over intervention quarters 5 and 6 for cohort two practices	All Medicare FFS beneficiaries assigned to treatment practices	3.75% (-)	27.0%	52.3%	$851	$73 ($44)	9.4%	0.95

Table V.4 *(continued)*

Source: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at the Centers for Medicare & Medicaid Services.

Note: The results for each outcome are based on a difference-in-differences regression model, as described in the text. Estimates are calculated for Medicare beneficiaries who are observable in the relevant time period: that is, beneficiaries who are enrolled in Medicare FFS (Part A and B), are alive, and have Medicare as their primary payer.

[a] The power calculation is based on actual standard errors from the analysis. For example, in the last row, a 3.75 percent effect on Medicare Part A and B spending (from the counterfactual of $851 + $73 = $924) would be a change of $35. Given the standard error of $44 from the regression model, we would be able to detect a statistically significant result 27.1 percent of the time if the impact was truly -$35, assuming a one-sided statistical test at the $p = 0.10$ significance level.

[b] We show statistical power to detect a very large effect (twice the size of the substantive threshold), because this provides additional information about the likelihood that we will find effects if the program is indeed effective. If power to detect effects is less than 75 percent even for a very large effect, then the evaluation is extremely poorly powered for that outcome.

[c] The counterfactual is the outcome the treatment group would have had in the absence of the HCIA-funded intervention. Our estimate of the counterfactual is the treatment group mean minus the regression-adjusted difference-in-differences estimate.

[d] Percentage difference is calculated as the regression-adjusted difference between the treatment and comparison group, divided by the adjusted comparison group mean.

[e] *p*-values test the null hypothesis that the regression-adjusted difference-in-differences estimate is greater than or equal to zero (a one-sided test). Because it is a one-sided test, as the difference-in-differences estimate approaches positive infinity, the *p*-value approaches 1, whereas it would approach 0 in a two-sided test. We adjusted the *p*-values from the primary test results for the multiple (two) comparisons made within the service use domain.

ED = emergency department; FFS = fee-for-service; HCIA = Health Care Innovation Award; WIPH = Wyoming Institute for Population Health.

Quality-of-care outcomes. The treatment group's average number of inpatient admissions for ACSCs was 14.4 per 1,000 beneficiaries per quarter during the primary test period, which was estimated to be 0.1 more admissions than the counterfactual. (Our estimated counterfactual—the outcome the treatment group members would have had in the absence of the HCIA intervention—is the treatment group mean minus the difference-in-differences estimate.) This was a 0.7 percent unfavorable difference, which is less than the substantive threshold of 5.0 percent. We cannot assess the statistical significance of this difference because we used one-sided statistical tests, testing only for favorable effects. The statistical power values in Table V.4 show that our analysis had limited power to detect differences in ACSC admissions: 18.0 percent to detect a difference the size of the substantive threshold, and 29.2 percent to detect an effect twice the size of the substantive threshold.

Service use. The treatment group's average number of all-cause inpatient admissions was 74.7 per 1,000 beneficiaries per quarter during the primary test period. This was 4.6 more admissions per 1,000 beneficiaries per quarter than the estimated counterfactual. This is a 6.5 percent unfavorable difference, which is greater than the substantive threshold of 3.75 percent. For ED visits, the treatment group averaged 148.7 visits per 1,000 beneficiaries per quarter during the primary test period, which represents a 12.3 percent unfavorable difference—also higher than the substantive threshold of 5.0 percent. The mean percentage difference across all-cause inpatient admissions and ED visits was 9.4 percent (the average of 6.5 percent and 12.3 percent). We adjusted the *p*-values for both tests for the multiple statistical tests in this domain. However, as with ACSC admissions, we cannot assess whether these unfavorable differences are statistically significant because we tested only for favorable effects. We had poor statistical power to detect true impacts the size of the substantive thresholds.

Spending. Medicare Part A and B spending per beneficiary per month averaged $851 for the treatment group during the primary test period, which was estimated to be $73 higher per beneficiary per month than the counterfactual. This 9.4 percent unfavorable difference is greater than the substantive threshold of 3.75 percent. As with the other domains, we cannot assess statistical significance, and power to detect effects was poor.

4. Results for secondary tests

Estimates during the first intervention year (January 1, 2013, to December 31, 2013 for cohort one and January 1, 2014, to December 31, 2014 for cohort two). Results from the secondary tests indicate unfavorable effects across all outcomes during I1 through I4 (Table V.5). Given that the treatment and comparison groups were well matched at baseline (Table V.2) and treatment group outcomes did not worsen substantially over time (Table V.3), both the primary and secondary test results suggest that outcomes for the comparison group improved faster than those for the treatment group during the intervention period. *Ex ante,* we would have expected little to no change in outcomes for both treatment and comparison groups in the first year after the intervention started. The large unfavorable effects early in the intervention period diminish our confidence in the comparison group as a reasonable representation of the counterfactual.

Estimates limiting the sample to prevent sample addition. The secondary test results limited to those beneficiaries attributed at the start of the baseline or intervention period (Table V.5) are generally consistent with the primary test results. The results show substantively large unfavorable effects across three of the four outcomes for the treatment group (inpatient admissions, ED visits, and spending). The effect sizes for these outcomes are larger in magnitude than the effect sizes from the primary tests. For the fourth outcome, ACSC admissions, the results for this sample are in the favorable direction, but they are not substantively important or statistically significant. If sicker beneficiaries were being added to the treatment group over time because of the WIPH WyRCT program component, we would expect these secondary tests to show more favorable results than those from the primary tests. In contrast, the substantively large unfavorable effects for most outcomes suggest that WyRCT is not driving the primary test results as hypothesized, and these secondary test results further suggest a limitation in the comparison group.

Results for NCQA PCMH recognition and EHR Meaningful Use during the first intervention year (January 1, 2013, to December 31, 2013 for cohort one and January 1, 2014, to December 31, 2014, for cohort two). There is some evidence that comparison practices were on a different path to quality improvement compared with treatment practices, as reflected by the number of comparison practices that obtained NCQA PCMH recognition within the first year of the intervention (0 percent of treatment practices versus 16 percent of comparison practices; results not shown). Before matching, we limited our potential comparison pool to practices that did not have NCQA PCMH recognition in the baseline period, so the differences between the groups in the first year of the intervention were not observed in the baseline period. Similarly, among practices that were not meaningful users of EHRs in the baseline period, there were differences between treatment and comparison practices in the first year of the intervention. A higher proportion of treatment practices than comparison practices became new meaningful users during this period (15 percent of treatment practices versus 8 percent of comparison practices; results not shown). Although this practice change might be expected to improve outcomes, it might not lead to changes in patients' outcomes that could be observed during the impact evaluation period (I5 to I10 for cohort one and I5 and I6 for cohort two). Overall, the results for NCQA PCMH recognition and EHR meaningful use provide further evidence that unobserved differences between the treatment and comparison groups might have biased patients' outcomes during our evaluation period.

Table V.5. Results of secondary tests for WIPH

Secondary test definition				Results			
Domain	Outcome (units)	Time period for impacts	Population	Treatment group mean	Regression-adjusted difference between treatment and the estimated counterfactual[a] (standard error)	Percentage difference[b]	*p*-value[c]
Estimates during the first intervention year (January 1, 2013–December 31, 2013, or January 1, 2014–December 31, 2014)							
Quality-of-care outcomes	Inpatient admissions for ambulatory care sensitive conditions (#/1,000 beneficiaries/quarter)	Intervention quarters 1,2	All Medicare FFS beneficiaries assigned to treatment practices	16.7	2.5 (2.2)	17.7%	0.87
	Inpatient admissions for ambulatory care sensitive conditions (#/1,000 beneficiaries/quarter)	Intervention quarters 3,4	All Medicare FFS beneficiaries assigned to treatment practices	14.3	1.7 (2.3)	13.6%	0.78
Service use	All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	Intervention quarters 1, 2	All Medicare FFS beneficiaries assigned to treatment practices	75.5	1.3 (5.6)	1.7%	0.59
	All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	Intervention quarters 3, 4	All Medicare FFS beneficiaries assigned to treatment practices	73.6	8.0 (5.2)	12.1%	0.94
Service use	Outpatient ED visits (#/1,000 beneficiaries/quarter)	Intervention quarters 1,2	All Medicare FFS beneficiaries assigned to treatment practices	135.1	16.9 (24.5)	14.3%	0.75
	Outpatient ED visits (#/1,000 beneficiaries/quarter)	Intervention quarters 3,4	All Medicare FFS beneficiaries assigned to treatment practices	139.5	5.1 (25.9)	3.8%	0.58
Spending	Medicare Part A and B spending ($/beneficiary/month)	Intervention quarters 1, 2	All Medicare FFS beneficiaries assigned to treatment practices	$813	$71 (48)	9.6%	0.93
	Medicare Part A and B spending ($/beneficiary/month)	Intervention quarters 3, 4	All Medicare FFS beneficiaries assigned to treatment practices	$812	$68 (46)	9.2%	0.93

Table V.5 *(continued)*

Secondary test definition				Results			
Domain	Outcome (units)	Time period for impacts	Population	Treatment group mean	Regression-adjusted difference between treatment and the estimated counterfactual[a] (standard error)	Percentage difference[b]	*p*-value[c]
Estimates limiting the sample to prevent sample addition after the first baseline or intervention quarter							
Quality-of-care outcomes	Inpatient admissions for ambulatory care sensitive conditions (#/1,000 beneficiaries/quarter)	Intervention quarters 5–10 for cohort one; intervention quarters 5–6 in cohort two	Medicare FFS beneficiaries assigned to treatment practices in the first baseline or first intervention quarter	15.1	-0.5 (2.1)	-3.4%	0.40
Service use	All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	Intervention quarters 5–10 for cohort one; intervention quarters 5–6 in cohort two	Medicare FFS beneficiaries assigned to treatment practices in the first baseline or first intervention quarter	75.7	10.2 (5.2)	15.6%	0.98
Service use	Outpatient ED visits (#/1,000 beneficiaries/quarter)	Intervention quarters 5–10 for cohort one; intervention quarters 5–6 in cohort two	Medicare FFS beneficiaries assigned to treatment practices in the first baseline or first intervention quarter	149.9	28.9 (27.9)	23.9%	0.85
Spending	Medicare Part A and B spending ($/beneficiary/month)	Intervention quarters 5–10 for cohort one; intervention quarters 5–6 in cohort two	Medicare FFS beneficiaries assigned to treatment practices in the first baseline or first intervention quarter	$856	$120 (45)	16.2%	1.00

Source: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at the Centers for Medicare & Medicaid Services.

Notes: The results for each outcome are based on a difference-in-differences regression model, as described in the text. Estimates are calculated for Medicare beneficiaries who are observable in the relevant time period: that is, beneficiaries who are enrolled in Medicare FFS (Part A and B), are alive, and have Medicare as their primary payer.

[a] The counterfactual is the outcome the treatment group would have had in the absence of the HCIA-funded intervention. Our estimate of the counterfactual is the treatment group mean minus the regression-adjusted difference-in-differences estimate.

[b] Percentage difference is calculated as the regression-adjusted difference between the treatment and comparison group, divided by the adjusted comparison group mean.

[c] The *p*-values from the secondary test results were not adjusted for multiple comparisons within or across domains.

ED = emergency department; FFS = fee-for-service; HCIA = Health Care Innovation Award; WIPH = Wyoming Institute of Population Health.

5. Consistency of impact estimates with implementation findings

The impact estimates in the primary tests are not plausible given the implementation findings. As described in Section III, WIPH directed its HCIA funds for the PCMH program to TransforMED for practice facilitation services and to small grants to practices to help pay NCQA application fees. Over time, the focus of the program shifted to NCQA application review instead of practice facilitation services. Although the PCMH program did not include an intensive intervention, it is unclear why beneficaries in treatment practices would have such large unfavorable outcomes relative to beneficiaries in comparison practices as suggested by the quantitative estimates. Most practices participating in the PCMH program were working toward, and half ultimately achieved, NCQA recognition during the HCIA award period, indicating that select practices implemented core elements of the PCMH model successfully. A number of practices had difficulty with aspects of practice transformation, particularly related to EHR adoption and operation, which might have distracted them from optimal patient care or made it challenging to focus on other care improvements that would be expected to affect the outcomes examined. However, we would not expect to see such large unfavorable impact estimates for the outcomes examined, even if practices experienced problems in these areas.

Therefore, based on the implementation findings, we conclude that the large unfavorable quantitative effects are not plausible given that (1) the intervention delivered was minimal and (2) the aspects of the program that practices *did* adopt during the intervention period could not plausibly have produced unfavorable effects as large as the ones observed. It is possible that practices were limited in their practice transformation and had reduced availability to treat patients due to difficulties with EHR adoption or that practices provided more comprehensive care that resulted in detection of additional health issues to address. However, we have no reason to believe that these difficulties could lead to substantively large unfavorable impacts on the outcomes examined during the time periods examined.

6. Conclusions about program impacts, by domain

Based on all evidence currently available, we determined that we could not draw conclusions about program impacts on patients' outcomes for any domain. Table V.6 summarizes these conclusions and their support.

The primary test results showed substantively large, unfavorable differences between the treatment and comparison groups for inpatient admissions, ED visits, and spending. However, the secondary tests also indicated substantively large unfavorable effects for time periods and samples for which no effects were expected—suggesting a limitation in the comparison group. Although the treatment and comparison practices were well matched on observable characteristics at baseline, our findings lead us to believe there could have been unobserved differences between the groups or other confounding factors that influenced the results. Given the results from the secondary tests and the implementation findings, the impact results are not plausible, and therefore, we cannot draw any conclusions about program impacts on patients' outcomes.

Table V.6. Conclusions about the impacts of WIPH's HCIA program on patients' outcomes, by domain

		Evidence		
Domain	Conclusion	Primary test result(s)	Primary test result plausible given secondary tests?	Primary test result plausible given implementation evidence?
Quality-of-care outcomes	None	No substantively important effect; power was low to detect an effect on the single outcome in the domain	No	No
Service use	None	Differences between treatment and comparison groups were substantively large and unfavorable for the combined impact estimate (across two outcomes) in the domain	No	No
Spending	None	Differences between treatment and comparison groups were substantively large and unfavorable for the single outcome in the domain	No	No

Sources: Tables V.4 and V.5.

HCIA = Health Care Innovation Award; WIPH = Wyoming Institute of Population Health.

VI. DISCUSSION AND CONCLUSIONS

WIPH received HCIA funding to create medical neighborhoods across Wyoming, which the awardee sought to achieve with five distinct program components: (1) the PCMH program, which provided training and facilitation to primary care practices to support PCMH transformation; (2) the WyRCT program, which offered transitional care services to patients recently discharged from hospitals who were 65 and older with a qualifying condition; (3) the telehealth program, which provided infrastructure for clinicians at hospitals and practices to provide remote-access care; (4) the Medication Donation Program, which collected donated medications from public and health agencies and distributed them to participating providers to offer to eligible patients; and (5) the Virtual Pharmacy program, through which pharmacists provided medication management services and reported patients' compliance back to prescribers. Collectively among the five components, WIPH aimed to reduce ED visits, hospital admissions, and total spending.

Our evaluation focuses on the PCMH component of the HCIA intervention. WIPH contracted with TransforMED to facilitate the PCMH component, but otherwise provided little direct funding to transforming practices. Consequently, practices' approach and engagement in the transformation process varied. Practices generally worked to implement the six standards necessary to achieve NCQA PCMH recognition—for example, by scheduling team huddles and creating patient reports. By June 2015, 10 of 20 participating practices had achieved NCQA PCMH recognition. Most clinicians reported that the PCMH component made care more patient-centered, but about a third perceived negative effects on efficiency. Commonly cited barriers to transformation were limited physician engagement, staff capacity, and challenges implementing necessary EHR functionalities.

We were unable to draw conclusions about the impact of WIPH's PCMH program on Medicare FFS beneficiaries' outcomes. The primary test results from our impact evaluation showed that the differences between the treatment and comparison groups were substantively large and unfavorable for inpatient admissions, ED visits, and spending, as the comparison group's outcomes improved more quickly over time than the treatment group's outcomes. However, the results from secondary, corroborating tests did not follow the patterns we expected *ex ante,* and they diminished our confidence in the comparison group as a valid representation of the counterfactual. As a result, we believe unobservable factors or statewide differences between Wyoming and Montana might have driven the impact estimates. At the same time, the implementation evidence suggests the PCMH intervention was not intensive, and there is no evidence that it could have caused unfavorable effects as large as those observed. A number of practices had difficulty with aspects of practice transformation, particularly related to EHR adoption and operation, which might have distracted them from optimal patient care or made it challenging to focus on other care improvements that would be expected to affect the outcomes examined. However, it is not plausible that these challenges could have caused such large unfavorable impact estimates for the outcomes examined.

The WIPH intervention as a whole was designed to be diffused statewide to create medical neighborhoods throughout Wyoming. This included five components reaching providers throughout the entire state. We did not assess impacts on patients' outcomes for any program component other than the PCMH component because we either lacked identifiers for providers who participated in the programs, lacked claims for the majority of patients who benefited from the other components, or we were unable to replicate the enrollment criteria, making it difficult to construct a meaningful comparison group. However, the PCMH component also presented challenges to evaluation. WIPH did not use any specific criteria to select practices for the PCMH component—the intervention included any practice that volunteered to join the PCMH component. As with other voluntary programs, selection bias is a potential issue, so practices that joined the PCMH intervention might have systematically differed from other practices in Wyoming that chose not to join. For this reason, we selected our matched comparison group from outside of Wyoming to attempt to minimize selection issues, but our inability to reconcile the primary test results with secondary test results and implementation findings raises concerns that unobservable factors or statewide differences drove the unexpected impact estimates.

CMMI and other stakeholders could consider a number of changes to the design of similar programs in the future to increase the potential to draw conclusions about program impacts on patients' outcomes. For example, for a PCMH practice transformation program, administrators could randomize practices that volunteer so that some participate in the program and others do not. If the program has many components, as the WIPH program did, delivery of those additional components could be tied to the PCMH practice transformation program so that participation in the other components depends on location near a practice randomized into the PCMH treatment group. Alternatively, because randomization can be challenging and expensive in many interventions, program administrators could also consider using explicit selection criteria for PCMH programs so that evaluators can replicate those criteria to construct an appropriate comparison group.

It is useful to draw lessons from TransforMED, another awardee in the HCIA-PCR portfolio presented in Chapter 9. TransforMED had a similar but distinct intervention to WIPH's PCMH component and used its award to provide population health management and cost-reporting software, and technical assistance, to complement practices' PCMH transformation. Two important elements enabled that evaluation to draw conclusions about TransforMED's impacts on patients' outcomes. First, TransforMED provided a focused intervention for treatment practices and did not include any other intervention components within the same regions as the treatment practices. This meant it was possible to find comparison practices that did not receive the intervention within the same region as each treatment practice. Second, TransforMED's treatment practices were located in 15 states. This meant that comparison practices were drawn from the 15 states that corresponded to the areas where the treatment practices were located. Including practices from those 15 states minimized the potential that local practice conditions in one particular geographic area could drive the impact results, as might have been the case in our evaluation of WIPH.

The findings from TransforMED's HCIA evaluation suggest it is possible to draw conclusions for practice-based interventions even when randomization is not used. Conclusions can be possible if the programs under evaluation are focused and allow for selection of comparison practices from similar geographic areas to those of the treatment practices.

This page has been left blank for double-sided copying.

REFERENCES

Boards of Trustees, Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds. "2013 Annual Report of the Boards of Trustees of the Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds." Table V.D1. Washington, DC: Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds, 2013. Available at http://downloads.cms.gov/files/TR2013.pdf. Accessed August 13, 2014.

Centers for Medicare & Medicaid Services. "CSV Flat Files—Revised: Readmissions Complications and Deaths—National.csv." Baltimore, MD: CMS, 2014. Available at https://data.medicare.gov/data/hospital-compare. Accessed August 14, 2014.

Chronic Conditions Data Warehouse. "Table A.1.a. Medicare Beneficiary Counts for 2005–2014." Baltimore, MD: Centers for Medicare & Medicaid Services. Available at https://www.ccwdata.org/web/guest/medicare-tables-reports. Accessed June 29, 2016.

Ehrlich, Emily, Andrea Wysocki, KeriAnn Wells, Boyd Gillman, Greg Peterson, Catherine DesRoches, Sandi Nelson, Laura Blue, Keith Kranker, Kate Stewart, Frank Yoon, Jelena Zurovac, and Lorenzo Moreno. "Evaluation of the Health Care Innovation Awards (HCIA): Primary Care Redesign Programs. Second Annual Report: Findings for Wyoming Institute of Population Health at Cheyenne Regional Medical Center—Patient-Centered Medical Home Component." Princeton, NJ: Mathematica Policy Research, December 11, 2015.

Gerhardt, Geoffrey, Alshadye Yemane, Keri Apostle, Allison Oelschlaeger, Eric Rollins, and Niall Brennan. "Evaluating Whether Changes in Utilization of Hospital Outpatient Services Contributed to Lower Medicare Readmission Rate." *Medicare & Medicaid Research Review,* vol. 4, no. 1, 2014, pp. E1–E13.

Health Indicators Warehouse. "Average Age of Medicare Beneficiaries (mean)." Hyattsville, MD: National Center for Health Statistics, HIW, 2014a. Available at http://www.healthindicators.gov/Indicators/Average-age-of-Medicare-beneficiaries-mean_308/Profile/ClassicData. Accessed November 19, 2014.

Health Indicators Warehouse. "Hospital Inpatient Medicare Admissions (per 1,000 beneficiaries)." Hyattsville, MD: National Center for Health Statistics, HIW, 2014b. Available at http://www.healthindicators.gov/Indicators/Hospital-inpatient-Medicare-admissions-per-1000-beneficiaries_2001/Profile/ClassicData. Accessed August 13, 2014.

Health Indicators Warehouse. "Medicare Beneficiaries Eligible for Medicaid (percent)." Hyattsville, MD: National Center for Health Statistics, HIW, 2014c. Available at http://www.healthindicators.gov/Indicators/Medicare-beneficiaries-eligible-for-Medicaid-percent_317/Profile/ClassicData. Accessed August 4, 2015.

Institute of Education Sciences. "What Works Clearinghouse: Procedures and Standards Handbook, Version 3.0." Washington, DC: U.S. Department of Education, IES, 2014. Available at http://ies.ed.gov/ncee/wwc/DocumentSum.aspx?sid=19. Accessed September 15, 2014.

National Committee for Quality Assurance. "PCMH 2011—PCMH 2014 Crosswalk." Washington, DC: NCQA, 2011 Available at http://www.ncqa.org/programs/recognition/practices/patient-centered-medical-home-pcmh/pcmh-2011-pcmh-2014-crosswalk. Accessed July 13, 2016.

National Committee for Quality Assurance. "Recognition Directory." Washington, DC: NCQA, 2016. Available at http://recognition.ncqa.org/PSearchResults.aspx?state=WY&rp=6. Accessed July 13, 2016.

Peikes, Deborah, Stacy Dale, Eric Lundquist, Janice Genevro, and David Myers. "Building the Evidence Base for the Medical Home: What Sample and Sample Size Do Studies Need? White Paper." AHRQ Publication No.11-0100-EF. Rockville, MD: Agency for Healthcare Research and Quality, October 2011.

Stuart, Elizabeth A. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science,* vol. 25, no. 1, 2010, pp. 1–21.

Truven Health Analytics. "AHRQ Quality Indicators, Prevention Quality Indicators v5.0 Benchmark Data Tables." Prepared for the Agency for Healthcare Research and Quality, U.S. Department of Health and Human Services. Santa Barbara, CA: Truven Health Analytics, March 2015. Available at http://www.qualityindicators.ahrq.gov/Downloads/Modules/PQI/V50/Version_50_Benchmark_Tables_PQI.pdf. Accessed August 18, 2015.