March 2017

Health Care Innovation Awards (HCIA) Meta-Analysis and Evaluators Collaborative

Annual Report
Year 2

Prepared for
Timothy Day
Centers for Medicare & Medicaid Services
Center for Medicare & Medicaid Innovation
7500 Security Boulevard
Baltimore, MD 21244-1850

Prepared by
Anupa Bir, ScD MPH
Kevin Smith, MA
Leila Kahwati, MD MPH
Heather Kane, PhD
Nikki Freeman, MA
Kyle Emery, MA
Jim Derzon, PhD
Holly Stockdale, MPH
Shivani Reddy, MD MPH
Julia Cohen, MA
Benjamin Koethe, BA
Josh Green, BA
RTI International
3040 E. Cornwallis Road
Research Triangle Park, NC 27709

RTI Project Number 0212790.009.005

RTI INTERNATIONAL

[This page intentionally left blank]

HEALTH CARE INNOVATION AWARDS (HCIA) META-ANALYSIS AND EVALUATORS COLLABORATIVE

SECOND ANNUAL REPORT

Federal Project Officer: Timothy Day

RTI International

CMS Contract No. HHSM500201000021i
August 15, 2016

This project was funded by the Centers for Medicare & Medicaid Services under contract no. 500-00-1234. The statements contained in this report are solely those of the authors and do not necessarily reflect the views or policies of the Centers for Medicare & Medicaid Services. RTI assumes responsibility for the accuracy and completeness of the information contained in this report.

CONTENTS

Appendices

List of Figures

List of Tables

EXECUTIVE SUMMARY

Section 1115A of the Social Security Act (added by Section 3021 of the Affordable Care Act [ACA]) authorizes the Center for Medicare and Medicaid Innovation (CMMI) to test innovative health care payment and service delivery models that have the potential to lower Medicare, Medicaid, and Children's Health Insurance Program (CHIP) expenditures while maintaining or improving the quality of beneficiaries' care (42 U.S.C. 1315a). Under the law, preference is to be given to models that improve coordination, efficiency, and quality. CMMI has launched a number of demonstrations to test innovative models that aim to improve care. Beyond the models that are currently being tested, CMMI has funded Health Care Innovation Awards (HCIA) to encourage additional grassroots innovation that addresses locally perceived needs. The first round of HCIA awards was made in July 2012 for a 3-year period of performance. These HCIA awardees have proposed compelling new service delivery and payment models that will drive system transformation and deliver better outcomes for Medicare, Medicaid, and CHIP beneficiaries. The initiative was not prescriptive, but rather open-ended, with specific, shared goals of improving outcomes and reducing costs.

The Centers for Medicare & Medicaid Services (CMS) grouped the HCIA awards by similarity of objective into 10 groups that fall into 3 broad categories of intervention focus (i.e., managing medically fragile populations in the community, interventions in hospital settings, and community interventions) and awarded evaluation contracts to 7 frontline evaluators (FLEs). After an open competition, in 2013 CMS awarded RTI a meta-evaluation contract to use FLE findings and perceptions to summarize HCIA awardees' implementation experiences and the impact of HCIA awards on four core outcomes: total cost of care, hospitalizations, hospital readmissions, and emergency care utilization. This second annual report presents our findings on the first round HCIA awardees' innovation implementation and impacts based on FLE's reporting of awardees through the second year of their award.

Our second annual report (AR2) summarizes work that has been conducted since our first annual report in March 2016. In that time, the number of unique HCIA innovations in our database has grown to 135 unique interventions implemented by the 108 awardees[1] and many of these now have at least 2 years of post-intervention follow-up results. More innovations may soon have 2 years of post-intervention data because 44 percent of the innovations are continuing under no-cost extensions granted in 2015.

Impact Findings. The impact of HCIA innovations on the four core outcomes are estimated by the project's FLEs. Updated forest plots for estimated impacts on costs and utilization show the same mixture of positive, negative, and mostly near zero effects observed in our first annual report. Although a few awardees produced significant cost savings (and a few had significant dissavings), within ambulatory, post-acute care, and hospital-based settings, the average effect of the HCIA award was no significant impact on total cost of care. We observed similar results for three groups of special interest—innovations in rural settings, those addressing

[1] In this report we distinguish between interventions (unique treatments or strategies) and innovations (typically referring to the 108 awardees who received Round 1 HCIA awards); 49 of the 108 awardees (45.4%) used their awards to implement and test multiple interventions.

pediatric populations, and awardees who were granted no-cost extensions. Awardees in these groups returned results showing no savings or dissavings as a result of their innovation.

To confirm these findings were not an artifact of the meta-analytic methods used to summarize awardee achievement, we did analyses using time-series models of longitudinal trends and constructed a Bayesian fixed-effects model. The linear trends in the time-series model supported the meta-analytic results, and the Bayesian model largely duplicated the effect estimates and 90 percent confidence intervals we found using our standard frequentist approach.

To test if the methods used to construct comparison groups may have affected results, we catalogued the five major methods FLEs used to create comparison groups for their difference-in-difference analyses. Analysis of these methodological choices did not detect any systematic bias in HCIA results attributable to the particular comparison group method selected.

Although impact findings should be considered preliminary, results in performance distributions for 8 of the 11 analyses found greater between-awardee variation than would be expected from sampling error. This suggests there may be particular features of innovations that impact performance on the four core measures. To systematically assess what features of innovations are associated with greater and lesser effectiveness, we conducted a series of meta-regressions testing the relations of innovation features with the year 2 magnitude of cost savings reported for ambulatory awardees. Out of a variety of structural, intervention component, and implementation features, the following features of innovations were associated with improved cost savings:

- For-profit tax status
- Health informatics component
- Community Health Workers
- Clinically fragile patients.

New innovations (in contrast to extensions of existing, ongoing programs) and awardees experiencing problems hiring or retaining frontline staff, on the other hand, had greater costs than their comparisons.

We linked innovation features, implementation effectiveness, and three core outcomes together in a unified meta-path model. The results demonstrated that hospital admission effects had a greater impact on cost savings than emergency department (ED) use effects did. The key mediator in the path model was new innovation status. Compared to existing innovations, new innovations had lower levels of implementation effectiveness and produced less favorable hospital admission and ED effects.

Implementation Findings. Implementation experience and effectiveness at the awardee level were assessed with an FLE survey (Annual Awardee Summary Form) and through thematic analysis of FLE quarterly and annual reports. By the end of the second year, over 80 percent of innovations were considered implemented to a great or moderate extent by FLEs. Several themes

identified in the first annual report continued, and other new themes were identified; these are summarized below. Despite relatively high measures of implementation effectiveness, many awardees continued to face both anticipated and unanticipated challenges. The lack of reimbursement for non-traditional staff and services remains a challenge and has implications for sustainability once CMS funding ends. Cultural barriers (e.g., language barriers, lack of trust) were challenges for innovations delivering care or placing self-monitoring technologies in patients' homes. Vulnerable patients' needs for additional resources and support affected recruitment and treatment maintenance. Several awardees needed additional staff to support innovation implementation, and were challenged in recruiting those staff. Lastly, the time necessary to forge strong relations with new partners was an unanticipated challenge for many awardees.

Awardees adapted their innovations in response to these and other challenges, with several benefits emerging as innovations matured. In particular, implementing effective and workflow-friendly health information technologies (HIT) was a commonly expressed challenge in FLE reports. However, by the end of the second year of implementation, these challenges were generally judged to be small to modest and the benefits of robust HIT infrastructures were becoming apparent. Also expressed were improvements in staff satisfaction, retention, empowerment, and staff relations. Reasons given for these improvements include awardee cross-training; physical colocation of staff; and improved recruitment, hiring, and training practices.

Staff appreciation of community health workers (CHWs) increased in the second year as staff recognized the contributions CHWs made in improving workflow, connecting with patients, and enhancing implementation. However, lack of reimbursement for CHWs, care coordinators, and new staff types is a significant barrier to sustainability, and awardees adopting these innovations are turning to future payment reform pilots or demonstrations—mainly Accountable Care Organization (ACO) pilots—for sustainability.

Perhaps most importantly, clinical staff satisfaction with and support for the innovations increased markedly in the second year as the value of innovations in improving workflow and patient care became increasingly apparent. Independent of success on the four core outcomes, several innovations will likely be sustained in whole or in part on the basis of staff support and satisfaction.

Nonetheless, several challenges, often beyond awardee control, continue to affect awardee performance and sustainability. Few awardees found formal improvement or change management processes useful for monitoring innovation implementation and such processes were adopted to a great extent by only 12 awardees (14.6 percent). Awardees also reported that existing organizational capacity affects resilience to challenges and may affect sustainability, as do states' decisions around Medicaid expansion and their support (or lack thereof) for alternative payment models.

With innovations successfully implemented and their benefits observed, innovation and organizational leaders increasingly turned their attention to sustaining all or part of their innovations once CMS-support ends. Some awardees turned to state and federal funding streams for ongoing support while others secured financing from commercial health plans. Although sustaining their innovation is often conditional on a demonstrated return on investment or

documented improvements in patient health outcomes, awardees in large provider institutions expect their workflow-integrated innovations to continue once HCIA funding ceases. For many awardees, partners played an active and strategic role in planning for sustainability by agreeing to adopt and integrate key innovation components into their existing work.

Rapidly and effectively implementing effective innovations remains a significant challenge for delivering innovative health care. Using Qualitative Comparative Analysis (QCA) and path modeling (a form of correlational analysis), we attempted to isolate programmatic features associated with implementation success. QCA did not identify any necessary or sufficient programmatic features or combinations of features: all tested features were present in both effective and ineffectively implemented interventions. The path model identified three key features independently associated with greater implementation success. Awardees were more successful in implementing their innovations when implemented at a single-site, engaged in more staff training, and engaged in more robust implementation planning. For innovations implementing HIT, filling frontline staff roles and recruiting and retaining staff were significant challenges, and staff who were hired to fill technical, research, or administrative roles were significantly more likely to work semi-independently than clinical staff. New programs were somewhat less effective in implementing their awards, faced greater challenges in implementing HIT, and were somewhat more likely to hire technical, research, or administrative staff to support their innovation although these features did not significantly impact implementation effectiveness.

SECTION 1:
BACKGROUND AND METHODS

Section 1115A of the Social Security Act (added by Section 3021 of the Affordable Care Act [ACA]) authorizes the Center for Medicare and Medicaid Innovation (CMMI) to test innovative health care payment and service delivery models that have the potential to lower Medicare, Medicaid, and Children's Health Insurance Program (CHIP) expenditures while maintaining or improving the quality of beneficiaries' care (42 U.S.C. 1315a). Under the law, preference is to be given to models that improve coordination, efficiency, and quality. CMMI has launched a number of demonstrations to test innovative models that aim to improve care. Beyond the models that are currently being tested, CMMI has funded Health Care Innovation Awards (HCIA) to encourage additional grassroots innovation that addresses locally perceived needs. The first round of HCIA awards was made in July 2012 for a 3-year period of performance. These HCIA awardees have proposed compelling new service delivery and payment models that will drive system transformation and deliver better outcomes for Medicare, Medicaid, and CHIP beneficiaries. The initiative was not prescriptive, but rather open-ended, with specific, shared goals of improving outcomes and reducing costs.

The Centers for Medicare & Medicaid Services (CMS) seeks to learn from the efforts of the diverse group of awardees. For evaluation purposes, CMS categorized awardees into 3 groups based on their principal focus and into 10 groups for their similarity of objective. These 10 groups were then assigned to 7 frontline evaluators (FLEs) who conducted process and impact evaluations. In addition, in 2013 CMS awarded RTI a meta-evaluation contract to synthesize results from FLE reports and observations to obtain an overarching perspective on the learning from all HCIA awardees. This will allow general conclusions to be drawn across these all interventions about which approaches are most promising, for which populations, and in what conditions and settings they are most appropriately implemented. For this evaluation, we rely on FLEs' data and analyses reported in the FLEs' quarterly and annuals reports. The meta-evaluation does not collect data directly from awardees, and has no direct contact with awardees.

In addition, the meta-evaluation will address specific cross-cutting service delivery issues across awardees in developing strategies for pediatric populations, rural populations, and populations with behavioral health needs. Moreover, the meta-evaluation examines how interventions can be scaled up to wider practical use and how they can best be subjected to broad-based testing and ongoing quality improvement. In addressing these questions, we use the entire awardee portfolio, allowing comparisons between groups and within and between specific subpopulations of interest.

To maximize efficiency, the scientific value, and the utility of findings for CMS, we coordinate with the FLEs, evaluating the different awardee groups in aggregate. We work with the FLEs to ensure that (1) the full set of available outcomes and data is understood and carefully managed, (2) we thoroughly understand the interventions and study designs across the projects, (3) we have the opportunity to suggest and influence changes or additions to data collection through CMS representatives for the frontline evaluation, and (4) we collect the analytical outputs from the FLEs that inform the overarching evaluation. For outcomes based on claims data, we focus on developing and collecting standardized measure calculations. From awardee measurement and monitoring plans, we assessed the extent to which awardees across groups

include the same measures. For additional outcomes, particularly qualitative ones, we also engage in upfront coordination with FLEs to maximize the set of available and relevant measures for characterizing the key overarching features of interventions, settings, and contexts.

This report is presented in three major sections. The first covers the background of the initiative, our role in its evaluation, and the data and methods we used to assess awardees' implementation experience and the impact of awards on the four core outcomes: total cost of care, hospital admissions, emergency department use, and hospital readmissions. The second major section presents our findings on how award implementation was experienced by awardees and examines whether award features or conditions were systematically associated with implementation effectiveness. The third major section presents findings related to awardee effectiveness in improving the four core outcomes. For these analyses we partitioned the HCIA interventions into three broad classes: ambulatory care, post-acute care, and hospital-setting as each represents distinct intervention approaches for different populations with different health care needs. We turn now to the data sources and analytic techniques used in this annual report to examine the implementation and effectiveness of HCIA awardees in improving health care delivery and their impact on health care costs and utilization.

1.1 Data Sources

Primary data acquired for analyses in this report include data elements from the first and second annual awardee summary forms, the no-cost-extension statuses from CMMI, and the quarterly means data from the FLEs on each of their awardees. As meta-evaluator for the HCIA model, however, the majority of our data is secondary data collected from the HCIA awardees by the FLEs and the Lewin Group, which is the implementation contractor. Section 1.1.1 outlines our primary data sources, the data elements derived from them, and their uses in this report; section 1.1.2 does the same for each of our secondary data sources.

1.1.1 Primary Data

Outcome Data. From the beginning of our meta-evaluation, we have been collecting quarterly means data from the FLEs for each of their awardees[2] using a quarterly template data tool. These data include regression-adjusted quarterly difference-in-differences scores (DiD), and means and standard deviations for baseline and intervention quarters for each of the four core measures of analysis: total cost of care (TCOC), all cause hospital admissions, all cause hospital readmissions, and emergency department (ED) utilization. FLEs were requested to calculate Medicare total costs using only Medicare Parts A and B, quarterize (prorate) payments for patients with less than a full quarter's eligibility (except patients who die or for the first inpatient admission in a quarter), and to not standardize, risk-adjust or down weight for partial eligibility. All cause hospital admissions were defined as the number of patients who were admitted to the medical-surgical units and excludes patients kept overnight in observation beds. Patients with multiple admissions in a quarter were counted each time they were admitted, and estimates were

[2] The seven FLEs were expected to provide these data for each of their awardees although there were many instances in which not all data were available or were not provided to our specifications, thereby limiting our ability to include a number of awardees in our meta-evaluation. More detail on quantitative data availability and quality is in section 3.1.

to be quarterized. Risk adjustment was to be done during construction of the comparison group, although further adjustment using diagnostic characteristics was possible for estimating intervention effects. All cause readmissions were to be similarly quarterized and risk adjusted and were defined as an unplanned follow-up admission to any short-term acute general or long-term care hospital within 30 days of a discharge from another hospital of the same type. Finally, all cause ED utilization includes any overnight ED visits without a hospital admission including overnight ED observation visits without a hospital admission.

In this report we use data collected via the template to estimate the influence of sample size on cost and utilization effect sizes, to estimate the overall impact of the initiative, to estimate reasons for variation in result, and for our continued time-series analysis.

Annual Awardee Summary Forms. Seeking additional information from the FLEs on each of their awardees, we collected data from FLEs in 2014 through a structured assessment form, or the Annual Awardee Summary Form (AASF). This form asked FLEs to provide information about key awardee characteristics, staff deployment models, program design, and project history. For the analyses in this report, we supplemented the primary data collection conducted in 2014 by using a revised version of the form in 2015. The Second Annual Awardee Summary Form (AASF2) asked FLEs more targeted questions about awardees' implementation process and solicited a more uniform assessment of both implementation and intervention effectiveness. Using Likert scales, the tool asked FLEs to assess 4 to 12 different measures in each of 7 domains: innovation complexity, implementation planning, implementation process, staff training, organizational leadership, implementation effectiveness, implementation findings, and intervention impact. (Additional details on the AASF2 assessment tool are provided in Appendices A–C).

These data were used to construct composite scores in each of the domains for various implementation effectiveness analyses—including Qualitative Comparison Analysis—discussed in section 2, and for meta-regression analyses (section 3.4).

No-Cost Extension Statuses. We also collected data from CMMI on their decisions for each awardee regarding a no-cost extension of the awardee's intervention period. These data were used in meta-regression and path model analyses.

1.1.2 Secondary Data

FLE Reports. As mentioned above, most of our data is secondary data collected from the awardees by the FLEs. For the analyses in this report, we utilized qualitative data gathered from our thematic analysis of FLE second annual reports and eighth quarterly reports to CMS. We developed a structured and systematic coding scheme for innovation components and characteristics to standardize qualitative data extraction; these data elements comprise our structured qualitative coding (SQC) data, which is used extensively in our implementation and intervention effectiveness analyses.

The FLE reports also provided the quantitative data used in our intervention effectiveness analyses: FLEs were asked to provide summative intervention effect size estimates using DiD regression modeling for their awardees for each of four core measures. These data were used to estimate the overall impact of the innovations for ambulatory, post-acute, and hospital-based samples (section 3.2), meta-regression (section 3.4.2) and path modeling analyses (section 3.4.3).

Lewin Group (Implementation Contractor) Data. Analyses in this report also use several variables collected in awardees' quarterly report to CMS, collected by the Implementation Contractor, Lewin: awardee tax status (for-profit or not-for-profit), organization type (academic institution or not), direct participant enrollment, awardee geography (urban, suburban, or rural), CMMI award spending, and barriers to enrollment. These variables were used in meta-regression and path model analyses.

1.2 Analytic Techniques

1.2.1 Data Coding and Transformation

We conducted a thematic analysis of the seven FLE's second annual reports and fifth, sixth, and seventh quarterly reports, and associated appendices using NVivo to code text associated with implementation findings. Three pilot coding passes were undertaken to calibrate the qualitative coding, which led to minor adjustments to the coding scheme. We achieved intercoder reliability kappas[3] of 0.7 or greater for most codes. The various content and format differences across the FLE reports and appendices generated the major challenge in text coding and synthesis. Some of these differences may be unavoidable because of the variation in types of awardees and nature of the different awardee innovations. Differences were managed through adjudication and a final inventory of themes reified. Nonetheless, many major themes emerged from this process and are presented in section 2.1.2 of this report. In addition to thematic analysis, we also updated a structured and systematic coding scheme for innovation components and characteristics originally developed for the first annual report.

1.2.2 Qualitative Comparative Analysis

We used qualitative comparative analysis (QCA) to examine combinations of innovation features, including characteristics of the target population, components and characteristics of the innovation, and contextual features related to the implementation that might contribute to implementation effectiveness. Drawing from mathematical set theory, QCA examines which features—individually or in combination—are necessary or sufficient for producing an outcome (Schneider & Wagemann, 2012). This report uses QCA to identify the necessary and sufficient features associated with effective (versus ineffective) implementation. Effective implementation refers to the awardee's consistent delivery of the planned innovation care or services to the intended target population at the intended level of quality and intensity. In QCA, a feature (or combination of features) is considered "necessary" if it is a consistent feature among awardees with effective implementation. A feature (or combination of features) is considered "sufficient" if effective implementation is a consistent outcome among awardees with the feature (Ragin,

[3] Kappa is a statistical measure of interrater agreement that ranges from 0 to 1, with 1 indicting perfect agreement.

2000). Relationships of necessity and sufficiency are a type of complex relationship that traditional qualitative and quantitative methods are not able to identify.

1.2.3 Quantitative Impact Analysis

In this report, we use an expanded repertoire of quantitative methods to evaluate the impact of the HCIA innovations. Section 3 contains results from six distinct methods.

Forest Plots. The standard approach to meta-analysis is to compute a mean intervention effect and standard error for each awardee and then display results for all awardees in the form of a forest plot. We begin by updating the innovation effect plots from the last report, incorporating additional awardees and longer follow-up periods.

Bayesian Analysis. The Bayesian approach uses observed data to revise probability distributions. In this report, we use Bayesian techniques to illustrate how meta-analytic findings can be expressed in terms of the probability that an innovation will achieve savings in total cost of care.

Time-Series Analysis. For some time, we have been asking FLEs to report core outcomes on a quarterly basis. We now examine these longitudinal data using time-series analysis. In contrast to the difference-in-difference effects from the main regression analyses, the time-series approach measures effects in terms of different time trends between the HCIA and comparison groups.

Heterogeneity Analyses. We computed two formal statistical tests to determine (1) whether all innovations share a common effect size for a particular core measure, and (2) the proportion of the total effect variance that can be attributed to between-innovation differences.

Meta-Regression. We present meta-regression models in this report for the first time. In the analyses for ambulatory setting innovations, HCIA total cost of care effects are regressed on three types of innovation features (structural characteristics, innovation components, and implementation features).

Meta-Path Analysis. Finally, we used path analysis, a form of structural equation modeling, to conduct two multivariable analyses. The first path model examines the influence of innovation characteristics, challenges, and performance on FLE ratings of implementation effectiveness. The second path model links implementation features to implementation effectiveness and then to the effects for three core outcomes.

[This page intentionally left blank]

SECTION 2:
IMPLEMENTATION FINDINGS

2.1 Implementation Findings Overview

Implementation experience and implementation effectiveness of HCIA awardees was assessed using structured survey data provided by the FLEs and using FLE quarterly and annual reports submitted in the second year of the award. The findings from these sources are introduced below and described in detail in the following subsections.

According to the FLEs, the vast majority of innovations were implemented to either a great or moderate extent. Nonetheless, many awardees continued to struggle with both anticipated and unanticipated challenges including:

- Reimbursement for non-traditional staff and services
- Managing cultural barriers
- Complex and vulnerable patients' needs for additional resources and support
- The need for additional staff to support innovation implementation
- The time necessary to forge strong relations with new partners
- The ongoing challenges of implementing health information technologies (HIT).

To meet these challenges, many awardees adapted their innovations to improve staff satisfaction, retention, empowerment, and relations. These adaptations included:

- Staff cross-training
- Physical colocation of staff
- Improved recruitment, hiring, and training practices.

As these innovations have matured, the value of innovations in improving workflow and patient care has become increasingly apparent to clinical staff, and their satisfaction and support for the innovations has increased. In particular, the benefits of robust HIT infrastructure have become apparent for many awardees and, as clinical staff have worked alongside community health workers (CHWs), staff appreciation of their contributions in improving workflow, connecting with patients, and enhancing implementation has grown.

However, awardees continued to experience implementation challenges that are often beyond awardee control and these continue to impact awardee performance and sustainability. For example, prior and existing organizational capacity, state decisions around Medicaid expansion, and state support (or lack thereof) for alternative payment models all impact awardees' resilience to challenges and may affect innovation sustainability. Due in part to the

innovative nature of these interventions (few had formal protocols or manuals) and in part to the varied experience of awardees in implementing innovative practices, few awardees found formal improvement or change management processes useful or appropriate for monitoring implementation of their innovation.

In year 2, innovation and organizational leaders are increasingly turning attention to sustaining the innovation beyond the 3-year award period. Key strategies and approaches to sustainability include:

- Seeking state and federal funding for ongoing support
- Securing financing from commercial health plans
- Participating in future payment reform pilots or demonstrations
- For larger provider institutions, absorbing innovations into existing operations.

Awardees seeking payer support for sustainment have found that such support is conditional on a demonstrated return on investment or documented improvements in patient health outcomes. For several awardees, partners are taking an active and strategic role in planning for sustainability by agreeing to adopt and integrate key innovation components into their existing work.

2.1.1 Revised Annual Awardee Summary Form

The RTI meta-evaluation does not evaluate HCIA awardees directly, but gets its evidence of awardee performance from the FLEs assigned to each awardee. However, FLE reports did not provide consistent data from all domains relevant to implementation evaluation across all awardees. Although FLE reports follow a general template, authors of those reports have considerable latitude in what is presented, and the depth to which it is discussed. Therefore, in September 2015 a Second Annual Awardee Summary Form (AASF2) was administered to FLEs, most of whom completed awardee site visits in 2015. The survey asked FLEs for their assessment of awardee success in several domains associated with implementation effectiveness. The survey was developed using themes from the first annual report and the Consolidated Framework for Implementation Research (CFIR; Damschroder et al., 2009). The domains assessed include innovation effectiveness, complexity, planning, and process; staff training;

In September 2015, a Second Annual Awardee Summary Form was fielded to FLEs for their assessment of awardee success in several domains associated with implementation effectiveness. The survey was developed using themes from the first annual report and the Consolidated Framework for Implementation Research Damschroder et al., 2009). The domains assessed include innovation effectiveness, complexity, planning, and process; staff training; organizational leadership; general challenges; and challenges specific to HIT. Most items produced good variation in responses, although, since many awardees did not have partners or implement HIT, there were many valid missing responses for those items. In addition, many FLEs were unable to assess organizational leadership. Although the instrument was not originally designed to form subscales, with the exception of complexity (two factors), domains could each be reduced to a single factor. Many of these factors are used in subsequent analyses.

organizational leadership; general challenges; and challenges specific to HIT. A detailed description of the design of AASF2 and the by-item summary findings are given in Appendix A; the AASF2 instrument and instructions for its administration are provided in Appendix B.

Most items produced good variation in responses, although, since many awardees did not have partners or implement HIT, there were many "not applicable" responses for those items. In addition, many FLEs were unable to assess organizational leadership. A series of factor analyses was performed to assess the domain structure of the instrument. With the exception of complexity (two factors) and challenges, item responses within domains were consistent and items within the domain could be summarized by a single score for each intervention. The methods used and the factor analytic results for all domains are presented in Appendix C. The summary scores supported by the factor analyses are used in several subsequent multivariate analyses.

The main AASF results in three keys areas were as follows:

Implementation. In general, interventions were complex to a moderate or great extent across the 12 items assessing complexity. With the exception of "Required hiring technical, research, or administrative staff new to the organization," which split almost evenly, most respondents selected to a great or moderate extent in estimating degree of complexity. Most interventions were planned to a great or moderate extent, and awardees were split about evenly in having "Experience with implementing similar programs at a similar scale." According to FLE respondents, interventions were introduced in a planned and deliberate way with very few exceptions. Similarly, awardees reportedly worked well with all necessary entities both within and across organizations and executed self-monitoring plans, although few awardees used formal improvement frameworks or change management processes. A majority of interventions identified the need to train staff in new or additional skills to either a great or moderate extent, and most interventions provided high levels of training for staff. Of all domains assessed, FLE respondents expressed the least comfort in responding to items assessing organizational leadership's support for and interest in awardee interventions. Between 18.7 and 37.4 percent of respondents selected "unable to assess" for items assessing organizational leadership's support for and involvement in CMMI innovations.

Challenges. While qualitative coding of FLE annual and quarterly reports in our first annual report identified many challenges to innovation implementation, the second AASF shows that, while many challenges are evident, they may not be as severe or prevalent as implied in the FLE quarterly and annual reports. Of the 12 items assessing implementation challenges, only "Enrolling patient participants" was a great or moderate challenge for over 50 percent of interventions. The "Level of reimbursement for services" was a great or moderate challenge for a near parity of interventions, with 1 in 5 interventions (21.1 percent) rating the extent of the challenge as great.

While not applicable to all awardees, among those implementing HIT the challenges associated with implementing HIT were most frequently perceived to be moderate, slight, or not at all. "Building out or installing HIT to support the innovation," "Data standardization across systems," and "Interoperability across organizations" were perceived to be moderate to great

challenges for about half of the interventions and "data standardization across systems" was the most frequently cited major challenge (14.6 percent).

Implementation Effectiveness. As reported by the FLEs, the vast majority of interventions were implemented to either a great or moderate extent although the effectiveness of implementation tended to vary by site. Two-thirds of interventions were rated as having achieved "Full adoption by frontline staff" to a great extent, while over half reported "Completion of all tasks needed for full innovation implementation" and that "Innovation components [were] being delivered as intended and at the prescribed level of quality" to a great extent. "Rapid adoption of the innovation" showed the greatest variation in response, with 21.1 percent rating rapidness of adoption as either great or slight with 45.5 percent reporting that rapid adoption was achieved to a moderate extent; FLEs found implementation to be not at all rapid for 5.7 percent of interventions. Interestingly, FLEs were unable to assess some aspect of implementation effectiveness for between 4.1 and 9.8 percent of interventions.

In general, and in contrast with the first AASF, FLEs responding to the AASF2 were able to provide ratings for most items, and most items showed considerable variation in response. Respondents had the greatest difficulty providing ratings in the "Organizational Leadership" domain, but were still able to provide ratings in this domain for three-quarters of the interventions. Results from AASF2 indicate that most innovations were successfully implemented with generally positive ratings in all implementation domains and while many interventions experienced challenges in implementing their intervention, most challenges were judged to be small or modest challenges. Even HIT, a prominent challenge in our first annual report, was not a considered by FLEs to be a major challenge for many awardees by the end of their second year of award. Implementation planning was carried out in a structured and organized fashion, with sufficient attention to training for a strong majority of interventions. This attention to the mechanics of implementation likely contributed to the success awardees experienced in implementing these often complex innovations.

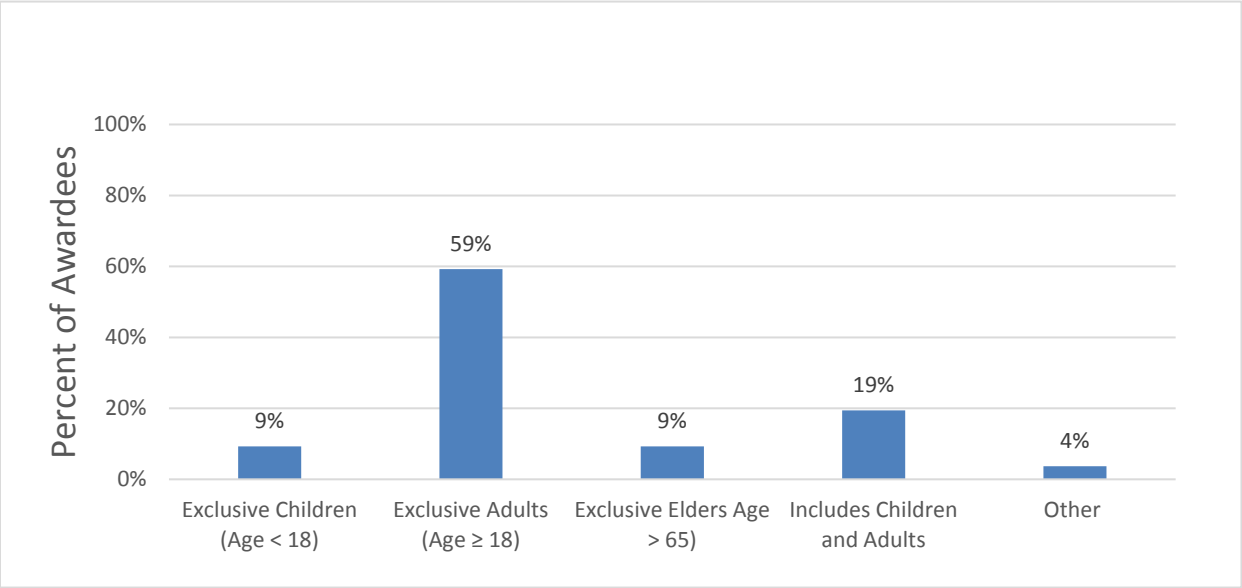
2.1.2 Emerging Themes Results

2.1.2.1 Target Population Findings

This section describes and discusses findings related to innovation target populations as reported by FLEs in their second annual reports. The terms used by the awardee to describe its target population vary. We updated our target population assessment based on our review of FLEs' second annual reports. For this section, we systematically evaluated the target population for each awardee to determine how innovations focused on different ages and different payer beneficiary groups. ***Figure 2-1*** shows that distribution of awardees by various age groups, ***Figure 2-2*** shows the distribution of awardees by payer status, and ***Figure 2-3*** shows the distribution of awardees by the fragility and complexity of the patient population. Most innovations (59 percent, N = 64) have the adult population aged 18 or older as an intended population; 9 percent (N = 10) focus exclusively on children. Another 19 percent (N = 21) included but were not limited to children. A small proportion (9 percent, N = 10) focused on elders (aged 65 or older) only. Many awardees include more than one payer, with the majority of innovations target Medicare fee-for-service, Medicaid, or dually eligible beneficiaries. In addition, 39 percent of awardees also include commercial plan members as a target population. Some awardees did not indicate any

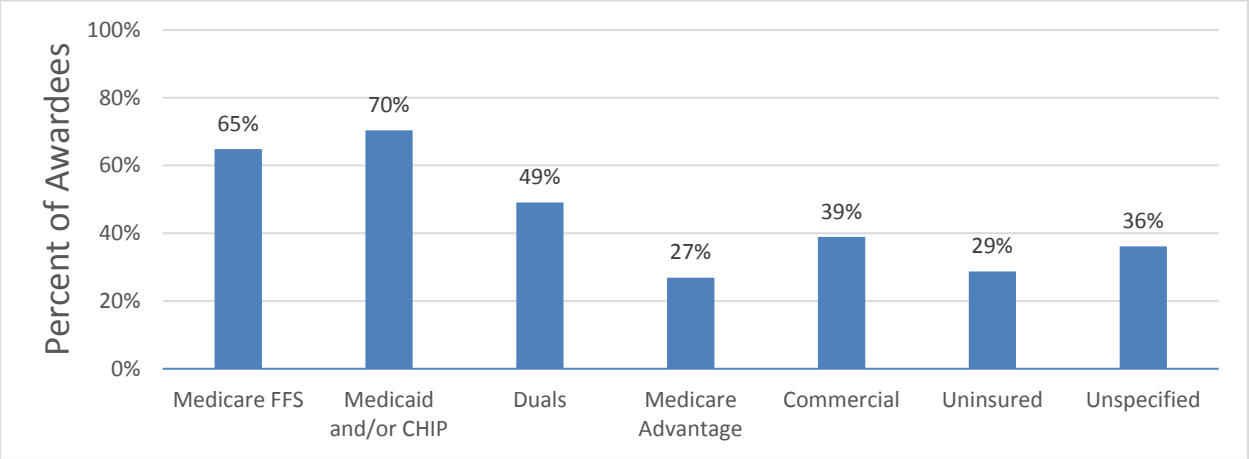
specific payer group as a target of their innovation, and others indicated that beneficiaries of any payer (e.g., Medicare, Medicaid, commercial) and the uninsured were eligible to receive the innovation. The majority (65 percent) of innovations that were designed to target and enroll individual patients defined criteria for enrollment that required multiple or complex clinical conditions or parameters that describe a patient's clinical statues (e.g., laboratory values). This includes mental health and substance abuse diagnoses. Thirty-two percent of innovations used criteria based on a history of "high" utilization, for example frequent ED visits, recent hospital admission, or recent hospital readmission. Lastly, 17 percent defined criteria for enrollment based on social characteristics, such as homelessness or patients with language, cultural, or transportation barriers.

Figure 2-1
Target Population by Age Category[a]



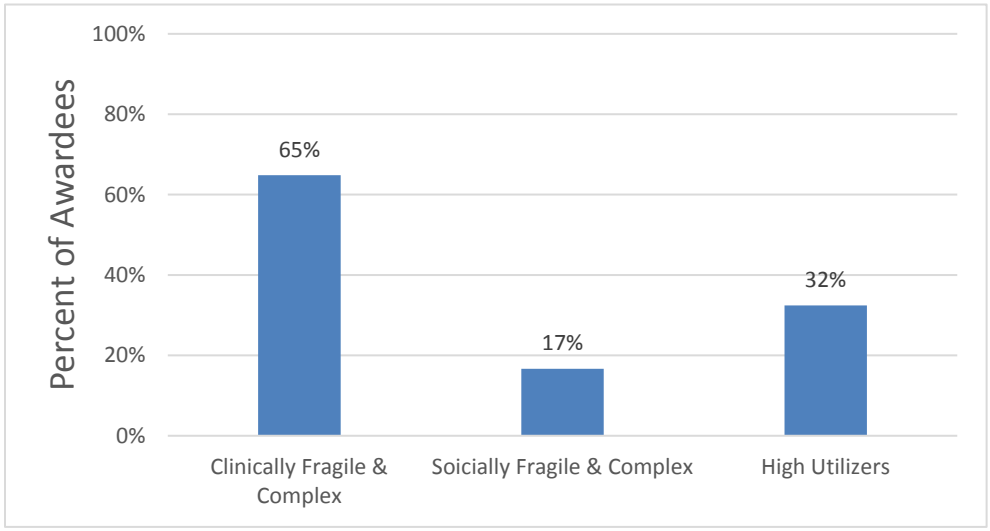
[a] Data obtained from structured coding of FLE quarterly and annual reports.

Figure 2-2
Target Population by Payer Status[a]



[a] Innovations may target more than one type of payer beneficiary. Data obtained from structured coding of FLE quarterly and annual reports.

Figure 2-3
Target Population by Fragility and Complexity of Patient Population[a]



[a] Innovations may target more than one type of patient population. Data obtained from structured coding of FLE quarterly and annual reports.

In the first annual report, we identified several barriers to identifying, enrolling, and engagement patients for the innovations, including identifying and engaging medically complex and socially vulnerable patients, inadequate or faulty administrative data (e.g., missing or outdated contact information), and misalignment between their enrollment projections and the limited pool of potential participants. These issues continued to challenge some awardees during the second year of implementation.

Some awardees encountered and managed cultural barriers in serving racial or ethnic minorities. A few awardees that worked with minority populations encountered barriers (e.g., language barriers, lack of trust in service providers because of cultural differences) in serving those populations. One awardee sought to install medical monitoring equipment in patients' homes that would allow patients to monitor their condition, but some patients did not speak English and were uncomfortable allowing a stranger into their homes. Those obstacles sometimes meant that providers could not install the medical equipment, and thus patients had less autonomy to monitor their health condition. Another awardee had difficulty working around patients' cultural differences and apprehension over using technology. For this awardee, staff found alternatives to address patient concerns. For example, when a patient could not read her glucometer, home visiting staff provided her with a device that gave audio readouts.

One awardee changed the staffing plan at one of its sites because of cultural concerns. The awardee originally planned to add CHWs as cultural advisors to its care teams in one site to advocate for its many Native American patients and include them in the innovation. However, the awardee's cultural liaison informed them of Native Americans' historically grounded distrust of health care institutions. The liaison suggested that the awardee start by raising awareness of Native American culture among staff, and using CHWs to build relationships with the Native American population. As of May 2015 the awardee was revising its plan to address this.

Vulnerable patients require high levels of resources and support to enroll and maintain engagement. Socially or clinically fragile patients required a lot of time, attention, and support to recruit and stay engaged in some innovations. Enrolling vulnerable patients sometimes entailed addressing other needs before working with them on health needs. As observed in the first annual report, meeting those needs required unanticipated time and resources, delayed the provision of health care, and contributed to staff frustration. For example, two awardees found that some rural populations needed help with basic social needs (e.g., transportation, stable housing, running water), which had to be addressed before staff could begin helping those patients manage their medical conditions. Another awardee had a social worker connect vulnerable patients with public benefits. One of the awardees tried to mitigate transportation barriers by sending nurses to patients' homes to provide care. This taxed nurses' schedules and exacerbated staffing shortages for some sites; in turn this decreased staff satisfaction with the innovation. After addressing the most pressing social needs, awardee staff often made multiple follow-up calls and contacts to maintain vulnerable patients' engagement. For example, in one awardee, care coordinators repeatedly contacted patients to encourage their continued participation; however, despite the many attempts, some patients still left the program before completing it.

2.1.3 Implementation Process Findings

In this section we describe findings related to the innovation implementation process. This includes a summary of findings from standardized reporting by FLEs on the AASF along with thematic findings from FLE annual reports. This section covers concepts related to innovation adaptation, self-monitoring, use of formal change management processes for implementation, involvement of partners, and HIT-related implementation issues. Through the AASF, some FLEs reported separately for awardees with distinct innovation components, as a result, we report findings for 123 innovations implemented by 108 awardees. As reported last

year, most awardees implemented innovations at multiple sites (N = 82, 67 percent). Sites refer to geographically or organizationally distinct locations for care; across awardees, the mean number of implementation sites was 13.2. Among multisite awardees, 46 percent (N = 38) had a management team at the awardee-level that oversaw implementation and 28 percent (N = 23) had site management teams. The remaining 8 multisite innovations used some other model of implementation oversight. Over half (57 percent) of participating sites in multisite innovations were external partners and not under the direct management control of the awardee organization.

Adaptation of innovations is a common consequence of implementation. Identifying which components of an innovation are essential and immutable, and which can be modified without adversely impacting innovation fidelity and effectiveness, may increase the generalizability of innovations and improve the identification and adoption of effective practices. Similar to last year, awardees refined components of innovations in response to unmet patient needs, differences in organizational/site capacity to deliver innovation, and preferences of patients and providers. Some awardees continued to struggle with fitting in new services to existing workflows and resources.

Awardees expanded their self-monitoring activities, and are increasingly using data to monitor implementation and impact. According to FLEs, 76 of 123 interventions (61.8 percent) were closely monitoring their implementation process to a great or moderate extent, while 85 of the 123 (69.1 percent) were judged to be implementing their self-monitoring plan to a great or moderate extent. As data systems improved, new challenges emerged. For example, to share results of ongoing self-monitoring with their partner sites, one awardee's staff generated site-level reports, initially providing data on every metric to each site on a quarterly basis. Most site representatives indicated that these voluminous results were "information overload" and requested more user-friendly reports. Program staff simplified the reports and provided leaders at remote sites with PowerPoint slide decks that they could then adapt for staff presentations.

Other examples of how awardees combined self-monitoring with quality improvement activities emerged during the second year of the awards. For example, one awardee constantly monitored performance using weekly benchmark reports generated by their electronic health records (EHR), which then populate the innovation's dashboard. This information is shared at weekly leadership meetings. The tracked information includes number of ED patients 65 or older; number of these 65+ patients with flags indicating enrollment in the innovation; proportion of hospital admissions; wait time to see provider; ED length of stay; estimated proportion of revisits; EHR notes recorded by the social worker, nurse practitioner, or pharmacist; physical therapist consult at discharge; number of Regional Health Information Organization (RHIO) notifications; and number of patients arriving at innovation ED from another local ED as a result of a RHIO notification. As another example, one awardee conducted a demand analysis to better align social worker availability with ED crisis need. This data-driven quality improvement activity increased patient capture from 50 percent to 75 percent.

In year 2, a broader array of self-monitoring data was described, covering both implementation monitoring and impact monitoring, including service delivery measures, quality measures, measures of impact on cost and utilization, and patient experience data. We also identified that innovations using time-sensitive workflows (e.g., a sepsis bundle to be implemented within 3 hours of patient admission) requires astute monitoring and constant

refinement of the clinical and HIT workflows. These were innovations largely being implemented in acute care settings.

Few awardees used a formal improvement or change management process (e.g., Plan-Do-Study-Act, Lean); innovations varied in terms of how easily associated health care practices can be implemented using such approaches. Only 24 of 123 awardee interventions (19.5 percent) used formal improvement or change management processes to a great or moderate extent. It is likely that these methods may be most appropriate for awardees focusing on specific conditions or adopting or adapting well-defined care protocols than they were for awardees developing innovations that require flexibility, cooperation across organizations, or patient engagement and participation. Further, formal approaches may require considerable investment; this may limit the benefits associated with formal change management systems, especially in under-resourced settings. An additional theme that emerged in year 2 reports concerns the benefit that a formal process offers for getting staff at a range of levels in the organization involved in both providing feedback to make changes and enhancing the visibility of the impact of the change across the hierarchy of staff.

Apparent in the second year was the need for additional staff to support many of the innovations implemented by awardees. Most interventions reported having to hire additional technical, research, or administrative staff who were new to the organization. Only 17 interventions (13.8 percent) reported not having to hire any of these support personnel. In addition, partners often provided training essential to innovation delivery, served as sites for patient recruitment, offered tools and technical expertise supporting the use of HIT, or enabled the provision of more comprehensive care or services to patients.

In year 2, we saw that it takes time to build strong relationships between awardees and partners, especially for new partnerships. Awardees used several strategies to build strong relationships and trust including the use of collaborative practice agreements, joint objectives, one-on-one meetings, face-to-face engagement, and talking points. Awardees and partners used various communication strategies to maintain their relationship; these include holding standing meetings, participating in training, collaborative care conferences, speaking at partner staff meetings, networking calls, and interorganizational work groups. Such activities instilled a collaborative approach and fostered continued partner buy-in and engagement. For some awardees, having shared office space with partners and working in close proximity fostered relationship building.

2.1.3.1 HIT Implementation

In the first annual report, we detailed the varied challenges in implementing HIT innovation components. For example, the time required to develop new HIT systems to support innovations did not necessarily match the time available to implement innovations and suboptimal or non-existent data-sharing capabilities hampered awardees' abilities to implement some HIT components of their innovations. Further, several awardees encountered difficulty in establishing stakeholder buy-in, aligning systems with clinical workflow, and managing functionality glitches.

To assess the frequency and severity of HIT implementation challenges, we included a section in the second AASF addressing these concerns (see ***Table 2-1***). Although HIT challenges are frequently expressed in FLE reports, and are no doubt frustrating to awardees, by the second year of the evaluation, FLE responses to the second AASF suggest these challenges are likely not as severe or as universal as the narrative reports suggest. Data standardization and interoperability emerge as major challenges for 14.6 percent and 12.2 percent of interventions respectively, and building out or installing HIT to support the innovation was a modest challenge for 27.6 percent of awardees. Nonetheless, by the end of their second year of award the challenges faced by most awardees in implementing HIT were considered by their FLEs to be modest, small, or not existent.

Table 2-1
FLEs' assessment of Challenges in Implementing Health Information Technologies[a]

	Major challenge	Modest challenge	Small challenge	Not a challenge	N/A	Unable to assess	Major or modest challenge	Small or not a challenge
Selecting or designing HIT to support the innovation	8.9%	21.1%	22.8%	23.6%	22.0%	1.6%	30.1%	46.3%
Building out or installing HIT to support the innovation	9.8%	27.6%	20.3%	18.7%	21.1%	2.4%	37.4%	39.0%
Identifying, hiring, or obtaining vendor support for innovation HIT requirements	4.9%	10.6%	12.2%	32.5%	30.9%	8.9%	15.4%	44.7%
Data standardization across systems	14.6%	18.7%	15.4%	18.7%	27.6%	4.9%	33.3%	34.1%
Interoperability across organizations	12.2%	18.7%	13.0%	16.3%	38.2%	1.6%	30.9%	29.3%
Alignment of HIT with clinical workflow	8.9%	20.3%	26.8%	17.9%	22.8%	3.3%	29.3%	44.7%
Acceptability of the HIT by frontline staff	6.5%	19.5%	24.4%	24.4%	20.3%	4.9%	26.0%	48.8%

[a] Data obtained from FLEs responding to the 2nd Annual Awardee Summary Form

In year 2, the challenges focused less on adoption of the HIT and more on refining its implementation. For example, several awardees created EHR versions of clinical tools but had trouble getting them to function correctly within the EHR. This required awardees to engage in constant "minor tweaking." For awardees this often meant "getting in the HIT development queue" (meaning IT requests would be addressed in the order received by the IT department) and dealing with unintended consequences, which could potentially lead to safety concerns. Another awardee met with patient resistance to engage in video teleconsultation, when it involved showing of his or her body by video to the teleconsulting physicians. Staff had mixed feedback

regarding the utility of some HIT that was adopted, and in some cases this impacted staff fidelity to implementation.

Although HIT challenges persisted in year 2, maturing HIT capabilities yielded benefits. Some of these benefits are related to the HIT innovation components themselves, while others are related to HIT activities occurring within the organization more generally. For example, one awardee implemented two HIT components specific to the innovation: population management software and a cost reporting software. As a result, the awardee was able to overlay a third analytic component that allowed two new tools to give providers, practices, and communities (patient-centered medical neighborhoods) real time data on patient referral patterns and provider efficiency related to specific conditions. These data identified that cost of care was more highly correlated with specialty care than primary care. This led the awardee to focus discussions with community leaders to spread the program to specialty, rather than primary, care providers. Another awardee's organization switched to a new EHR during the implementation period; this switch facilitated several process improvements in the primary care clinic and was an opportunity to reexamine and standardize some of their workflow processes, leading to greater clarity regarding the roles and responsibilities of each staff type.

2.1.4 Fidelity, Reach, and Dose

Measuring fidelity, reach, and dose helps evaluators make judgments about scope and conformity of the innovation to their original design. **Our review of the FLE second annual reports did not yield any new findings with respect to measuring these outputs, indicating that awardees continued to struggle with the same issues and challenges in year 2 as they did in year 1.** For example, innovations are not designed using specific evidence-based models and are instead "flexible by design," which makes assessing fidelity challenging and in many cases inappropriate. Fidelity as a construct for measuring implementation effectiveness may only be applicable to awardees implementing evidence-based interventions that have matured to the stage of being manualized, such that a clear understanding of what the intervention consists of, including who should be receiving it and at what dose.

By their nature, and through their use of rapid-cycle evaluation, most innovations were designed to be iterative and adaptable to meet the needs of the patient population or culture of the institution. Ensuring innovation fidelity across partners was sometimes challenging because of different organizational structures and processes, and occasionally different patient population needs. Awardees found that allowing for flexibility and adaptability of the innovation was well received by partners but acknowledged the variability introduced by such an approach. Nonetheless, FLEs rated interventions as delivering innovation components to a great extent as intended and at the prescribed level of quality for 63 of 123 interventions (51.2 percent), and at the intended level of intensity and frequency for 57 of 123 interventions (46.3 percent). No interventions were identified as not at all delivering innovation components as intended and at the prescribed level of quality, and only 2 interventions were rated as not at all delivering components at the intended level of intensity and frequency.

FLEs continued to have difficulty accurately evaluating reach. Because innovations had multiple components with different patient populations, multiple implementations sites, and indirect innovations, awardees could not easily establish appropriate numerators and

denominators for calculating reach; awardees sometimes lacked data sources or access to data sources for this calculation. For instance, awardees may not have access to data from private payers to add to the total number of eligible patients.

Measuring dose also proved troublesome. Innovations, by definition, entail implementing something new, and most HCIA innovations did not implement established programs or curricula with a significant evidence base for a "minimally effective dose." Because awardee staff could have multiple interactions with patients via a range of modes (i.e., in-office, in home, via telephone), many awardees could not capture the differences in intensity or quality of patient contact that would help identify a "meaningful dose" of an intervention.

2.1.4.1 Sustainability

During year 2, HCIA awardees considered their plans for sustainability and reported how they might continue the innovations or portions of the innovations at the end of the funding period. Awardees pursued an array of approaches for sustainability and these approaches varied depending on the nature of their innovations, state policy and political environments, and the demonstrated effectiveness of the innovations. For example, some implementation leaders negotiated with payers to fund the innovation; however, payers wanted to see evidence on the value of the innovation or a return on investment. For awardees implementing new care coordination and care management approaches, a lack of reimbursement for care coordination services from public and private payers will likely limit sustainability. Other awardees hope to join Accountable Care Organizations (ACOs) or other payment reform pilots occurring in the state, pursue Medicaid waivers to pay for HCIA activities, or receive financial support and buy-in from their provider site to sustain their innovation. Awardees also turned to public and private grant sources to maintain innovation components. Although we did not assess plans or activities associated with sustainability in the second annual AASF, we did ask FLEs if, based on their knowledge of the processes, activities, and management supporting this innovation, the innovation could be successfully disseminated for widespread adoption and implementation. Of the 123 interventions, FLEs believed that 96 (78.1 percent) could be successfully disseminated for widespread adoption and implementation. Based on evidence from thematic coding of FLE reports, we discuss several awardee strategies for sustainability below.

Innovation and organizational leaders increasingly focused on strategies for sustaining some or part of the innovations. These leaders primarily explored funding or reimbursement for an innovation's services after HCIA funding concluded. In one awardee, organizational leadership recognized the value of the innovation and initiated conversations with innovation leaders to strategize getting approval from payers to include innovation components in their bundled payments. Another awardee's innovation leaders assessed opportunities for value-based payment models with commercial insurers and state Medicaid officials within its pediatric ACO. In several awardees, innovation staff had conversations with organizational leaders to identify possibilities for reimbursement for services when state or federal policies impeded reimbursement. For example, federal policy does not recognize pharmacists as Medicare Part B health care providers, and one awardee is considering internal funding for its pharmacy teams after HCIA funding ends. Another awardee's innovation leader worked with organizational leaders to configure the EHR to allow pharmacists to use available medical codes to bill for services.

Some awardees secured financing from Medicaid and commercial health plans to sustain the innovation, but payers wanted to see a return on investment or improvements in health outcomes before entering contractual agreements. Some Medicaid managed care organizations and commercial health plans expressed an interest in funding HCIA innovations in select states. These private insurers agreed to sustain the innovation by paying for EHRs or medication management systems, financing care management fees or other health care services, or paying for various types of data modeling designed to measure outcomes and costs. One awardee secured funding to support HCIA staff, but many awardees identified sustaining support for staff salaries as an obstacle for sustainability (see theme below). Many awardees noted the importance of demonstrating the value of the innovation to the payer through impact assessments or program evaluations.

Awardees turned to state and federal financing streams to sustain some innovations. A few awardees noted sustaining their innovations through 1115 Medicaid waiver programs such as the Healthcare Transformation waiver and the Delivery System Reform Incentive Payment (DSRIP) program. For example, one awardee in New York could incorporate a provision in the state's DSRIP proposal to scale its intervention to seven additional sites. Other awardees discussed negotiating with the state to include provisions in their 1115 waivers that would provide reimbursement for certain health care professionals such as home care specialists and peer support counselors.

Participating in future payment reform pilots or demonstrations—mainly ACO pilots—are a viable mechanism for maintaining care coordination innovations. Several awardees sought to participate in future payment reform models tested in their state, such as ACOs, bundled payment pilot programs, total cost of care models, and others. ACOs were mentioned most frequently by awardees. Awardees noted joining ACOs currently in practice, and others are submitting applications to become their own ACO. Through ACOs, awardees could pay for the salaries and benefits of care managers and finance a telehealth care management system. According to one awardee, applying to become an ACO would likely help them leverage additional grant funding in the years to come. A few awardees also reported that the support of ACOs could enable them to scale their innovation to other counties or regions.

Large provider institutions plan to sustain many HCIA innovations after the grant concludes, particularly those innovations that are fully integrated into clinical workflows. Some awardees in large provider institutions garnered long-term financial support from organizational leaders during implementation. Awardees noted that these innovations have been successfully embedded in the culture and workflow of the institution making it easier to obtain long-term financial support for the intervention.

Some awardees are exploring creative, non-traditional funding sources to sustain their innovations. Several awardees applied for other grants to maintain their HCIA innovations, including grants from private foundations, public agencies, and universities. Others are considering charging dues or fees from partner sites or instituting small charges for beneficiaries to preserve the innovation in future years. One awardee was considering selling its training model to other organizations and universities to help scale and sustain the innovation.

Awardees identified lack of reimbursement for care coordination services and new staff types as key challenges to sustainability. Many awardees attributed challenges to ensuring sustainability to a lack of reimbursement for care coordination services or the inability of certain health care professionals to bill for health care services related to chronic disease or care management. For example, one awardee noted that a lack of reimbursement for medical services delivered via telehealth would likely impact sustainability. A few awardees indicated that they would have difficulties funding non-clinical staff (i.e., health coaches, patient navigators, community health workers). Although these staff help physicians and nurses perform a variety of care coordination activities, they cannot bill for these services. On a positive note, a few awardees did note that evolving Medicaid redesign and payment reform undertaken in their state could present a viable funding source for these types of staff and services in future years.

Partners played an active and strategic role in planning for sustainability by agreeing to adopt and integrate key innovation components into their existing work. They also secured additional funding or policy changes to continue activities beyond the HCIA award period. For example, one awardee and its partners agreed to continue providing the services and are collaborating to seek out public and private funders to support their ongoing efforts. Another awardee and its partners are in discussions with Medicaid managed care organizations about plans for reimbursement of the community health worker component of its innovation.

2.1.5 Context

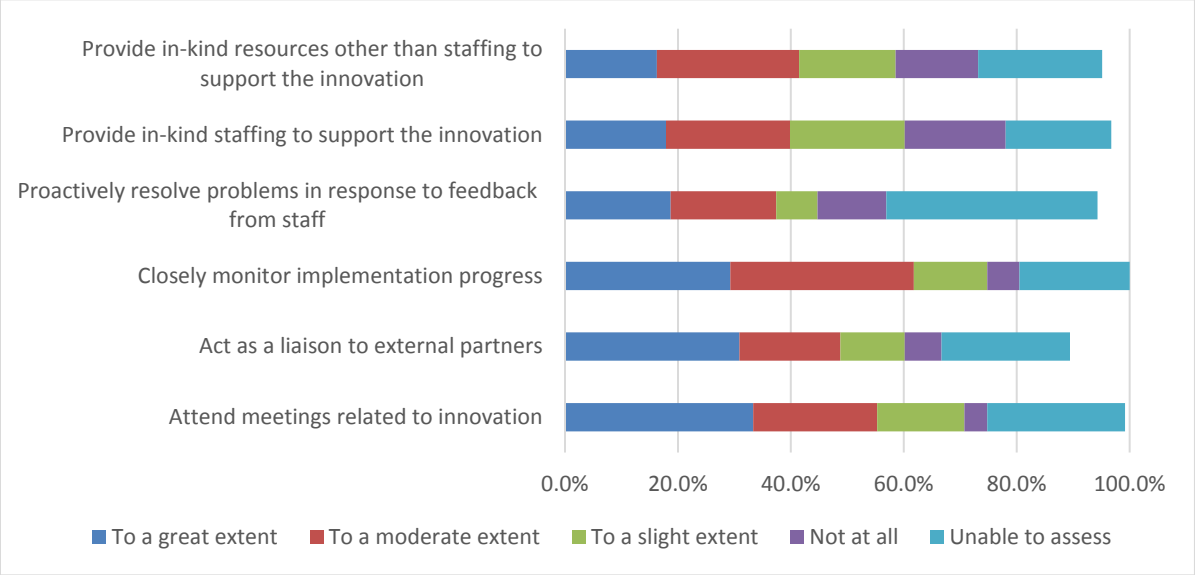
2.1.5.1 Leadership

This section describes the qualities of successful leaders and their efforts to sustain or scale the innovations. We discuss two types of leaders: innovation leaders and organizational leaders. The former are the innovation Project Director/Principal Investigator or other innovation team staff. Organizational leaders include the person or persons to whom the innovation Project Director/Principal Investigator reports and are individuals with the power to make resource allocation decisions within the organization (e.g., the CEO). In the first annual report we detailed how awardees' organizational and innovation leaders supported innovations by generating buy-in, engaging stakeholders, and fostering an environment amendable to implementation. Organizational leaders also marshalled and allocated resources. Leadership styles that encouraged feedback from staff and partners created a positive learning culture. That culture helped awardees to evolve and adapt the innovations to meet staff and patient needs.

Awardees identified characteristics of champions and strategies for cultivating them when they did not exist. Most effective leaders were well-known and trusted because of their years of working with their colleagues and partners. Innovation leaders built those relationships and experiences over years of work and collaboration, which gave them knowledge, respect, and legitimacy. Organizational leaders demonstrated their support for the innovations by engaging in innovation activities. The AASF survey responses indicated that organizational leaders demonstrated varying levels of involvement in the innovations. As illustrated in ***Figure 2-4***, although many FLEs did not assess the role of organizational leaders, those which did reported that organizational leaders attended innovation meetings, monitored progress, helped resolve problems, provided resources to support the innovation, or a combination of these. Their involvement signaled to staff and partners that they support the innovation, which helped foster

buy-in. The FLEs reported that when an organizational or innovation leader was well-respected, the innovations they supported would benefit. One awardee's well-respected Principal Investigator used her influence to gather support for the innovation. Similarly, at one awardee's site, the organizational leader had worked there for years and had strong institutional support. As a result, the awardee received immediate support from other organizational leaders when she proposed a new innovation component.

Figure 2-4
Involvement of Organizational Leaders[a]



[a] Columns may not total to 100 percent due to missing data. Data obtained from FLEs responding to the 2nd Annual Awardee Summary Form.

2.1.5.2 Organizational Characteristics

This section describes organizational characteristics at the awardee and site levels that affected the implementation of the interventions. In the first annual report we called attention to the value of integrated organizational structures, a culture of innovation, and experience with transformation. Some awardees and sites with integrated organizational structures (e.g., shared data systems) and streamlined administrative processes could more easily track patients and obtain approval up the chain of command to make changes. Having a culture of innovation encouraged staff to try new approaches to care delivery, and accustomed them to initiating change. Finally, awardees that piloted the innovation or had implemented similar programs encountered fewer challenges and delays in HCIA implementation, because those earlier efforts allowed them to anticipate potential obstacles and plan for them. Having experience with implementing similar programs emerged as the primary theme in the FLE's second annual reports, but with new details.

Awardees with organizational capacity before launch of the innovation could overcome unforeseen challenges and may have more sustainable and scalable programs in the end. Organizational capacity includes resources, a history of implementing similar programs, experienced staff, established partnerships, and infrastructure (e.g., HIT, administrative and

technical support) that enable an awardee to rollout health care transformation. The FLEs' second annual reports commented on how organizational capacity (or lack thereof) facilitated (or hindered) implementation for many awardees, partners, and the practices implementing the innovation. Two measures of resource capacity and commitment (providing in-kind staff and other resources) showed consistent relations with implementation effectiveness. Innovations with the capacity to provide in-kind staff and other resources were implemented more rapidly and more completely than innovations in organizations lacking those resources.

Prior experience with the innovation also improved innovation implementation. Some practices affiliated with one awardee already had shared decision-making programs, which reduced the level of effort needed for implementing a new shared decision-making program because the awardee knew how to roll out implementation and practices had experience engaging other patient types in shared decision making. Practices and awardees that lacked that capacity or infrastructure faced greater challenges getting started. One primary care redesign awardee built most components of its HCIA initiative from scratch, and staff had limited experience with, or understanding of, transformation processes. Starting behind led to additional delays in implementation (e.g., delays in establishing contracts with partners) and under-enrollment of patients. Although having experience benefited most awardees, in a few instances that experience also created some difficulties. For one awardee, implementing sites already had existing data collection tools and resisted adopting new tools. The sites found it time consuming to adapt those tool for HCIA efforts.

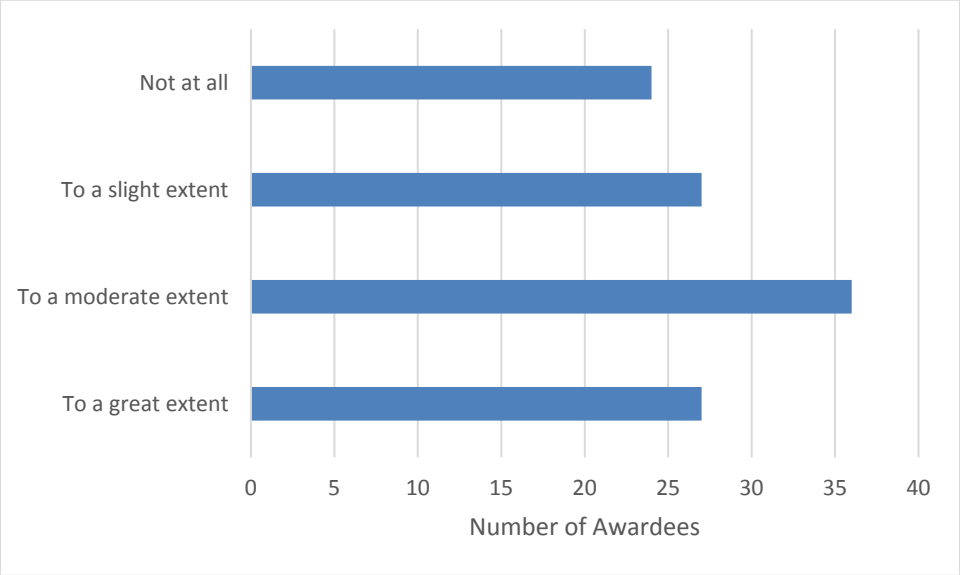
Staff familiarity with transformation facilitated implementation at the practice and awardee level. For clinical staff implementing an evidence-based care management model with one awardee, their experience with mental health care coordination and integration efforts laid the groundwork for their work under HCIA. They viewed their HCIA work as a natural extension of their ongoing activities and found it easy to incorporate HCIA processes into their normal care delivery structure. Staff comfort and familiarity with transformation eased integration of new roles and processes.

Adequate staff capacity to manage change processes made implementation a smoother process, but some awardees' lacked the staff necessary to implement their innovations. One awardee grappled with regional staffing shortages, and many of its health navigators were recent college graduates with limited experience working with patients and no training in health care navigation. With few experienced staff to draw on, this awardee faced significant capacity challenges in implementation and monitoring their innovation.

In the second AASF, FLEs assessed the extent to which an awardee's management team had experience with implementing similar programs at a similar scale. In total, 90 of the interventions (79 percent) had experience with implementing similar programs at a similar scale. FLEs indicated that 26 interventions (21.1 percent) had this experience "to a great extent," while another 33 were experienced "to a moderate extent" (26.8 percent). However, according to FLEs, nearly one in five interventions (24) were delivered by awardees with no prior experience (see ***Figure 2-5***). Awardees who had implemented similar programs were generally able to implement their innovations more rapidly than those with less experience ($r = 0.66$) and FLEs reported that this experience was associated with better success in delivering innovation components at the intended level of intensity and frequency ($r = 0.47$).

Awardees often drew upon their experiences to inform plans for sustainability and scalability. One awardee took the lessons it learned from implementing its program in three counties and used that knowledge to avoid some hurdles when expanding it to two additional counties. Awardees with the internal management and resources to implement interventions had an advantage in sustaining or scaling their initiatives in changing and sometimes uncertain financing environments. One such awardee used HCIA funding and its own internal investments to train staff and evaluate the initiative's impacts to make the case for scaling-up to additional sites post-HCIA. As noted above, demonstrating improvement in outcomes can build a case for sustaining or scaling a model.

Figure 2-5
Experience with implementing similar programs at a similar scale[a]



[a] Data obtained from FLEs responding to the 2nd Annual Awardee Summary Form.

2.1.5.3 Team Characteristics

This section discusses team functioning and dynamics that impact implementation. Key themes explored in RTI's first annual report included clarity around roles and responsibilities, integration of new staff, approaches to enhance coordination among team members, and workflow impact. In year 1, many HCIA awardees shifted and redefined many roles for existing staff as they implemented their innovations. In some cases, redefining roles allowed current staff to work more effectively and at the top of their license. In other cases, a lack of role clarity created competition and confusion among staff. In these cases, teams needed clear guidance from leadership around expectations to function effectively. Additionally, taking the time to educate staff about the innovation and its value enhanced implementation. In year 1, awardees reported certain barriers to team functioning such as staffing new employees in locations separate from current staff and difficulties adjusting to inconsistent communication styles and preferences. Not surprisingly, effective communication and coordination within and across teams was considered critical to success.

Some awardees experienced limitations using non-licensed or mid-level clinical staff to perform responsibilities related to complex patient care. At the start of implementation, many awardees experimented with using non-licensed personnel or mid-level health care professionals to perform certain responsibilities related to care coordination and disease management. However, as the innovation progressed, some awardees learned that non-licensed staff did not always have the necessary expertise or clinical skills required to address the medical needs of complex patients. For example, one awardee initiated its innovation tasking medical assistants with conducting home-based rapid response triage. The awardee received feedback that the medical assistants may miss critical aspects of a patient's health care needs and, as a result, transferred the responsibility for triage to registered nurses and eventually nurse practitioners. Another awardee reported having to deliver additional training to its medical office assistants (MOAs) when it learned that the MOAs could not conduct appropriate chart reviews for diabetic patients. Determining the most appropriate roles and functions for non-licensed or less experienced staff can be challenging when providing care to patients with complex conditions.

Cross-training staff to fulfill multiple roles and functions can enhance implementation and efficiency. Several awardees saw the value of cross-training clinical and non-clinical staff to perform multiple functions. Cross-training can minimize fragmentation of care or delays in services that might occur if teams are highly specialized and one staff member is absent. Awardees also noted that cross-training occurs when budgets are tight. One awardee reported training its medical care coordinators—who typically focus on clinical nursing tasks—to identify psychosocial concerns that may prevent a patient from seeking necessary medical care. Another awardee described training staff to perform administrative and management functions in addition to their clinical responsibilities. For example, all nurses and social workers at one awardee perform follow-up calls, answer questions from study patients, and understand how to interpret a care plan to assist in delivering care coordination services for participating patients. Lastly, in addition to performing clinical tasks, care navigators at another awardee negotiated contracts with insurance companies and educated payers about the importance of new services that were not typically covered.

Physical colocation of staff fosters strong working relationships and facilitates better coordination of care, which can translate into enhanced health outcomes for patients. Several awardees noted that colocating staff in the same physical location—even for only one day a week—can improve communication and foster positive relationships among team members. One awardee embedded its care managers in primary care practices one half day each week, which helped these staff members build trust with the local physicians and practice staff. Another awardee intentionally moved its offices for all core innovation staff to one central location to ensure that all team members could participate in daily team huddles, meetings related to patient cases, and trainings as appropriate. The awardee credits this move with enhancing team effectiveness and ultimately improving the quality of care delivered to patients. Regular face-to-face contact among team members also improved efficiency by minimizing the time clinicians spend reviewing and interpreting electronic and written communications.

Achieving the right balance between communicating electronically versus communicating in-person affects team functioning and care coordination. Several awardees

acknowledged the role EHRs, and other technological tools, play in improving communication among team members and enhancing patient care. Technology helped different types of health care providers communicate more effectively within a health care system or clinic, as well as externally with members of the care team who may be located off site. Technology also supported physician access to information about patient status and treatment. However, relying solely on technology without any in-person interaction could disrupt patient care if important information is missed or ignored. For example, one awardee cautioned against relying too much on e-mail communication when caring for patients in an intensive care unit because it does not allow physicians to ask questions and have a dialogue about a patient. Several awardees indicated that having regular in-person team meetings or daily huddles in concert with using technology to communicate about patient care ensured high quality care.

2.1.5.4 External Context

In this section, we describe the factors external to the innovations that impacted implementation. In the first annual report we described how national, state, and local policies helped and on occasion hindered implementation, and how changes in the health care market and payment models altered the implementation environment. Changes to policies impacted implementation of the innovations by influencing the eligible patient population, altering the public supports available to vulnerable populations, and creating an environment amenable to improving care. These changes adversely affected over one-third of awardees. Changes in the health care market (e.g., mergers, emergence of ACOs) also impacted the innovations, often by diverting attention from the innovations while staff worked through organizational changes and new partnering agreements. Awardees also struggled to provide value-based care with fee-for-service (FFS) payment models, and some failed to obtain reimbursement for all services. Those themes arose again in this year's data but with some differences.

States' decisions around Medicaid expansion affected the patient population and created capacity challenges. Medicaid expansion factored more prominently in the FLEs' second annual reports than in their first annual reports. Awardees across four types of FLE portfolios—behavioral health, complex care, community, and hospital—faced challenges with adapting to patients' needs as a result of states' decisions on Medicaid expansion. For example, Texas and South Carolina did not expand Medicaid, which resulted in large numbers of uninsured patients for HCIA awardees in those states. One such awardee struggled to meet the needs of its patients because most were poor and uninsured and required services beyond what HCIA-funded physicians could provide. Awardee staff cautioned that because patients' health-related social needs exceeded the services HCIA-funded physicians could provide, the innovation may not demonstrate improved health outcomes for this complex population. Another awardee focused on elderly Medicare patients because its state had not yet expanded Medicaid. The awardee found that the needs of this elderly population required them to use significant time and resources addressing patients' basic social needs (e.g., running water) before staff could begin treating their health conditions. The state later chose to expand Medicaid but only after the awardee started its HCIA innovation. Awardee staff explained that if the state had expanded sooner they would have enrolled younger, more moderate-risk patients, presumably with fewer complicated health needs. They indicated that a more balanced patient population would have allowed them to demonstrate greater program effectiveness.

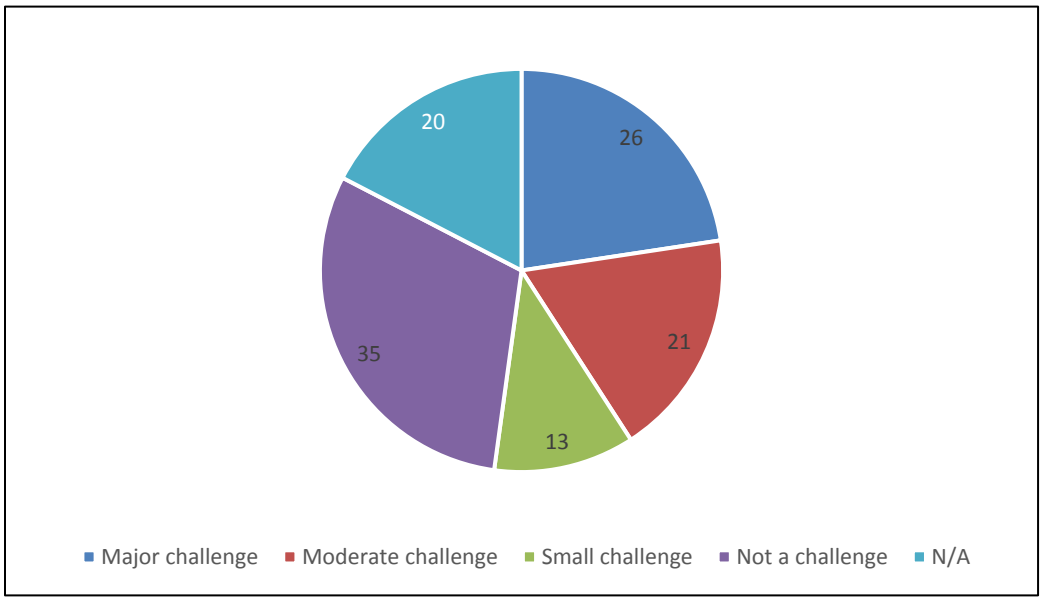
Some awardees struggled to meet the demand for services in states that expanded Medicaid. Expanding Medicaid increased a state's insured population, and the demand for health care services increased dramatically. However, the supply of providers did not concomitantly increase. One awardee and its partner organizations struggled to build their staff capacity to accommodate the increased demand for mental health services that arose after Medicaid expansion. Some awardees also had difficulty meeting the complex health needs of newly insured patients; these patients may have had pre-existing conditions that may have worsened or gone undiagnosed while they were uninsured. Providing and managing care for newly eligible people with untreated and complex health conditions created an unanticipated burden on awardee staff.

For one awardee, expansion also complicated the pool of patients eligible to receive the intervention. As part of its health care expansion efforts Illinois altered its policy to automatically enroll patients dually eligible for Medicare and Medicaid into a Medicare Advantage (MA) plan. As a result, one awardee struggled to find eligible patients for its intervention to provide care to high-risk patients before, during, and after hospitalization. Some enrolled patients became ineligible for the intervention when their MA coverage was initiated, and the awardee struggled to find individuals to enroll who were not recently part of a MA plan.

State support (or lack thereof) for alternative payment models (APMs) can facilitate (or hinder) implementation. The coordinated, comprehensive care approaches tested under HCIA generally did not align well with the current reimbursement practices and a FFS payment model. As noted in our first annual report, holistic care requires alternative approaches or payment models, such as shared savings, bundled payments, pay for performance, and other value-based strategies that may support the costs of providing comprehensive care. State support for alternative payment models could facilitate implementation and encourage sustainability. A small number of awardees noted the ways in which those APMs allowed them to move forward. One awardee reduced hospital admissions and emergency department visits, but suffered a financial loss under an FFS model. In 2014, the awardee's state legislature mandated an APM that rewards hospitals for improved patient outcomes. Hospital administrators in this awardee report that their efforts to reduce admissions and emergency department visits should now generate savings under the new APM. For another awardee, many participating practices joined local ACOs. Awardee staff said that move provided them with the financial support they needed for care management and transformation.

Awardees struggled to get reimbursement for all of their services, and some awardees absorbed costs. Results from the AASF revealed that the level of reimbursement for services was a challenge for nearly half of awardees (60, 48.8 percent); and a major challenge for 26 (21.1 percent; see ***Figure 2-6***). The primary challenges pertained to payment models that did not support value-based care, and a lack of reimbursement for telemedicine and services from certain types of health care workers. To compensate for this lack of reimbursement and in support of their innovations, 49 of 123 interventions (39.9 percent) were supported with in-kind staffing by awardees and 51 of the 123 interventions (41.5 percent) were supported with in-kind resources other than staffing.

Figure 2-6
Level of Reimbursement for Services[a]



[a] Data obtained from FLEs responding to the 2nd Annual Awardee Summary Form.

In places where value-based payment models were uncommon, some awardees absorbed costs. One awardee said transitioning practices to value-based care under a volume-based FFS payment model was difficult; awardees absorb the costs of comprehensive care services that physicians could not yet bill for under the existing payment model. Similarly, because New York State did not reimburse for case management services, one awardee paid for the labor costs for case managers in its New York site. Of note, awardees' implementation of round 1 HCIA innovations largely occurred before the availability of Medicare reimbursement for chronic care management services, which began January 1, 2015.

Several awardees struggled to obtain reimbursement for telehealth services. These awardees used telehealth technologies to connect with patients in rural or medically underserved areas, but their reimbursement options were limited. In one state, an awardee could not receive reimbursement for psychiatric services provided via telehealth. In another awardee, the costs associated with collecting the information required for reimbursement exceeded the actual reimbursement, which discouraged billing for telehealth services. Over the course of the HCIA funding period, Colorado and Idaho passed legislation supporting reimbursement for telemedicine. An awardee with implementing sites in one of those states reported that the new legislation gives them the opportunity to bill for telehealth. Another state's legislators established parity regulations to require commercial payers to cover telemedicine at a level equal to what is provided for in-person services, which facilitated the delivery of an awardee's behavioral health services.

Lastly, some awardees could not obtain reimbursement for services provided by particular types of health care workers. One awardee's innovation involved home visiting conducted by CHWs, but the awardee discontinued the CHW component because the state did

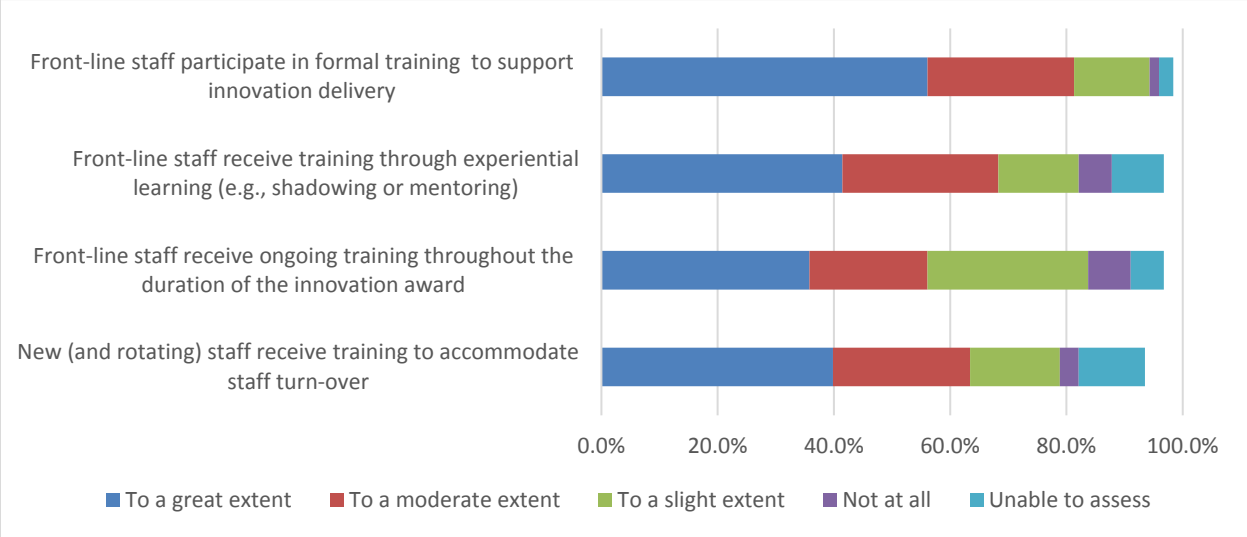
not adopt a CMS rule allowing for the Medicaid reimbursement of preventive services conducted by CHWs. Another awardee struggled with sustaining its pharmacist-based programs when federal policies did not recognize pharmacists as Medicare Part B health care providers and thus could not receive reimbursement for those services.

2.1.6 Workforce Development

2.1.6.1 Workforce Training

The development and deployment of new workforce to support health care innovation requires training to expand the roles of existing staff positions and integrate new staff members (see ***Figure 2-7***). In this section, we summarize key findings from previous reports that persisted and evolved as prominent themes at the time of follow-up site visits. Additionally, we highlight new findings related to workforce training and education. Several awardees continued to report a range of modalities to deliver staff training in support of innovation implementation, including formal training such as lectures, webinars, and workshops, and informal training through one-on-one shadowing and mentorship. Consistent with previous findings, staff often considered informal modalities to be more effective and relevant than didactic teachings, and awardees adapted training programs accordingly. For example, awardees ended up redesigning trainings for CHWs or paramedics to be more experiential. Furthermore, staff also responded well to training opportunities that accounted for skills and experience. Some licensed staff, such as pharmacists or nurses, came to the innovations with extensive background and initially reported that trainings duplicated what they already knew. Subsequently, awardees modified trainings to acknowledge existing skill sets. Trainings that were perceived as less applicable to a staff member's role were received with greater ambivalence.

Figure 2-7
Percent of Interventions Providing Staff Different Forms of Training[a]



[a] Totals may not equal 100 percent due to missing data and "Not Applicable" responses. Data obtained from FLEs responding to the 2nd Annual Awardee Summary Form.

Staff models continued to be a barrier to implementing workforce trainings. Many clinical roles involve shift work or rotations between hospital units, such that many staff cannot attend trainings. For example, residents rotate units every month, and nurses may temporarily float into units because of transient high patient volumes. These temporary staff may know little about the innovation and any training that occurs may be incomplete. However, as discussed in "Team Characteristics," awardees continued to identify cross-disciplinary training as a means of improving care and collaboration. By including staff with different licensure and backgrounds in trainings, staff could anticipate each other's needs. Additionally, cross-disciplinary training allowed team members fill in for other staff members if needed and overall improve respect for team members as a whole.

Many awardees focused training content on the skills needed to optimize patient interactions. Social workers helped staff interact with patients by showing team members how to communicate and manage patients with social and behavioral health needs. Several awardees highlighted the impact of motivational interviewing training on a range of staff, including pharmacists, nurses, and other support staff. Motivational interviewing promoted closer relationships with patients and improved staff confidence in engaging patients. One awardee attributed low attrition in the innovation to improved engagement through motivational interviewing.

Training not only transferred knowledge and developed skills, but established a level of comfort among staff tasked with implementing in the innovation. For staff to expand their roles, many required formal trainings on medical topics and skills. However, knowledge alone was not adequate to generate the confidence needed to apply these lessons in practice, particularly when interacting with a new, complex patient population, such as the critically ill or mentally ill. A supportive environment improved staff confidence in applying newly acquired knowledge; awardee staff even reported feeling more empowered by their training because it enabled them to take on new roles and engage patients (see section 2.1.6.3 for more information). Some awardees provided additional trainings to address uneasiness around unfamiliar clinical situations, offered additional observation or shadowing experiences, and performed assessments of comfort.

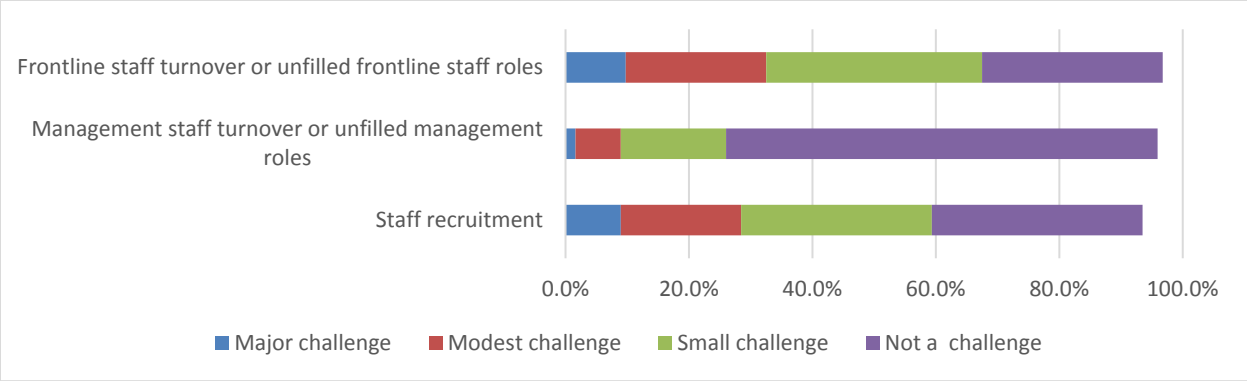
Awardees identified ways to make training more replicable and less resource intensive. Looking ahead, many awardees considered how to adapt training so innovations could be sustained or scaled up. Awardees often transitioned from in person, face-to-face training to video-recorded instruction or Web-based formats. For some awardees, implementation of these prerecorded formats occurred during the funding period; others planned to use those forms of training in the future. A few awardees also mentioned additional solutions, such as narrowing the scope of training activities, centralizing training to one organization, and employing a train-the-trainer approach to expand the base of trainers beyond licensed providers.

2.1.6.2 Recruitment, Turnover, and Retention

This section highlights key themes and findings related to recruiting and maintaining a stable workforce to support the innovation. Although the year 1 challenges persisted, several awardees highlighted how hiring staff to fit the demands of the innovation limited staff turnover and improved workforce retention. By year 2 recruitment, retention, and turnover challenges were considered mostly minor or non-existent (see ***Figure 2-8***). Nonetheless, several awardees

continued to face labor market constraints that affected recruitment and retention. Awardees in remote locations struggled with a limited pool of local talent. Additionally, patient volumes in these locations were too low to support a full complement of staff for team-based care financially. As a result, one or two staff members ended up performing multiple functions. Moreover, several types of staff—including mental health providers, nurses, critical care staff, and information technology specialists—were in high demand and could command competitive salaries with potentially lower workloads at other organizations. Some innovations involved specialized training for new staff, such as motivational interviewing or Lean certification. Offering additional education could improve recruitment and retention, but in some cases newly trained staff left for more lucrative employment or higher levels of medical training. One awardee took steps to limit the loss of new staff by requiring them to pay back the cost of training if they did not stay for at least 6 months.

Figure 2-8
Percent of Interventions Experiencing Challenges with Turnover, Retention and Recruitment[a]



[a] Totals may not equal 100 percent due to "Not Applicable" or Missing Data. Data obtained from FLEs responding to the 2nd Annual Awardee Summary Form.

New and established staff continued to struggle with the emotional burden of caring for vulnerable populations (i.e., those with complicated medical problems, behavioral health conditions, social needs). Stressful environments like the emergency department, and the demands of extending work hours to evenings and weekends also contributed to staff burnout and increased staff turnover. As in previous reports, turnover among clinical and non-clinical staff led to delays in implementation. However, a few awardees reported that turnover allowed for the hiring of clinical staff and leadership that were more supportive of the innovation.

Finding the right fit enhances retention and staff satisfaction. Several awardees stressed the need to find the "right" person for the job, referring to specific skills, personality traits, or both. Some innovations required staff with experience in specialized populations, such as patients with mental health conditions or chronically ill children. Turnover was more likely when staff lacked the requisite experience with these challenging populations. Several awardees highlighted the importance of certain personality traits and interpersonal skills. Across the spectrum of clinical providers from unlicensed staff to physicians, awardees reported the need for staff that were self-motivated, compassionate, and committed to the mission of working with

complex populations. Furthermore, over the course of the funding period, awardees learned to look for strong communication skills and became more discerning during the hiring process. For example, a few awardees realized that though innovations were set in the fast-paced environment of an emergency department, a traditional emergency room provider—accustomed to juggling several acute cases—might be inappropriate for an innovation focused on follow-up care. Although finding the "right" staff member can require more recruitment time and effort, several awardees reported improved job satisfaction, better staff retention, and limited turnover when they did.

Innovations created unanticipated demands. Many times, the amount of work required to implement innovations and meet CMS requirements exceeded awardee's initial expectations. Awardee staff identified high demand for behavioral health and social services most frequently, though also reported high telephone call volume for transitional care and high additional administrative burden. For example, in one awardee, children referred to the innovation often had a sibling with a behavioral health condition, which doubled the anticipated patient load; in another instance, patient interest in the innovation exceeded implementation staff expectations, which meant increased call volume for requests for information. For two awardees, adding a new assessment meant conducting the assessment and completing associated forms, which staff identified as initially burdensome. Awardees often responded to unanticipated workload by hiring new staff across the clinical spectrum to provide behavioral health services or assist with administrative work, and a few employed IT solutions such as automating referrals or redirecting phone calls to specific staff.

2.1.6.3 Staff Satisfaction and Acceptance

In the earlier stages of award implementation, some clinicians resisted innovations, expressing concerns about competition (i.e., losing revenue), suspicion around innovations replacing the traditional model of health care delivery, and frustration around integration of new processes, particularly information technology. Over the course of the award period, staff acceptance evolved with observing the positive impact of the innovation on patients, as well as their own professional satisfaction and development. By the second year, FLEs reported clinician buy-in and engagement was a major challenge for only 10 interventions (13.8 percent) and a modest challenge for an additional 17 (22.8 percent).

Staff satisfaction often hinged on how the innovation impacted workflow. As described in the first annual report, innovations had mixed effects on workflow. FLEs considered workflow redesign to be a major challenge for seven interventions (5.7 percent) and a moderate challenge for another 39 interventions (31.7 percent). When staff reported experiencing burden, they often also reported negative perceptions about the innovations. For example, adding the responsibilities of the innovation to existing clinical duties fueled burnout. As discussed in "Recruiting, Retention, and Turnover," for some awardees the workload for the innovation exceeded expectations, with many more patient interactions than expected or clinical tasks generated from an HIT-based innovation (such as numerous clinical alerts). For innovations where new staff or processes were better integrated into workflow, staff reported decreased burden and more acceptance. Well-integrated care coordinators could take on administrative and lower level clinical tasks, which gives physicians and nurses more time for clinical care. Some

HIT tools improved workflow, reducing duplicative data entry, improving patient tracking, and saving time for pharmacists and nurses.

Satisfaction increased when staff perceived a positive impact on patients. Across multiple settings such as emergency department, primary care, or behavioral health, staff expressed concern that processes like additional screenings or preclinic huddles would disrupt workflow and increase burden. Though innovations had mixed impacts on workflow as discussed above, staff increasingly accepted and reported satisfaction with innovations that improved patient care. For example, physicians may have been dubious about adding a care coordinator or social worker, but their attitudes changed when they saw patient successes, like consistent patient follow-up, improved asthma control, or more effective patient education. One awardee reported non-innovation staff nurses incorporating innovation activities into daily shift work after observing the positive results on patient care. Furthermore, in some awardees staff reported appreciating the opportunity to develop impactful relationships with patients; this opportunity enhanced satisfaction particularly among non-licensed staff that did not have as much patient contact prior to the innovation, such as pharmacy technicians. Nurses and care managers also reported a higher level of satisfaction delivering patient-centered care and establishing deeper relationships with patients, especially in settings outside of the traditional clinical setting.

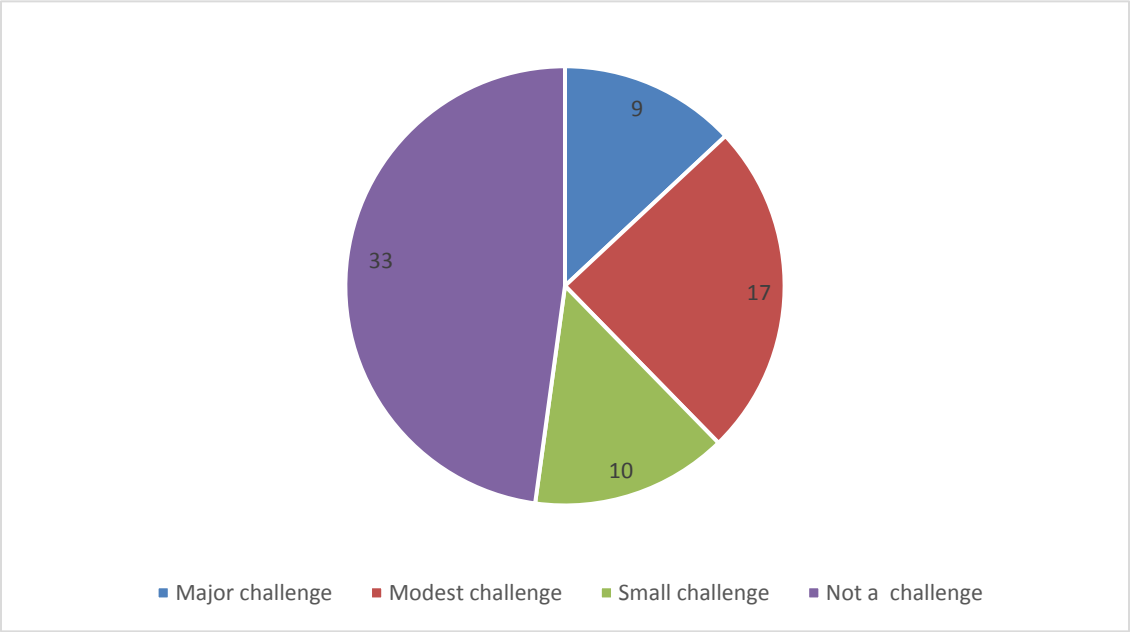
Clinical staff reported feeling empowered in their roles as a result of the innovation. Nurses appreciated additional training in clinical topics, increased critical thinking, and new caregiving responsibilities, such as recognizing sepsis. Practicing at the top of their license also empowered nurses in their interactions with physicians. Similarly, home health aides and pharmacy technicians reported satisfaction with fully using their skill set, compared to traditional roles of pill dispensing, as well as gaining additional skills in patient counseling. Some innovations also expanded the scope of care provided by physician generalists, like primary care or emergency medicine doctors. Although physicians initially expressed some reluctance to providing unfamiliar services such as chronic pain management, stroke care, or behavioral health care, physicians ultimately reported more confidence after gaining experience and receiving training.

2.6.1.4 Community Health Worker—Workforce Development and Integration

As of year 2, FLEs reported that 41 percent of awardees, and 56.5 percent of interventions were using CHWs as part of their innovation workforce. The credentials and education required for CHWs varied by awardee and were highly tailored to the innovation. In this section, we describe themes related to workforce development and integration of CHWs among awardee innovations. In the first annual report we described the roles and responsibilities of CHWs as part of the workforce implementing and delivering various innovations, along with challenges related to integrating CHWs into clinical health care teams. In year 2, awardees often described the role of CHWs as "community insiders" focused on providing emotional, social, and instrumental support necessary to bridge the gap between individuals and their care providers and organizations. However, for some awardees integrating CHWs into their workflow remained a challenge into the second year because of the lack of role clarity in day-to-day tasks and functions (see ***Figure 2-9***). This challenge was particularly evident among awardees without experience working with CHWs, who failed to provide sufficient training to both the care team

and CHWs, or who provided insufficient role and responsibility guidance to the care team and CHWs. Several awardees commented that in retrospect, earlier engagement of practices and providers in the design and planning phases, and more joint education, preparation, and mentoring related to integrating CHWs into clinical teams may have offset some of the challenges experienced. Several new themes related to CHWs emerged in year 2 and are described in the rest of this section.

Figure 2-9
Extent of Challenge Integrating Non-licensed Staff into the Care Team[a]



[a] Data obtained from FLEs responding to the 2nd Annual Awardee Summary Form.

CHW integration improved clinical workflow and enhanced implementation. Several innovations depended primarily on CHWs to coordinate additional patient services, facilitate the flow of patients through various health care settings, and support self-management activities. Several awardees reported quantifiable impacts on workflow, with clinicians spending between 30 and 50 percent less time arranging and coordinating social services and referrals than before the innovation. Additionally, awardees reported improved reach, "better" services, and improved standardization of coordination services as a result of CHW integration.

Working alongside CHWs created a change in mindset and a cultural shift among health care providers whom began to value the role and duties of CHWs in connecting with patients. Several awardees report that a cultural shift among health care organizations and providers occurred over time in attitudes towards and respect for the CHW role. Some awardees attributed this to the experience gained and rapport built between providers and CHWs over the implementation period. Within some awardee organizations, physicians and other clinicians who may have been initially indifferent or even skeptical have become champions for the CHW role. In another awardee organization that had previously used CHWs within the community, the innovation changed the entire dynamic of the relationship between clinical and non-clinical staff.

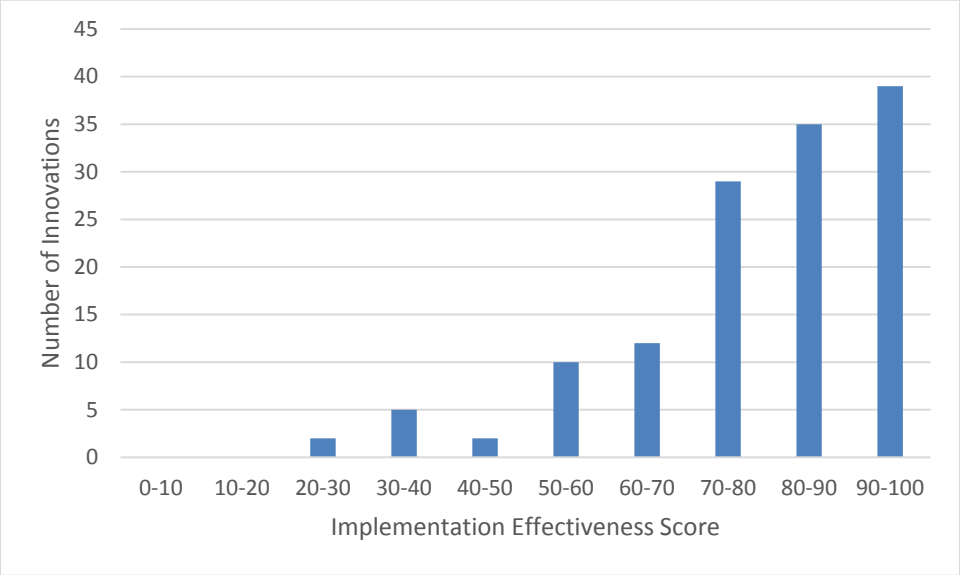
Identifying and responding to unmet patient needs was recognized as a key benefit of the CHW role, an issue that had previously been underappreciated by many clinical staff.

Careful recruitment and tailored training of CHWs contributed to successful CHW retention. Awardees that reported having high CHW retention attributed this success to a "rigorous approach to vetting and training" and recruitment of the "right" kind of individuals. One awardee described successful CHWs as "self-starters" and "gregarious." Another awardee attributed success with retention to a partnership between the awardee organization and clinical program that allowed clinical program leaders significant input into the training to create a relevant curriculum and support. Other awardees reporting training of CHWs to be a critical aspect of the innovation, and essential for effectiveness and retention.

2.2 Implementation Effectiveness Analysis

Five items from the AASF were used to develop a composite measure of implementation effectiveness for each innovation. These five items had factor loads greater than 0.70 on the implementation effectiveness construct and included primarily the assessment of adoption and fidelity. FLEs responded to each item using a four-point Likert scale and these responses were rescaled between 0 and 100, with 0 representing "not at all implemented" and 100 representing "complete and full implementation." The distribution of this composite implementation effectiveness score is presented in ***Figure 2-10***. The mean composite score among the 117 innovations for which FLEs could make assessments was 79.3 (SD 18.3).

Figure 2-10
Distribution of Implementation Effectiveness Composite Scores



2.2.1 Implementation Effectiveness Assessment

To further our understanding of HCIA interventions, we examined two different aspects of implementation effectiveness: (1) the relationship between rapid implementation of HCIA

interventions and awardee structural characteristics (section 2.2.2), and (2) the relationship between implementation effectiveness and program features (section 2.2.3).

2.2.2 Predicting Implementation Effectiveness: Path Modeling Approach

To assess the relations among implementation variables and factors and success in rapidly implementing HCIA interventions, we constructed a path model predicting implementation effectiveness. Path models are an extension of multiple regression and a special case of structural equation modeling that estimates the magnitude and significance of hypothesized causal connections between sets of variables. For this model, we considered all data collected from FLEs or abstracted from annual and quarterly reports. For AASF items, FLEs had to provide non-missing responses to half or more of the items in a scale to be included in the analysis. A second criterion for inclusion in the model was variation in respondent ratings (i.e., interventions were rated at different levels on a measure or scale). Finally, we considered the measure's theoretical contribution to predicting implementation effectiveness in this particular application. More information on these procedures is available in Appendix C.

In contrast to many tests of implementation, where a single intervention is being tested, the HCIA awardees represent a range of innovations and health care approaches being tested in a variety of sites. Moreover, some innovations are completely new while others represent extensions in scope, population, and/or setting. This heterogeneity in content, purpose, and innovation maturity creates challenges to fitting standard models of implementation effectiveness as the strength of relationship among factors and their association with successful implementation may vary based on any or all of these considerations. Therefore, while the variables available for testing were all drawn from Damschroder et al.'s (2009) theory of implementation and the thematic analysis conducted for the first annual report, the measured

How to rapidly and effectively implement innovations remains a significant challenge for delivering health care and other effective services. Using path modeling, we examined relationships among awardee structural characteristics and ratings reported by FLEs on the 2015 Annual Awardee Summary Form with FLE-rated implementation effectiveness.

- More staff training, greater implementation planning, and single-site implementation were directly related to improved implementation effectiveness.
- Innovations implementing HIT were significantly challenged in filling frontline staff roles and recruiting and retaining staff.
- Innovations whose implementation was formally planned engaged in more training and more formal staff training than other innovations.
- Staff who were hired to fill technical, research, or administrative roles were more likely to work semi-independently than clinical staff.
- Although not attaining significance, relative to awardees extending their programs, new programs
  - were somewhat less effective in implementing their awards
  - faced greater challenges in implementing HIT
  - were somewhat more likely to hire technical, research, or administrative staff to support their innovation.

relations and role of factors in predicting implementation effectiveness in prior research may not generalize well to the measured experience of first round HCIA awardees.

Initially, seven factors and four measures were identified as likely being related to success in effectively implementing interventions (see Appendix C for the items associated with each factor). However, due to missing data, one factor (Complex-Partnerships) and one measure, (Reimbursement) were dropped from the model. AASF factors and measures were also rescaled so that strongly endorsed items (initially coded 4) were scored as 100, while "not at all" was rescaled to zero. For example, responses for awardees implementing innovations with many complex duties or many HIT challenges were rescaled to 100 while scores for awardees implementing simple interventions or had no HIT challenges were rescaled to zero. The three binary measures (multisite, new program, and semi-independent staff) are scaled so that "No" equals zero and "Yes, the awardee has this characteristic" equals one. Descriptive statistics for the measures and factors included in the model are given in ***Table 2-2*** and descriptions of the retained measures and factors are provided below the table.

Table 2-2
Descriptive Statistics for Factors Included in the Implementation Effectiveness Path Model

Variable	Number of Respondents	Mean	Std. Dev.	Minimum	Maximum
High Implementation Effectiveness	117	79.26	18.28	25.00	100.0
High Clinician Buy-In	112	44.98	33.05	0.00	100.0
Many Complex Duties	122	58.89	21.81	0.00	100.0
Many HIT Challenges	121	30.37	25.29	0.00	100.0
High Implementation Planning	121	69.72	22.45	8.33	100.0
Multisite	123	0.67	0.47	0.00	1.0
New Program	123	0.41	0.49	0.00	1.0
New Hires	116	50.00	29.41	0.00	100.0
Recruitment Challenges	120	37.01	29.68	0.00	100.0
Semi-Independent Staff	123	0.33	0.47	0.00	1.0
High Staff Training	116	73.01	23.60	0.00	100.0

1. **High Implementation Effectiveness** is the dependent variable in this analysis and is an average of five of the six items from the second AASF assessing how rapidly and completely awardees fully implemented their intervention (full adoption by external partners was dropped due to the number of "not applicable" responses; $\alpha = .82$).

2. **High Clinician Buy-In** is a single item, taken from the second AASF, assessing the extent to which obtaining clinician buy-in and engagement was a challenge. Low ratings (0) on this item suggest it was not at all a challenge, while FLEs who felt it was a major challenge for their awardee rated it 100. We would expect greater difficulty in clinician buy-in and engagement to be negatively related to implementation effectiveness.

3. **Many Complex Duties** is based on a six-item factor that emerged from the second AASF, and describes the coordination and change in workflow (both in terms of staff and technology) required to implement the intervention ($\alpha = .69$). We expect that more coordination and need for staff adaptation to be negatively associated with implementation effectiveness.

4. **Strong HIT Challenges** is based on a seven-item factor that emerged from the second AASF. All items in this domain contributed to the factor, which assessed the extent awardees were challenged in implementing HIT ($\alpha = .91$). It can be observed that while many grantees found implementing HIT to be challenging, for the vast majority of grantees, it was not considered to be a great challenge. We expect greater challenges in implementing HIT to be negatively associated with implementation effectiveness.

5. **High Implementation Planning** is based on the four items from the second AASF assessing awardee's prior experience with similar innovations and the extent to which innovation implementation was coordinated and managed using planning documents ($\alpha = .71$). Greater implementation planning is expected to moderate the impact of complex duties on implementation effectiveness and be positively associated with implementation effectiveness.

6. **Multisite** is a single binary item abstracted from awardee reports assessing whether the intervention was implemented at one or more sites. Awardees implementing the intervention in a single site were scored zero while those implementing their intervention in multiple sites were scored one. Multisite should be considered a proxy variable, capturing the myriad of additional coordination, communication, and implementation complexities associated with implementing innovations in multiple distinct geographic settings. As such, multisite is expected to be negatively related to implementation effectiveness as multisite awardees are expected to have greater difficulty in fully implementing their awards.

7. **New Program** is also a single binary item abstracted from awardee reports (0 = extends current innovation; 1 = new program). As mentioned earlier, several awardees used their awards to extend current activities. This experience is expected to facilitate implementation; thus, a new program is expected to be negatively associated with implementation effectiveness.

8. **New Hires** is a single item from the second AASF that asked FLEs to rate the extent to which the intervention required hiring new technical, research, or administrative staff in support of the innovation. Note that this item does not include new staff hired to deliver clinical health care services, but new staff necessary to support the innovation. While new non-clinical staff are not expected to directly influence implementation effectiveness, based on the thematic findings from the first annual report they are expected to mediate the challenges associated with implementing HIT and be associated with the hiring of semi-independent staff.

9. **Recruitment Challenges** is a single item from the second AASF that assesses the challenges experienced by awardees in recruiting staff, staff turnover, and unfilled frontline staff roles. Greater difficulty in recruiting staff is expected to be negatively associated with implementation effectiveness as unfilled staff roles place greater strain on existing staff and impedes awardee capacity to fully implement their innovation.

10. **Semi-Independent Staff** is a variable from the first AASF in which FLEs were presented with several options for characterizing staff deployment models. This variable is an indicator for awardees that "use new staff semi-independently," that is, to a large extent, using staff with technical expertise to provide non-clinical services in support of innovation implementation.

11. **High Staff Training** is based on the four-item factor which emerged from the five items assessing staff training in the second AASF (the item "New (and rotating) staff receive training to accommodate staff turnover" was dropped due to the number of "not applicable" and "unable to assess" responses; $\alpha = .79$). Greater attention to staff training is expected to moderate the impact of complex duties on implementation effectiveness and is expected to be positively associated with implementation effectiveness.

The bivariate correlations between implementation effectiveness and variables hypothesized to be related to that effectiveness are included in the model are presented in ***Table 2-3***. Correlations range from -1 to 1 with zero indicating no relationship and -1 and 1 indicating, respectively, perfect negative and positive relationships. Put simply, the closer the correlation is to -1 or 1, the stronger the evidence the two items are related. The direction of the relationship (positive or negative) is derived from how the measures are scored (e.g., "high" staff training is expected to be positively correlated while "low" staff training would likely be negatively correlated with implementation effectiveness. Correlations close to zero indicate that movement on one variable has no relation to movement on another variable. Many of the relations are modest, with the strongest positive relations being found for staff training and implementation planning, HIT challenges with complex duties, implementation effectiveness with implementation planning and staff training, and the need for new hires with semi-independent staff. Modest correlations can be observed for implementation planning with both new and multisite programs, complex duties with implementation planning, and semi-independent staff with new program.

Table 2-3
Bivariate Correlations Among Factors included in the Implementation Effectiveness Path Model[a]

	New Program	Multisite	Implementation Planning	Complex Duties	Staff Training	HIT Challenges	Recruitment	Implementation Effectiveness	New Hires	Clinician Buy-in	Semi-independent Staff
New Program	1										
Multisite	-0.17	1									
Implementation Planning	-0.19	0.24	1								
Complex Duties	-0.14	-0.02	0.21	1							
Staff Training	-0.08	0.14	0.42	0.10	1						
HIT Challenges	-0.17	-0.11	0.03	0.31	-0.11	1					
Recruitment	-0.06	0.12	-0.06	-0.15	-0.06	0.14	1				
Implementation Effectiveness	-0.13	-0.13	0.40	0.10	0.38	-0.03	-0.15	1			
New Hires	0.09	0.04	0.04	0.08	0.15	0.14	0.15	0.01	1		
Clinician Buy-in	-0.01	-0.05	-0.09	0.06	-0.09	0.05	0.05	-0.06	0.07	1	
Semi-Independent Staff	0.22	0.10	0.12	0.03	-0.12	0.05	0.10	-0.05	0.38	0.22	1

[a] Data obtained through structured coding of 1st and 2nd FLE annual reports and administration of the 1st and 2nd Annual Awardee Summary Forms.

2.2.2.1 Path Model Analysis Methods

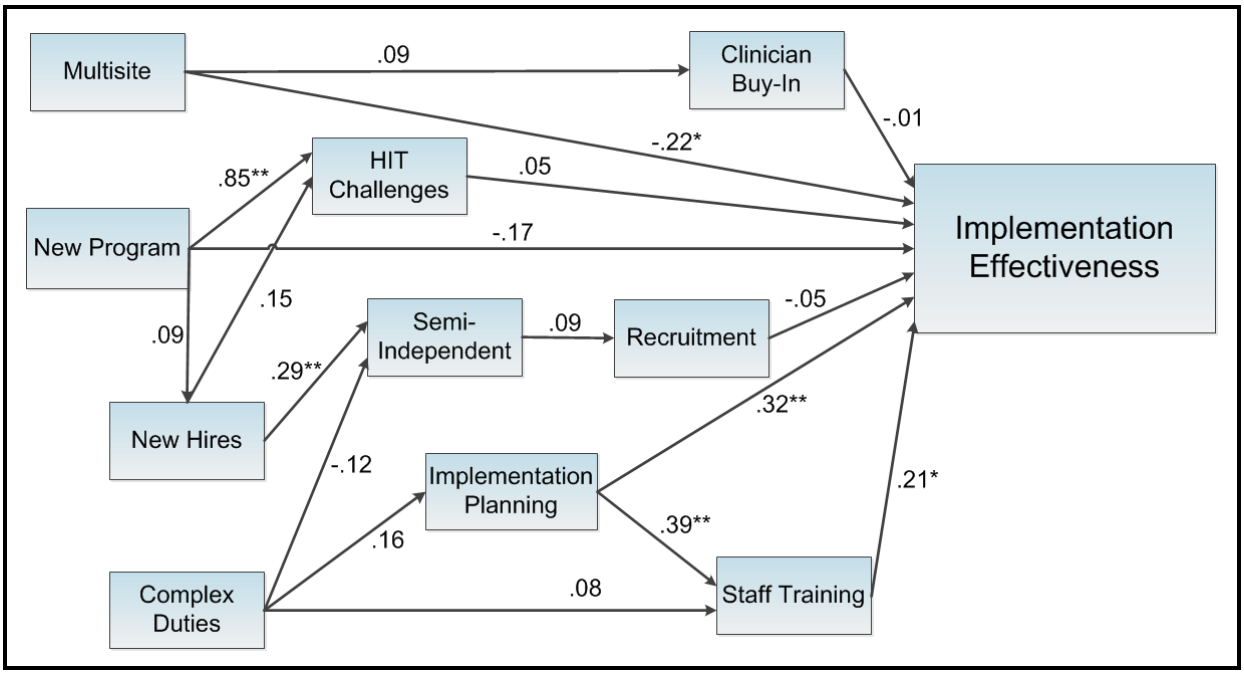
We analyzed the relationships among the AASF variables using path analysis. Path analysis is a statistical technique for estimating associations among a set of variables arranged in a presumed, hierarchical causal sequence; that is, variables on the right side of the model are assumed to be a result of variables entered on the left side of the model. Variables in the middle are expected to influence the relationship between the variables on either side of the model. The results of multiple regression equations are displayed in the form of a model that summarizes the key relationships (or paths) in the data. The magnitudes of individual effects are measured by standardized regression (beta) coefficients. After listwise deletion for missing data, of the 123 interventions for which we had second annual AASF data, 98 interventions (80.3 percent) from 87 awardees contributed to the model.

2.2.2.2 Path Model Results

The final path model is shown in ***Figure 2-11***. The exogenous covariates are aligned on the left side of the model, mediators ordered sequentially in the middle, and the degree of implementation outcome is to the far right. Lines (paths) between two variables indicate a hypothesized direct relationship between those two measures, while the numbers above, and

occasionally below the lines indicate the strength of relationship (with 0 indicating no relationship and -1 or 1 indicating a perfect relationship). The figure displays the measures and factors affecting implementation effectiveness. The strength of the path from multisite to clinician buy-in is 0.09, from multisite to implementation effectiveness is -0.22, from new program to HIT challenges is 0.85 (statistically significant), from new program to implementation effectiveness is -0.17, from new program to new hires is 0.09, from new hires to semi-independent staff is 0.29 (statistically significant), from complex duties to semi-independent staff is -0.12, from complex duties to implementation planning is 0.16, from complex duties to staff training is 0.08, from HIT challenges to implementation effectiveness is 0.05, from semi-independent staff to recruitment is 0.09, from implementation planning to implementation effectiveness is 0.32 (statistically significant), from implementation planning to staff training is 0.39 (statistically significant), from clinician buy-in to implementation effectiveness is -0.01, from recruitment to implementation effectiveness is -0.05, and from staff training to implementation effectiveness is 0.21 (statistically significant).

Figure 2-11
Measures and factors affecting implementation effectiveness



*p<.05, **p<.01

Overall, the data collected from FLEs and abstracted from annual reports fit the hypothesized model of implementation effectiveness relatively well ($\chi^2 = 52.41$, $p = 0.02$). The root mean squared error of approximation (RMSEA) is equal to 0.074 (90% CI = 0.028–0.11), which, while considered a mediocre fit by MacCallum, Browne, and Sugawara (1996), is well below the 0.10 cutoff suggested by others as indicative of a good fit (Kenny, 2015).

It was hypothesized that multisite awardees would have more difficulty than others in achieving clinician buy-in, and that this would slow implementation. However, this does not

appear to be the case. Multisite awardees show little difference relative to others in achieving clinician buy-in, indicating that buy-in was not an issue limited to multisite awards and clinician buy-in was not associated, in this path, with implementation effectiveness. However, in contrast with findings obtained from the first annual awardee summary, by the end of the second year of their award multisite awardees were reported as having significantly lower levels of implementation effectiveness relative to other awardees ($p = 0.01$). The difference between the first and second year implementation effectiveness findings is likely due to the greater opportunity for innovations to mature and awardees to demonstrate (or not) implementation effectiveness. However, why multisite interventions are less effectively implemented remains an unanswered question. It seems clear that multisite awardees face additional challenges relative to single-site awardees, but how those challenges are uniquely different from single-site implementations remains unanswered by this analysis. As the AASF was designed to collect elements common across multi- and single-site innovations, we unfortunately lack unique evidence on conditions and challenges specific to multisite awardees.

We also hypothesized that the many challenges facing those implementing new programs would impact implementation effectiveness. Perhaps surprisingly, in this model new programs were no more or less likely to demonstrate implementation effectiveness ($p = 0.35$), face challenges implementing HIT ($p = 0.56$), or have to hire new technical, research or administrative staff ($p = 0.38$), than awardees extending or expanding their innovations.

Although having challenges with HIT was not associated with implementation effectiveness, it was associated with a somewhat greater need to hire new technical, research, or administrative staff ($p = 0.12$) who were often semi-independent ($p = 0.00$), and this recruitment was often challenging ($p = 0.01$). It was not, however, more challenging to recruit these staff than other clinical staff ($p = 0.36$), and recruitment, in this model, was not significantly associated with implementation effectiveness ($p = 0.63$).

We had expected that the amount of coordination necessary and workflow changes required to implement the intervention ("complex duties") would involve additional training, more careful planning, and might entail hiring additional staff to support the innovation. None of these hypotheses were supported by the data. Interventions scoring higher on complex duties were no more likely than those scoring lower to engage in more staff training ($p = 0.39$); use more formal, coordinated, or written planning documents ($p = 0.10$); or hire semi-independent staff ($p = 0.19$).

Both staff training and implementation planning were strongly and positively associated with implementation effectiveness ($p = 0.03$ and 0.00, respectively). Awardees who engaged in more formal and extensive staff training and implementation planning were reported to have higher levels of implementation relative to awardees who reportedly had lower levels of staff training and implementation planning.

2.2.2.3 Path Modeling Summary

We used path analysis to examine the relationships among awardee structural characteristics and ratings reported by FLEs on the 2015 AASF. Below we review the key findings from this approach to understanding implementation effectiveness.

Path modeling identified three factors that impacted implementation effectiveness directly and significantly. These were staff training, implementation planning, and single-site implementation. Implementation effectiveness was rated lower among interventions implemented in multiple sites although this relationship may be attributable to numerous complications—not measured in the annual survey—associated with communicating goals and coordinating activities across multiple distinct social and geographic settings.

Significant relations also emerged among three of the assumed correlates of implementation. Innovations implementing HIT were significantly challenged in filling frontline staff roles and recruiting and retaining staff. Staff who were hired to fill technical, research, or administrative roles tended to work semi-independently. And innovations for which implementation was formally planned engaged in more training and more formal staff training.

An additional three relations did not attain significance, but may merit consideration for future implementation planning. Awardees who were extending their innovations showed somewhat greater implementation effectiveness than those implementing new programs. New programs tended to face greater challenges in implementing HIT and somewhat more likely to hire technical, research, or administrative staff than awardees who were extending their innovations to other settings or populations.

Contrasting the limited findings from this analysis with the rich results from the thematic analysis presented in section 2.1 may seem at first confusing. The thematic analysis highlights many implementation features and challenges that are not present or are not supported in the path model presented above. One possible explanation is that many of the issues emerging from the thematic analysis are not expressed in the context of implementation effectiveness, but as issues that awardees encountered; how and whether they impacted implementation is generally not assessed by the thematic analysis. It is also likely that many of the conditions and issues raised in the thematic analysis affect subsets of awardees, and may affect them quite substantially, but any measured relationship would be diminished in a statistical analysis if the feature tested in the model influences implementation effectiveness differently depending on the subgroup in which it occurs. Finally, it is also likely in this sample that implementation effectiveness is impacted by many awardee features and challenges, only a few for which we had data and which are tested in this model. In statistical parlance, these features and challenges represent confounds—unmeasured conditions that influence the implementation effectiveness finding, but are not accounted in the model. To the extent any of these suppositions occur, the strength of relations will be attenuated in the tested model.

How to rapidly and effectively implement innovations remains a significant challenge for delivering health care and other effective services. Multiple models of implementation have been proposed (see Nilsen, 2015 for a review) and a variety of factors have been identified as associated with effective implementation, but relatively few models have been rigorously tested across multiple diverse settings and interventions. Much of the current understanding of implementation effectiveness comes from research that has tested specific interventions or strategies at similar levels of maturity, implemented across similar settings, and addressing similar outcomes.

The current model reveals several factors associated with implementation effectiveness for this unique set of awardees, but other factors we thought might be critical did not consistently influence intervention effectiveness. While recruitment and HIT challenges were often prominent themes in FLE reports, these challenges did not adversely affect implementation effectiveness. And while innovation complexity was thought to be related to the need for more careful planning and additional staff training, these relations were not observed. The domains with the largest direct effects on effectiveness were staff training and greater planning efforts. By the end of their second year, most awardees had implemented their innovations to a great or moderate extent. This path model identified several features that tended to hinder or facilitate implementation. Many of the features identified may be modified through careful planning and thoughtful anticipation of the many needs and challenges associated with effectively implementing innovations.

2.2.3 Predicting Implementation Effectiveness: Qualitative Comparative Analysis (QCA) Approach

We used data from the structured qualitative coding of the FLEs' first and second annual reports and data from the AASF2. The annual reports and the AASF2 contained information on innovation components and characteristics (e.g., care coordination, health IT, complexity of the innovation), characteristics of the target population (e.g., pediatric focus, socially fragile), and context of the implementation (e.g., multisite implementation). We used our knowledge of the awardee innovations and principles from implementation science to select 22 features that might influence implementation effectiveness (see Appendix H). We used the FLE assessment of reach, dose, and fidelity from the AASF to determine the degree of implementation effectiveness for each awardee.

2.2.3.1 QCA Analysis Methods

We specified 120 different QCA models, each with three or four features as described in the above section and in Appendix H. We conducted analyses that included all HCIA awardees that had an implementation effectiveness score (N = 106). We also conducted analyses of all models on 12 smaller groups of awardees defined by similar characteristics (e.g., all awardees that used CHWs, all awardees that implemented new programs) to reduce heterogeneity across the included cases. We conducted analyses using both crisp and fuzzy sets, which refer to the way in which features of interest and the outcome are specified for analysis.

2.2.3.2 QCA Results

No single feature was necessary or sufficient among awardees with effective implementation in any of the models. Our analyses also yielded no combinations of features that were sufficient for effective implementation. In other words, we could not identify any patterns of features that were consistently found among awardees with effective implementation. What we found was that patterns in combinations of features were as likely to be present in awardees with ineffective implementation as they were to be found among awardees with effective implementation. From a set-theoretic perspective, this is a "null" finding.

We offer several explanations for these findings. First, the data we used for these analyses were designed to solicit common data elements from a wide variety of interventions. It

is possible that elements in the second AASF were too broad for the purpose of classifying awardees into unique and discrete sets. It may be that more specific measures targeting specific innovation features and components would improve the capacity of QCA to discover common elements of success; however, that would reduce the number of awardees that could be included in any single analysis to those with that specific feature or component. Second, the FLE assessment of implementation effectiveness resulted in data for this outcome that was quite skewed with FLEs generally rating implementation effectiveness highly for their innovations. Consequently, little to no variation in implementation effectiveness could be explained by innovation features. Finally, it is possible that no complex relationships between innovation features and implementation effectiveness actually exist and the "null" QCA findings accurately reflect the nature of the relationship between innovation features and implementation effectiveness; however, we think this explanation is the least likely.

2.2.4 Next Steps

We plan to evaluate the relationship between innovation features and impact outcomes using QCA in the third annual report. The impact outcomes include total Medicare expenditures, hospitalizations, readmissions, and emergency department visits. We will reassess which innovation features to specify in the analysis, based on those we would expect to influence the impact outcomes the most.

2.3 Implementation Findings Summary

The second year saw the vast majority of innovations implemented to either a great or moderate extent. Nonetheless, many awardees continued to struggle with both anticipated and unanticipated challenges including reimbursement for non-traditional staff and services, managing cultural barriers, vulnerable patients' needs for additional resources and support, the need for additional staff to support innovation implementation, and the time necessary to forge strong relations with new partners. Innovation adaptation was a common response to these and other challenges, with several benefits emerging as innovations matured. In particular, despite ongoing, often small to modest challenges from implementing HIT, the benefits of robust HIT infrastructures began to be apparent. In addition, many improvements were observed in staff satisfaction, retention, empowerment, and relations through awardee cross-training; physical colocation; and improved recruitment, hiring, and training practices. Working alongside of CHWs increased staff appreciation of CHWs for their contributions in improving workflow, connecting with patients, and enhancing implementation. And finally, as the value of innovations in improving workflow and patient care became apparent to clinical staff, their satisfaction and support for the innovations increased.

However, several challenges, often beyond awardee control, continue to affect awardee performance and sustainability. Perhaps due to the learning-by-doing nature of many of these innovations, the limited prior experience in implementing innovations of a similar size or scope by several awardees, and the need to adapt innovations based on implementation experience, few awardees found formal improvement or change management processes useful for monitoring innovation implementation. Awardees reported that existing organizational capacity affects resilience to challenges and may affect sustainability, as do state decisions around Medicaid expansion and their support (or lack thereof) for alternative payment models.

As awardees successfully implemented their innovations and observed the benefits of these practices, innovation and organizational leaders' attention turned increasingly to sustaining all or part of their innovations once CMS-support ends. Awardees in some innovations turned to state and federal funding streams for ongoing support; some secured financing from commercial health plans; and awardees in large provider institutions expect their workflow-integrated innovations to continue once HCIA funding ceases, although continued support from large providers is often conditional on demonstrated return on investment or documented improvements in patient health outcomes. As lack of reimbursement for care coordination and new staff types is a significant barrier to sustainability, awardees adopting these innovations see participating in future payment reform pilots or demonstrations—mainly ACO pilots—as a viable mechanism for sustainability. Finally, partners played an active and strategic role in planning for sustainability by agreeing to adopt and integrate key innovation components into their existing work.

Rapidly and effectively implementing innovations remains a significant challenge for delivering health care and other effective services. We were unsuccessful in identifying any necessary or sufficient programmatic features or combinations of features using QCA. All features tested were present in both effective and ineffectively implemented interventions. Using path modeling, we examined relationships among awardee structural characteristics and ratings reported by FLEs on the 2015 AASF with FLE-rated implementation effectiveness. Implementation effectiveness was significantly better among awardees implementing their innovation in a single-site and who engaged in more staff training and implementation planning. Filling frontline staff roles and recruiting and retaining staff was a significant challenge for innovations implementing HIT, and staff who were hired to fill technical, research, or administrative roles were significantly more likely to work semi-independently than clinical staff. Although not attaining significance, relative to awardees extending their programs, new programs were somewhat less effective in implementing their awards, faced greater challenges in implementing HIT, and were somewhat more likely to hire technical, research, or administrative staff to support their innovation.

[This page intentionally left blank]

SECTION 3:
IMPACT FINDINGS

In this section, we report the results of applying multiple quantitative approaches to evaluate the HCIA data we have assembled. Compared to the end of last year, there was little change in the findings summarizing HCIA effects. These results, summarized in forest plots, continue to show a wide range of favorable, unfavorable, and mostly null effects, with mean impacts for all three core outcomes near zero. Despite increased sample sizes and additional follow-up quarters, wide confidence intervals point to considerable persisting imprecision in the results for many innovations. We showed once again that the highest and lowest effect sizes emanate from innovations with small sample sizes. An alternative method of estimating effects using quarterly time series found HCIA effects that were very similar to the regression-based difference-in-difference estimates reported by the FLEs. A Bayesian fixed-effects model led to the same conclusions.

We conducted a detailed review of the approaches that FLEs used to create comparison groups, delineating five different methods. There were no systematic differences between propensity score weighting and matching procedures with respect to HCIA effects on total cost of care. We also created a synthetic comparison group to be used in cases in which no comparison was available. The synthetic comparison produced evaluation results that were substantially different from the pre-post results for four awardees reported by the FLEs, differing by at least $500 per awardee. This substantiates our decision not to use pre-post effect estimates in our analyses.

Our analyses of the effect variation between innovations produced strong evidence of heterogeneity for Total Cost of Care, Inpatient Admissions, and in two of the three settings for ED Use. This implies that much of this variation observed may be attributable to characteristics of the innovations themselves, which we examined using three separate meta-regression models for Total Cost of Care. Four innovation features were found to be significantly associated with costs savings: for-profit tax status, a health informatics component, community health worker staff, and clinically fragile patients. The meta-path model demonstrated that hospital admission effects had a greater impact on cost savings than ED use effects did. The key mediator in the path model was New Innovation status. Compared to existing innovations, new innovations had lower levels of implementation effectiveness and produced less favorable hospital admission and ED effects.

3.1 Evaluability Assessment

We continue to assess evaluability for each HCIA intervention. This entails monitoring the availability and quality of difference-in-difference (DiD) estimates for the four core outcome measures. From our assessment, 86 of the 135 interventions (63.7 percent) implemented by 68 of the 108 awardees (63.0 percent) had a DiD estimate for at least one core measure (i.e., total cost of care, hospitalizations, hospital readmissions, and emergency department visits) that could be used in our meta-analysis. This is an increase from our last report in which we had at least one estimate from 59 of 122 interventions (48.4 percent of the interventions; 49 of the 108 awardees or 45.4 percent of awardees). In this section we detail how we defined and classified HCIA interventions for meta-analysis, the types of estimates we received from the FLEs, and how we

determined which estimates were appropriate for meta-analysis. It should be noted that these results are based on findings achieved during the second year of the award and the HCIA programs are ongoing. These findings must be considered preliminary results, which we expect may change during the third year of awardee funding. We will continue to collect and solicit evidence of effectiveness from FLEs and hope, in the third year, to increase the number of interventions for which we have data.

3.1.1 Classification of HCIA Interventions

In our first annual report, we conducted our analyses at the awardee-level. Since then FLEs have begun to report findings with greater specificity, reporting awardee findings at the intervention level. That is, in situations where an awardee has implemented two or more interventions, separate impact estimates are reported. Consequently, we are also reporting at the intervention level and consider those interventions as independent trials even though they are funded through a common mechanism. Our rationale is that they serve different samples, have implemented different innovation components, and are typically administered by separate staff. Additionally, if an FLE reports separate estimates by payer (e.g., Medicare and Medicaid), then we treat those estimates individually. Of the 108 awardees, 22 (20.4 percent) implemented and tested multiple interventions. This increased the number of possible innovations tested from the 108 original awards to 135 interventions potentially available for testing. Of those 135 interventions we have received usable DiD estimates for at least 1 outcome from 68 awardees.

To improve our capacity to identify awardee-related sources of heterogeneity, we partitioned awardees into the three broad classes: ambulatory care, post-acute care (including post-acute, ED, and long-term care settings), and hospital-setting. This partitioning is being done for both substantive and methodological reasons. Substantively, these represent distinct intervention approaches for different populations with different health care needs. Methodologically, these substantive differences produce highly disparate results on the four core outcome measures. For example, the potential scale of cost savings from hospital-based innovations is much greater (given the higher cost of care) than those obtainable from ambulatory care innovations. Even within these large groupings, the actual interventions and populations are diverse and may create variation in outcomes independent of actual effectiveness. In section 3.4, we use meta-regression to examine how intervention and population diversity impacts estimates of ambulatory care effectiveness

Ambulatory care innovations generally identify and enroll eligible patients on a rolling basis and then follow them for the reminder of the innovation period. They provide ongoing preventive, primary care, and specialist services in health care facilities or in the patient's home. Post-acute care innovations are typically transition programs that target patients recently released from hospitals, nursing homes, or skilled nursing facilities. Notably, post-acute care data are generally episode-based with follow-up periods ranging from 30 to 120 days. Because of the episode-based nature of these interventions, we also included in the post-acute group the small number of episode-based interventions that take place in the ED setting or long-term care setting. The hospital-setting group involves a unique set of innovations based in hospitals, long-term care facilities, and intensive care units that are unlike any of the facilities in the other categories. The interventions in the hospital-setting group come exclusively from the hospital-setting HCIA awards. All hospital-setting HCIA awardee interventions are in the hospital-setting group except

UChicago. UChicago was identified by the FLE as having rolling entry rather than episode-based entry into the program and providing services from enrollment onward. Because of this substantive difference, we classified UChicago with the ambulatory-setting interventions. We based our post-acute care and ambulatory care designations on the classifications reported by FLEs. When these classifications were not provided, we based the assignment on FLEs' descriptions of the innovation components and on our qualitative coding of individual programs.

3.1.2 Types of Estimates Received

We are using summative DiD estimates of the core measures for meta-analysis. DiD estimates the average difference in performance between innovation and comparison groups over time, accounting for any pre-intervention period trends. Consequently, the availability of summative DiDs is essential, and we have monitored their availability from the FLEs carefully. ***Table 3-1*** summarizes the estimates that were available for this report across the ambulatory, post-acute, and hospital settings.

Table 3-1
Types of Estimates Received

Estimate Type	Ambulatory Care	Post-Acute Care	Hospital-Setting	All Innovations
DiD estimates for core measures*	39 (40.2%)	18 (62.1%)	8 (88.9%)	65 (48.1%)
4 core measures	23 (23.7%)	11 (37.9%)	***	34 (25.2%)
3 core measures	12 (12.3%)	7 (24.1%)	8 (88.9%)	27 (20%)
2 core measures	4 (4.1%)	0	0	4 (3%)
1 core measure	0	0	0	0
Non-DiD estimates for some or all core measures**	32 (33%)	8 (27.6%)	0	40 (29.6%)
Quarterly DiD	18 (18.6%)	4 (13.8%)	0	22 (16.3%)
Other measure types**	14 (14.4%)	4 (13.8%)	1 (11.1%)	20 (14.8%)
No core estimates reported	26 (26.8%)	3 (10.3%)	0	29 (21.5%)
FLE plans to report on at least some core measures	5 (5.2%)	0	0	5 (3.7%)
Problems with data availability	9 (9.3%)	3 (10.3%)	0	12 (8.9%)
Comparison group not feasible/large barriers to comparison group formations	12 (12.4%)	0	0	12 (8.9%)
Totals	97	29	9	135

* DiD estimates reported by the FLEs

** Pre-post, quarterly means, non-DiD summative estimates

*** Most hospital-setting innovations provide their intervention in the hospital; the FLE does not report hospital admissions

DiD estimates were available for 39 of 97 (40.2 percent) ambulatory care interventions. Of ambulatory interventions reporting at least one DiD estimate, most are reporting for all core four measures. Eighteen ambulatory interventions reported quarterly DiDs rather than summative DiDs. Other ambulatory care interventions reported pre-post estimates or regression adjusted estimates, and three only reported unadjusted quarterly means and standard deviations for the intervention and comparison groups.

Most of the post-acute interventions had at least one DiD estimate (18 of 29 interventions or 62.1 percent) and all of those reported at least three of the core four measures. For four interventions, quarterly DiDs were reported; four others reported pre-post estimates, regression adjusted estimates, or quarterly means for at least one of the core measures.

The hospital-setting interventions occur in the hospital or to recent inpatient discharges. Consequently, the hospital admission outcome is not appropriate for this group and was not reported by the FLE. All but one intervention in the hospital setting had DiD estimates for all of the other core measures. This intervention provided instead quarterly means for the intervention and comparison groups.

No estimates, DiD, quarterly, or otherwise, were available for 29 interventions. The most frequently cited reasons for a lack of estimates were data availability and problems constructing a valid comparison group. For five interventions, the corresponding FLEs are optimistic that they will be able to report on the core measures in the future.

3.1.3 Appropriateness of Estimates for Meta-Analysis

In addition to monitoring the availability and type of estimates being reported by the FLEs, we have monitored the quality of the estimates being reported and the ability of estimates to be included in the meta-analysis. Specifically, we have identified two situations in which we have decided that estimates should be used with great caution and one situation in which we were able to generate usable estimates from FLE-reported data.

Although all estimates contain uncertainty, some FLEs have noted estimates that should be interpreted with particular caution. We found five instances of this occurring. The reasons to treat these estimates with caution ranged from having a small sample size in conjunction with the inability to find comparators that match to high utilizers in the intervention group to serious concerns regarding the comparison group's ability to match on characteristics essential to the intervention selection process. For one intervention, the FLE reported DiD estimates but later found it to be erroneous; the FLE plans to correct their estimates in the future. Because exceptionally unreliable estimates and estimates known to be wrong can jeopardize the quality of our meta-analytic findings, these five estimates are not included in our analyses.

A handful of interventions that reported summative DiD estimates for at least some of the core measures serve populations that are unlike those in any of the other interventions. The interventions serve palliative care patients (PCCSB), chronically ill children (Houston), hospice patients who are mostly in their last 30-days of life (Sutter-AIM), or mothers with infants whose individual effects cannot be summed meaningfully (Finity). These populations have expenditures and utilization rates unlike those typically observed in HCIA interventions. We refer to these

interventions as serving "unique populations," and although we report their estimates in section 3.2, we do not use them in our grand means analyses nor meta-regressions.

The availability of methodologically consistent estimates is a key strength of our meta-evaluation and a major reason why we specified that FLEs report summative DiD estimates. For this report, two FLEs did not report summative DiD effect sizes for any of their interventions and one FLE did not report summative DiD effect sizes for five of their interventions. In these cases, however, the FLEs reported quarterly DiD effect sizes. We were able to use the quarterly DiDs to calculate summative estimates usable for meta-evaluation; the calculation was applied to 18 ambulatory and 4 post-acute interventions, increasing the number of interventions included in our analyses by 22. Our post-hoc calculation of summative DiDs from quarterly DiD estimates is not ideal as we cannot accurately model the correlation between quarterly DiD estimates resulting in standard errors that are likely smaller, but possibly larger than would be obtained from a summative DiD (see Appendix E for details). The implication is that, in the meta-analysis, these calculated estimates may appear to have more certainty than is warranted. We have followed up with these FLEs requesting they provide summative DiD effect sizes and will use those summative estimates in our next report.

After our assessment of the data available and the appropriateness of it for meta-analysis, we have at least one core measure DiD estimate for 86 of the 135 interventions (63.7 percent) that can be used in our analyses. This is an increase from our last report in which we had at least one estimate from 59 of 122 interventions (48.4 percent of the interventions; 49 of the 108 awardees or 45.4 percent of awardees). We will continue to assess the availability of data from the FLEs for our meta-evaluation. This includes the 60 interventions (49 awardees) that have received no-cost extensions of which we expect FLEs to report evaluation findings for 34 interventions (27 awardees) during the extension period. Finally, we will continue to monitor the reliability of the measures we receive and their implications for our meta-evaluation of the HCIA program.

3.2 Meta-Analysis

3.2.1 HCIA Innovation Impacts on the Four Core Outcomes

In this section, we present the impacts of individual HCIA innovations on the four core outcomes. As described in Appendix E, impact estimates were drawn from DiD regression analyses reported by the FLEs. The results are summarized in the form of forest plots. At the end of their second year of award, interventions are producing a range of effects on total cost of care, hospitalizations, hospital readmissions, and emergency department visits. At the 90 percent confidence interval, a few interventions show significantly reduced costs and/or utilization, a few show significantly increased costs and/or utilization, but most innovations show no difference in impact on the four core outcomes relative to their comparators. Across settings and outcomes, the average effect of interventions remains close to zero, indicating no appreciable difference in outcomes attributable to the HCIA award.

In this section there are separate sub-sections for each outcome, with the results broken out separately by ambulatory care, post-acute care (including post-acute, ED, and long-term care settings), and hospital-setting innovations. We also present a plot of interventions identified as

serving unique populations. Interventions were deemed to serve unique populations by review of awardee reports by subject matter experts. An overall mean and standard error averaged over all the interventions appears at the top of each plot. These values are calculated by weighting each contributing estimate by the inverse of its variance (fixed effects). All innovations are identified by a unique identification number. A complete list of these numbers is provided in Appendix G.

3.2.1.1 Total Cost of Care

The first of the core outcomes is Total Cost of Care (TCOC). These are the costs associated with Medicare Parts A and B, Medicaid expenditures, or encounters for Medicare Advantage beneficiaries. TCOC effects are regression-adjusted difference-in-difference estimates contrasting the innovation and its comparison group. All effects were converted into average differences per beneficiary per quarter. Negative effects represent cost savings, while positive effects are dissavings.

The TCOC forest plot for ambulatory care programs, the largest group of awardee innovations, is shown in ***Figure 3-1***. The effects of the 49 unique interventions range from $2,455 in savings (St Francis, 1056A) to $3,117 in dissavings (UChicago, 1033, not shown), and were fairly evenly distributed around the vertical line denoting an effect of $0 difference. The weighted overall effect was an additional $13 per beneficiary per quarter (SE = $6). The 90 percent confidence intervals indicate that eight of the innovations had savings that were significantly greater than zero (UEMS, 1026; FirstVitals, 1072; Bronx, 1055; Y-USA, 0965; IOBS, 0969; Le BonHeur, 1046; St. Francis, 1056A; Kitsap, 1062), while another eight innovations reported dissavings significantly greater than zero (Mineral, 1058; J-CHIP, 1053A; Curators, 1001; Altarum, 0976; PPMC-HRP, 0985B; Delta Dental, 0980; Carilion, 1010; Intermountain-C3, 0978C). It can also be observed that there is notably greater precision (smaller confidence intervals) for awardees reporting null findings.

Figure 3-1
Total Cost of Care: Ambulatory Care Innovations

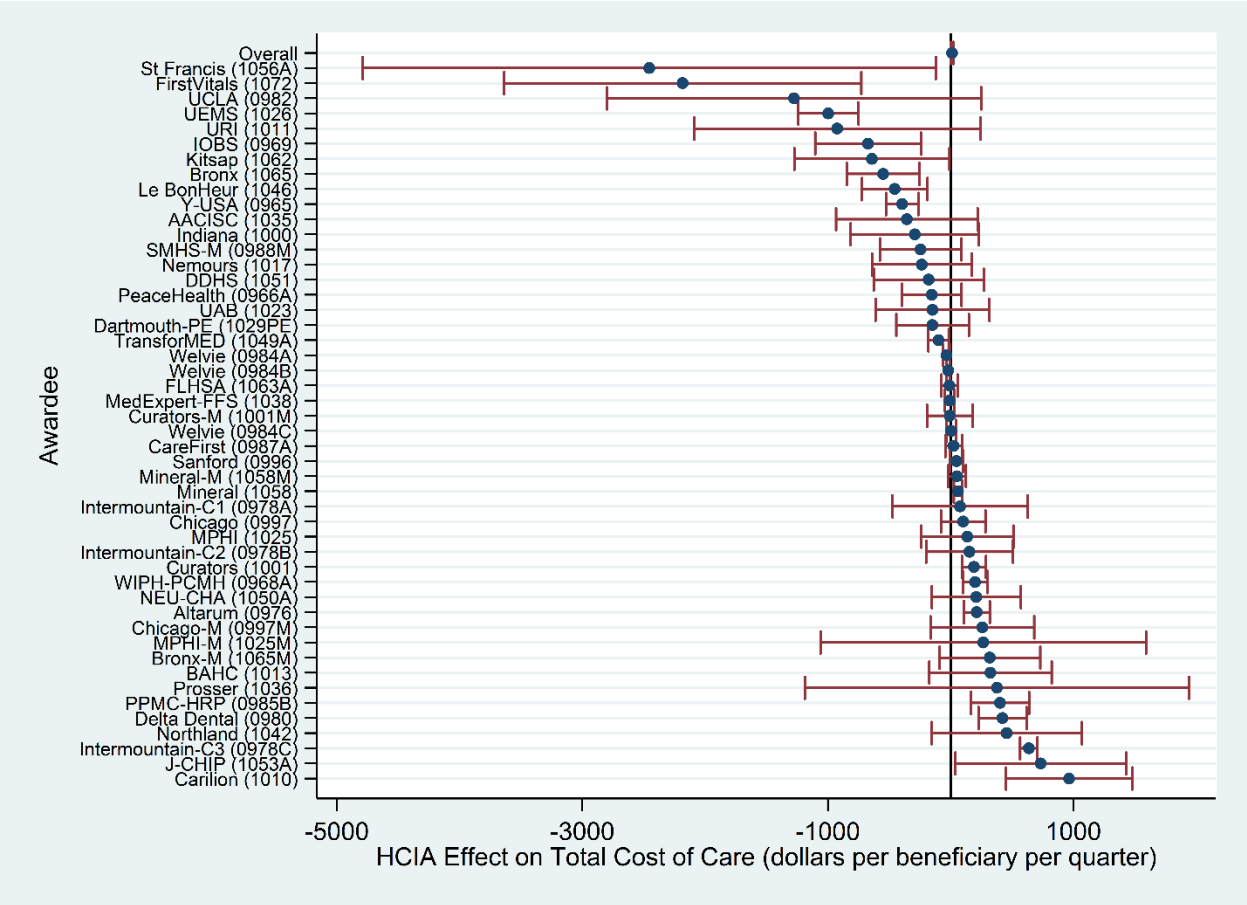
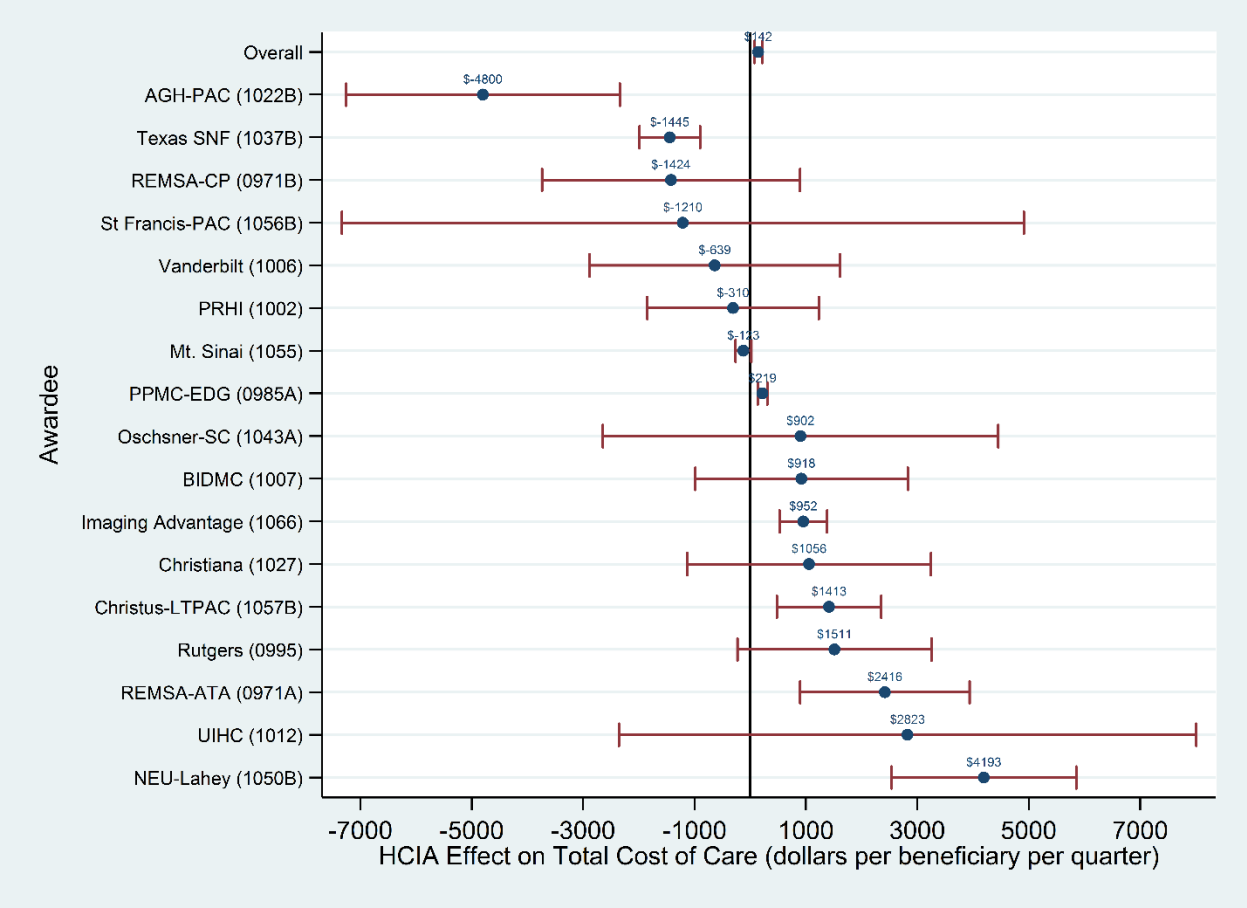


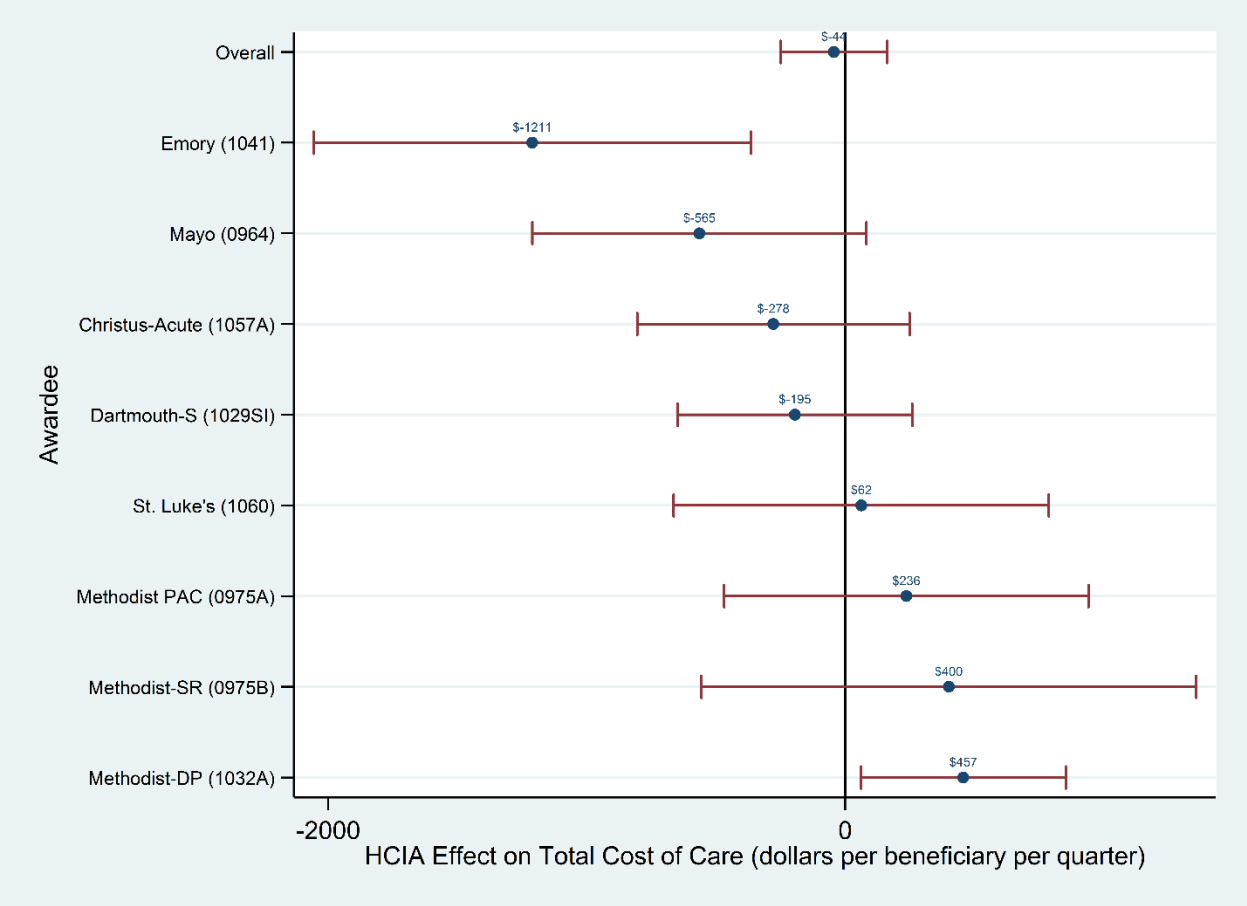
Figure 3-2 shows the results for the 11 post-acute setting interventions, 4 ED setting interventions, and 2 long-term care setting interventions. These innovations also exhibited a broad range of TCOC effects, and eight showed significance. Two showed significant savings (AGH-PAC, 1022B; Texas SNF, 1037B), and five reported significant dissavings (PPMC-EDG, 0985A; Imaging Advantage, 1066; Christus-LTPAC, 1057B; REMSA-ATA, 0971A; NEU-Lahey, 1050B). The mean overall effect for the post-acute care innovations was an additional $142 per beneficiary per quarter (SE=$43).

Figure 3-2
Total Cost of Care: Post-Acute Care Innovations



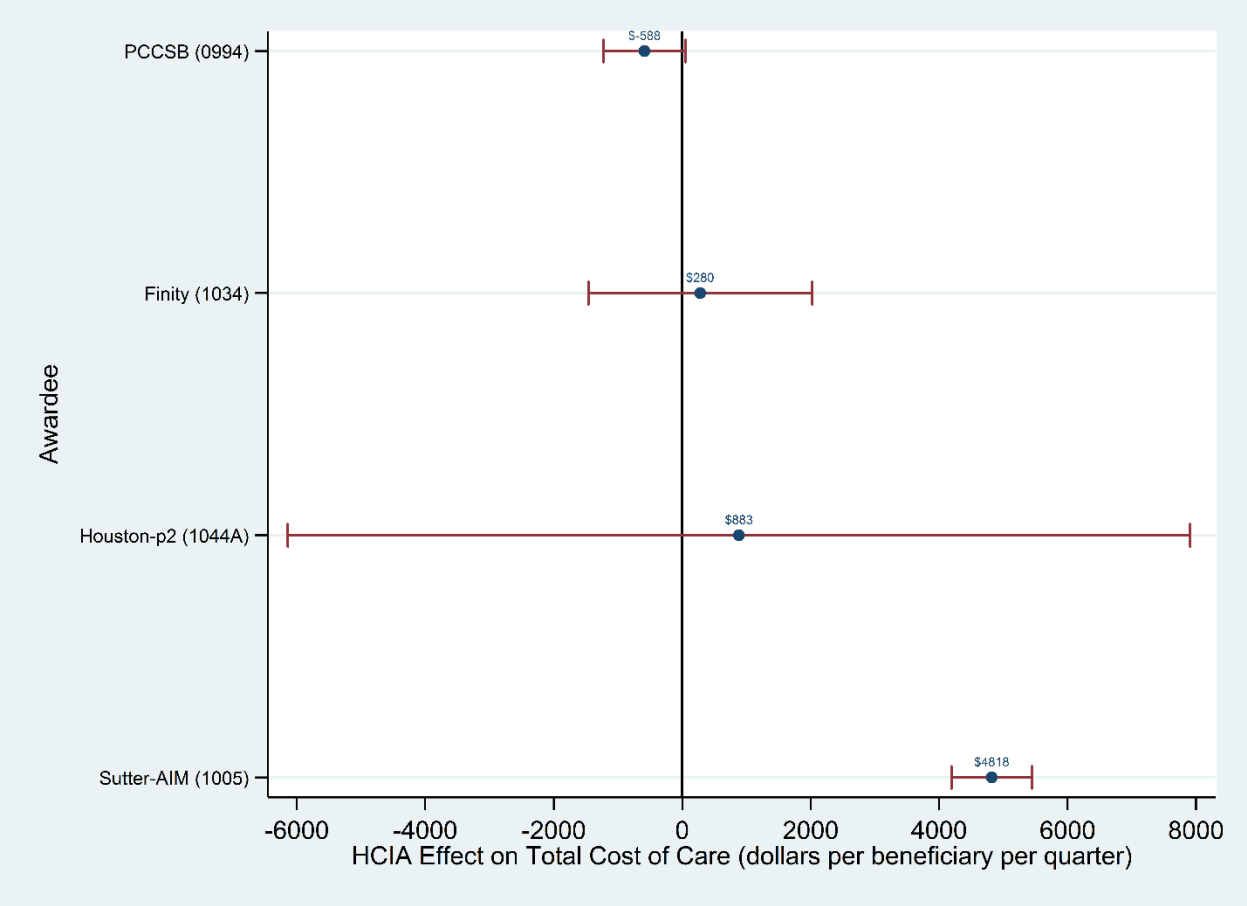
The plot of the eight hospital-setting interventions is shown in ***Figure 3-3***. These effects are expressed in terms of 60-day spending per episode. The grand mean effect of hospital-setting innovations on TCOC was not significantly different from zero (-$44 per beneficiary per quarter, SE = $125). Only one of the estimated effects shows dissaving significantly different from zero (Methodist-DP, 1032A), and one intervention showed significant savings (Emory, 1041).

Figure 3-3
Total Cost of Care: Hospital-Setting Innovations (60-Day Lookback)



Finally, we present the TCOC effects for four interventions that we have identified as serving unique populations in ***Figure 3-4.*** These interventions are dissimilar enough from other interventions that it does not make sense to include them in the prior analyses. Because these populations are not comparable to the other settings, or to one another, no overall grand mean was calculated. Only one of these four interventions produced a significant effect, (Sutter-AIM, 1005), a dissavings of $4,818 per beneficiary per quarter.

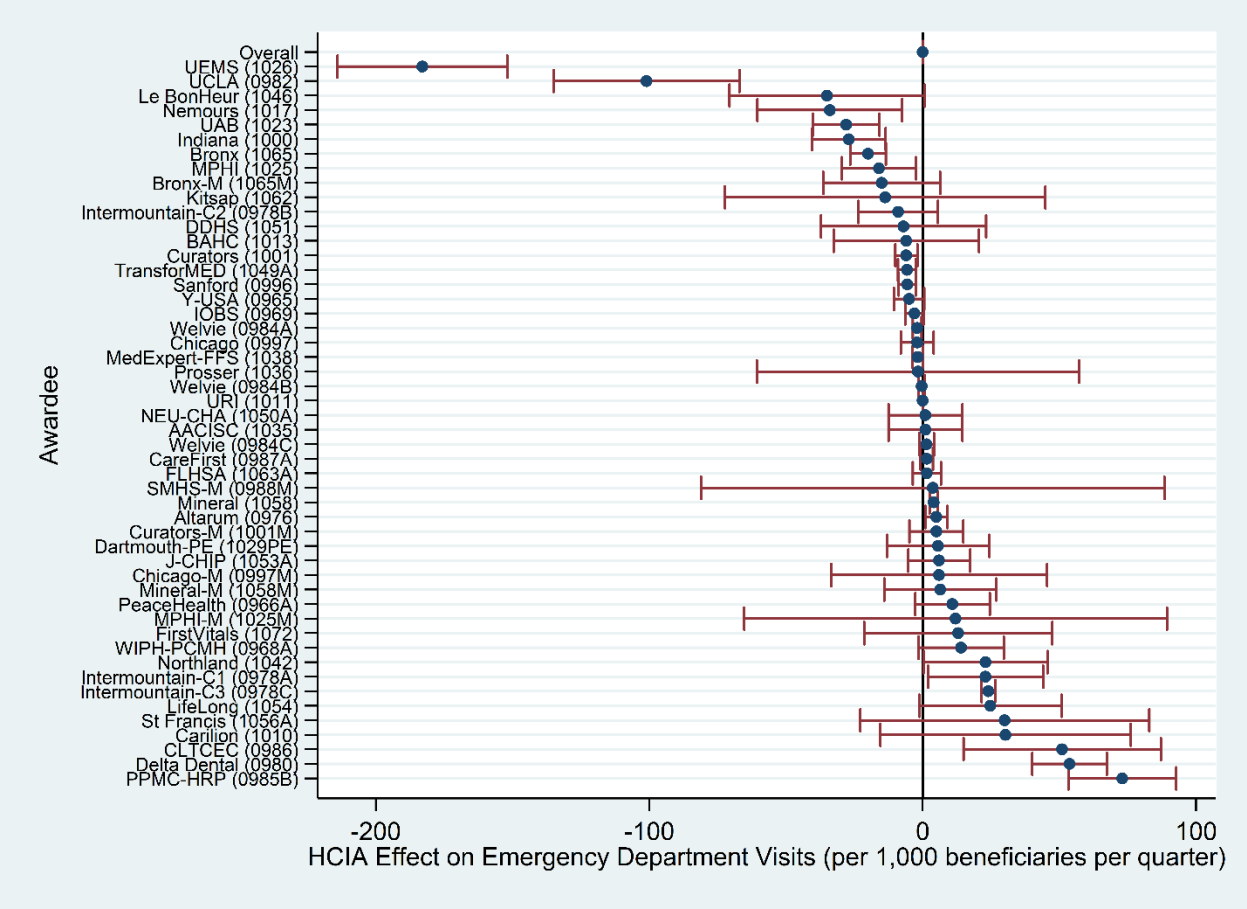
Figure 3-4
Total Cost of Care: Innovations with Unique Populations



3.2.1.2 Emergency Department Visits

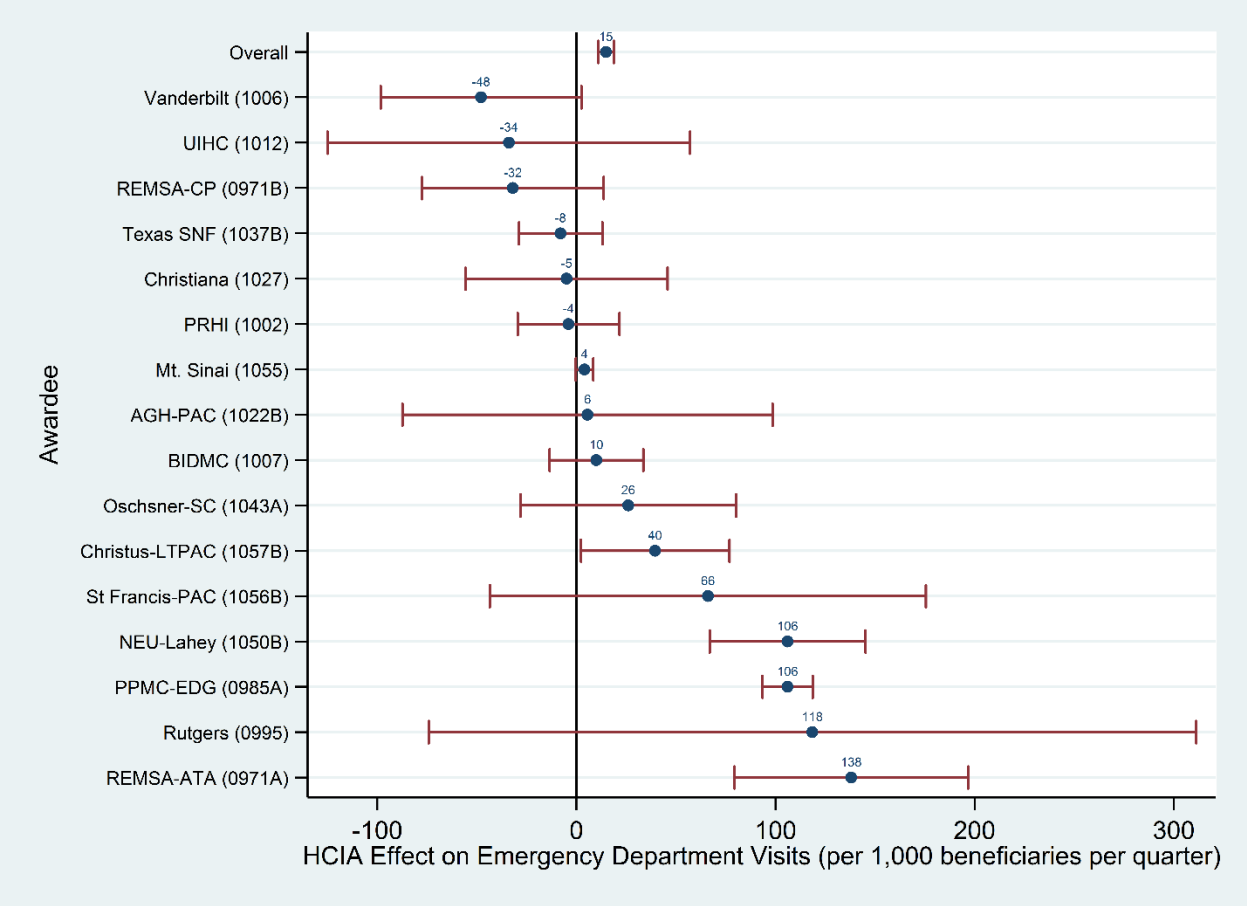
The second core outcome was visits to emergency departments per 1,000 beneficiaries per quarter. ***Figure 3-5*** summarizes the average quarterly effects for the 51 ambulatory care interventions. The number of interventions depicted here is not the same as the TCOC plot because FLEs do not consistently report all outcome measures for all interventions. One intervention (UEMS, 1026) had a particularly large and significant reduction of 183 visits per 1,000 beneficiaries. UCLA (0982) also showed a large reduction of 101 ED visits per 1,000 beneficiaries. Seven other interventions had smaller but statistically significant decreases ranging from 0.02 to 28 visits per 1,000 beneficiaries per quarter (UAB, 1023; Indiana, 1000; Bronx, 1065; Nemours, 1017; MPHI, 1025; Curators, 1001; URI, 1011). The effects for eight interventions showed significant increases in ED visits (Northland, 1042; Altarum, 0976; Intermountain-C1, 0978A; Mineral, 1058; CLTCEC, 0986; Intermountain-C3, 0978C; Delta Dental, 0980; PPMC-HRP, 0985B) while the remaining awardees reported no significant increases or decreases than their comparators. The overall weighted effect for these ambulatory care awardees was just -0.02 visits per 1,000 beneficiaries per quarter (SE = 0.01).

Figure 3-5
Emergency Department Visits: Ambulatory Care Innovations



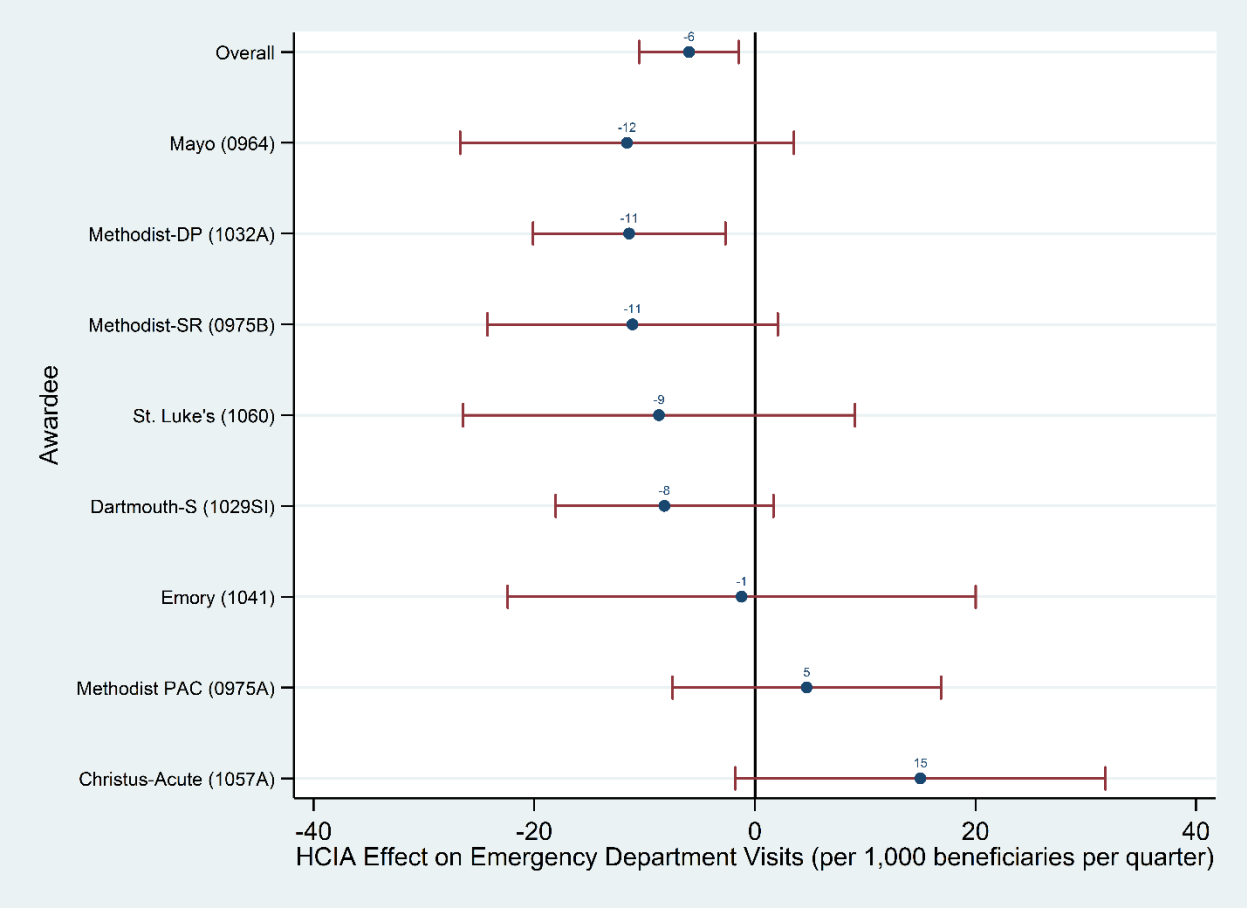
The results for the post-acute care awardees are shown in ***Figure 3-6***. The total of 16 effects come from 12 post-acute setting interventions, 3 ED setting interventions, and 1 long-term care setting interventions. Four interventions showed significant increases in ED visits (PPMC-EDG, 0985A; Christus-LTPAC, 1057B; REMSA-ATA, 0971A; NEU-Lahey, 1050B). None of these interventions demonstrated significant reductions in visits. The weighted mean across awardees was a significant increase in 15 visits per 1,000 beneficiaries per quarter (SE = 2.4).

Figure 3-6
Emergency Department Visits: Post-acute Care Innovations



Finally, the hospital-setting results are shown in ***Figure 3-7***. Of the eight interventions, one had significantly lower ED visit rates (Methodist-DP, 1032A) than its comparison, while the effects for the remaining interventions did not differ significantly from zero. The grand mean of hospital-setting interventions was significant at 6 fewer visits per 1,000 beneficiaries per quarter (SE = 2.7).

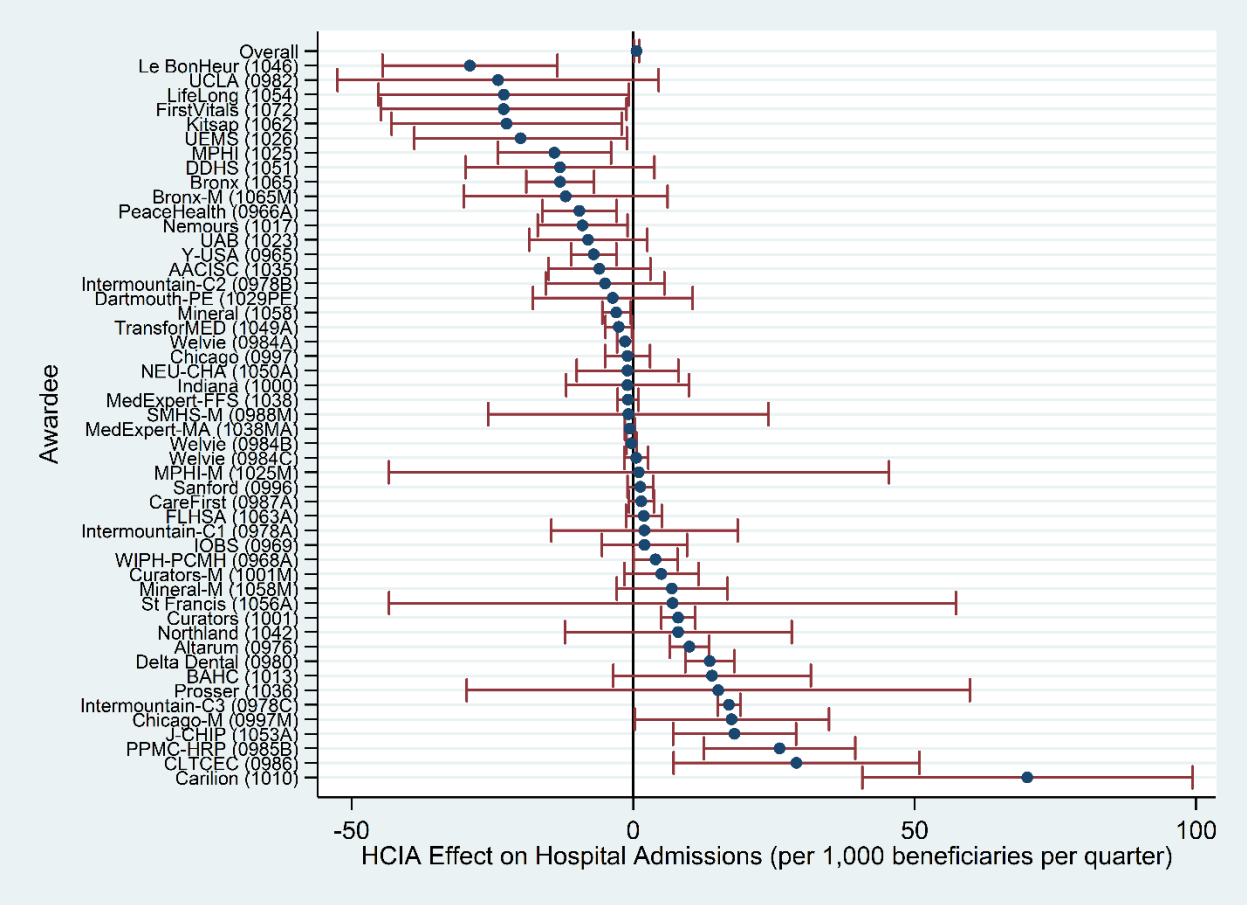
Figure 3-7
Emergency Department Visits: Hospital-Setting Innovations



3.2.1.3 Hospital Admissions

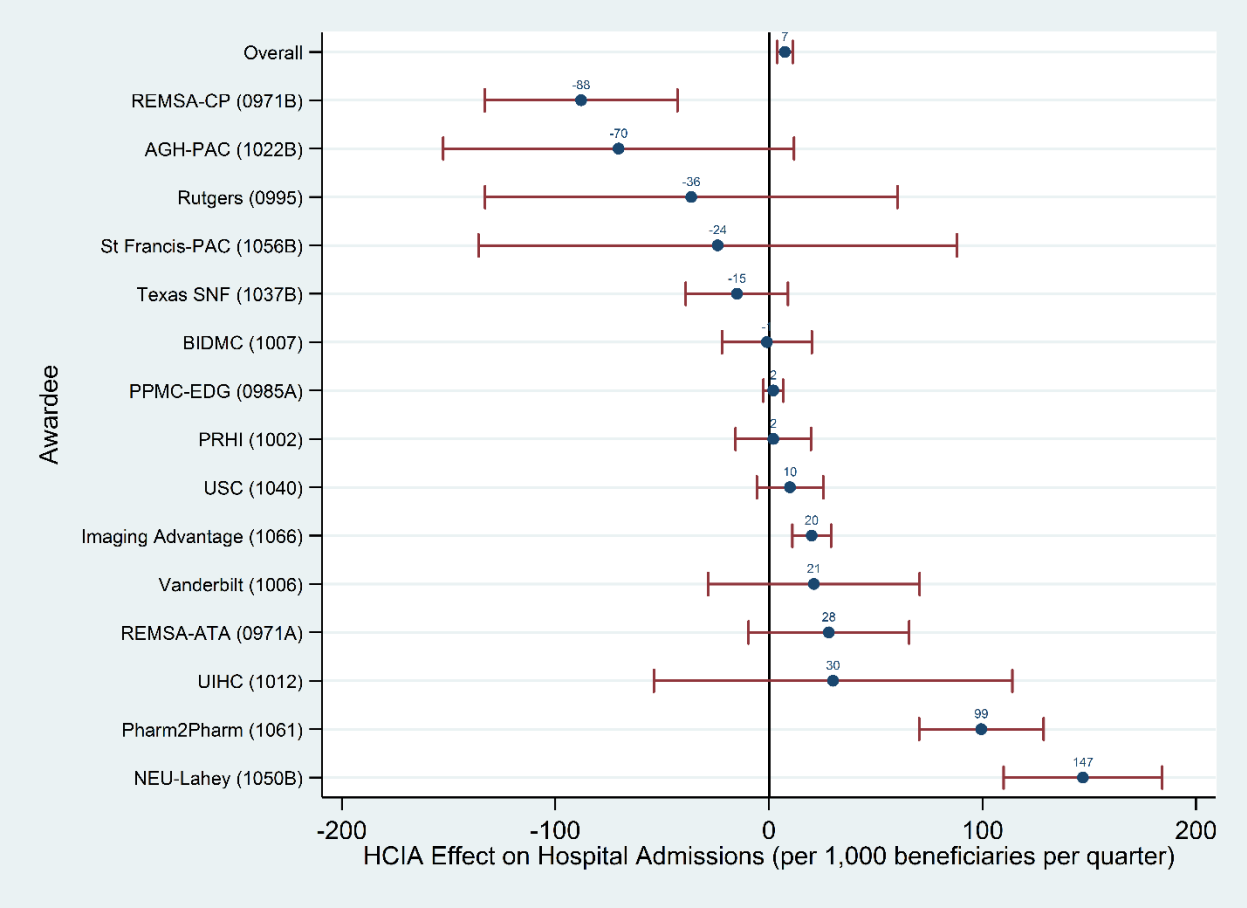
Hospital Admission effects represent differences between innovation and comparison beneficiaries expressed as DiD adjusted rates per 1,000 beneficiaries per quarter. The results for the 51 ambulatory care interventions (***Figure 3-8***) show a mix of favorable and unfavorable effects. Eleven programs achieved significantly lower hospitalization rates (Le BonHeur, 1046; Bronx, 1065; MPHI, 1025; Y-USA, 0965; Kitsap, 1062; FirstVitals, 1072; UEMS, 1026; Nemours, 1017; LifeLong, 1054; Mineral, 1058; PeaceHealth, 0966A), while another nine had significantly higher rates (Chicago-M, 0997M; Curators, 1001; Altarum, 0976; J-CHIP, 1053A; CLTCEC, 0986; Delta Dental, 0980; PPMC-HRP, 0985B; Intermountain-C3, 0978C; Carilion, 1010). The overall average was close to zero, with an average increase of 0.6 additional hospitalization per 1,000 beneficiaries per quarter (SE = 0.26).

Figure 3-8
Hospital Admissions: Ambulatory Care Innovations



Among the 15 interventions in the post-acute care group (***Figure 3-9***), most of the statistically significant effects were in the direction of higher hospitalization rates. Three awardees showed significantly more hospital admissions than their comparisons (Pharm2Pharm, 1061, NEU-Lahey, 1050B; Imaging Advantage, 1066). REMSA-CP (0971B) was the only intervention with significantly lower hospitalizations of -88 hospital admissions per 1,000 per quarter. The grand mean cross-awardee effect was an increase of 7 admissions per 1,000 beneficiaries per quarter (SE = 2.2).

Figure 3-9
Hospital Admissions: Post-acute Care Innovations

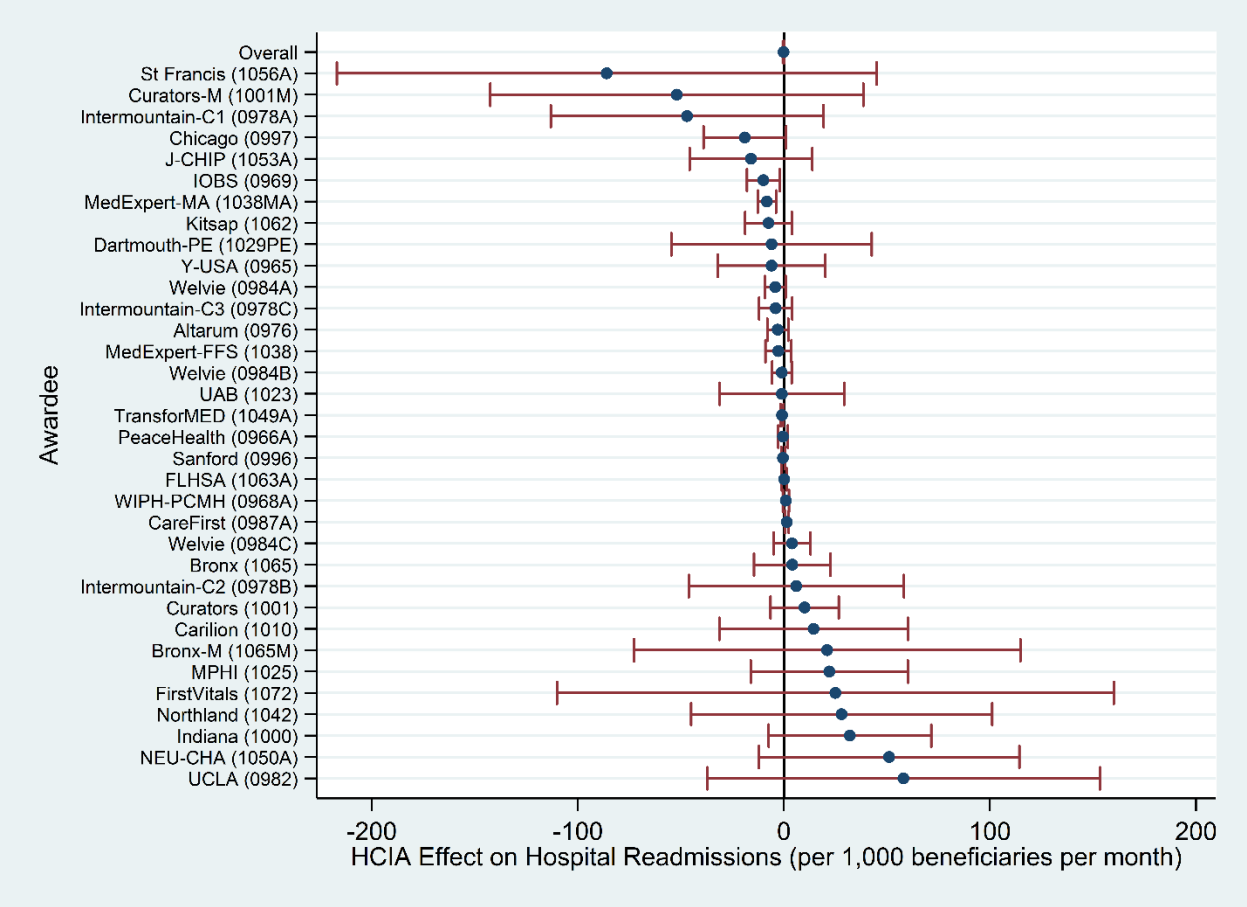


Admission rates were not assessed in the hospital-setting group because patients treated by these awardees were hospitalized around the time of the start of the innovations.

3.2.1.4 Hospital Readmissions

The final core outcome is readmissions within 30 days of an index hospitalization. These rates are relatively imprecise because their denominators are limited to recent hospitalizations. In general, anywhere from 5 to 30 percent of awardees' target populations are hospitalized each year. Accordingly, only 2 awardees of the 52 in the ambulatory care group had effects that significantly differed from zero (***Figure 3-10***). IOBS (0969) and MedExpert-MA (1038MA) reported relatively fewer hospital readmissions per beneficiary per quarter. The overall grand mean was -0.2 readmissions per 1,000 per quarter and did not differ significantly from zero (SE = 0.24).

Figure 3-10
Hospital Readmissions: Ambulatory Care Innovations



Two effect sizes of the 16 in the post-acute care (***Figure 3-11***) group showed significant decreases in readmissions (Rutgers, 0995; Pharm2Pharm, 1061). None of the effect sizes for the hospital-setting (***Figure 3-12***) group differed significantly from zero. Weighted grand mean effects were -0.7 per 1,000 per quarter for the post-acute care group and -1.5 per 1,000 per quarter for the hospital group. Neither of the grand means differed significantly from zero. Readmission rates may be related to the hospital admission core outcome because readmissions are counted in the hospitalization totals.

Figure 3-11
Hospital Readmissions: Post-acute Care Innovations

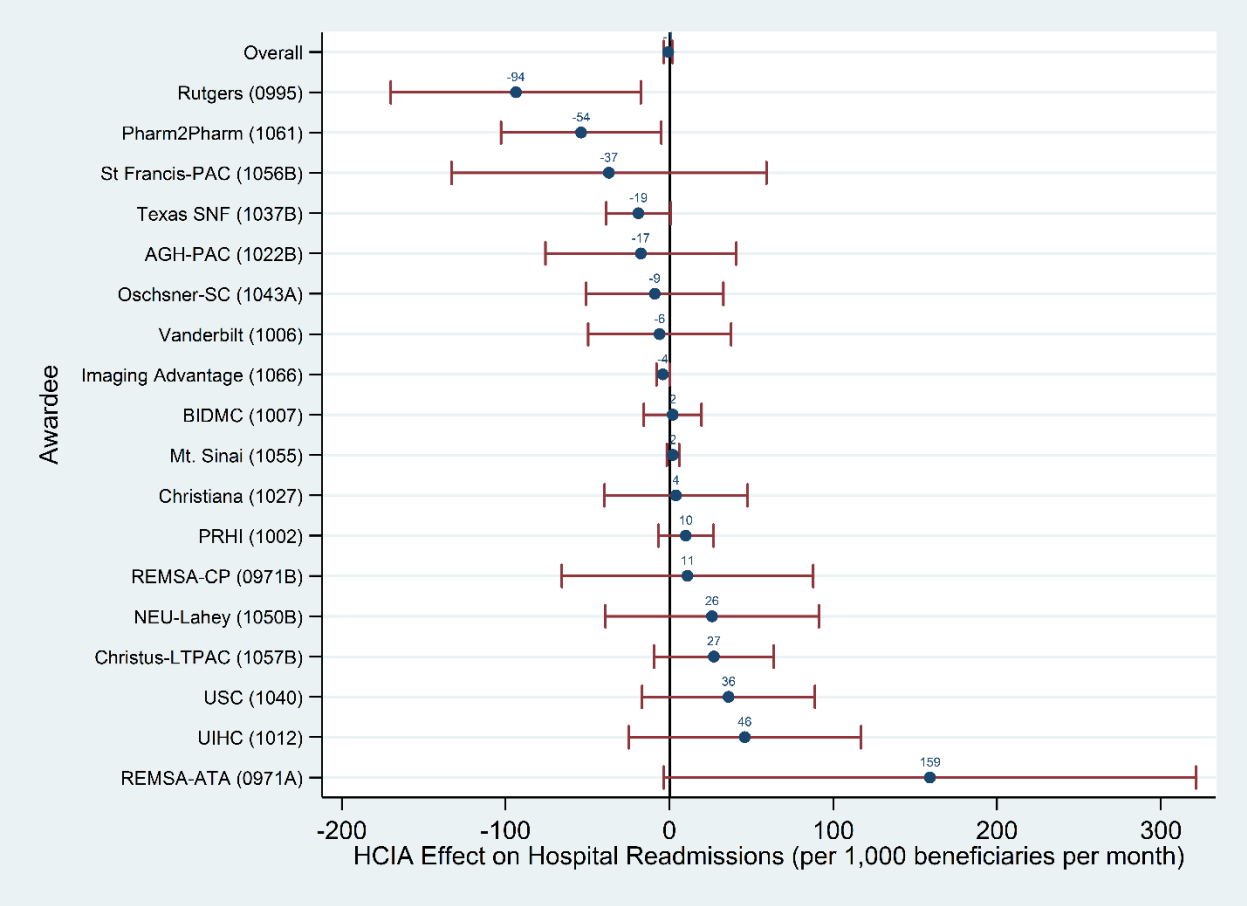
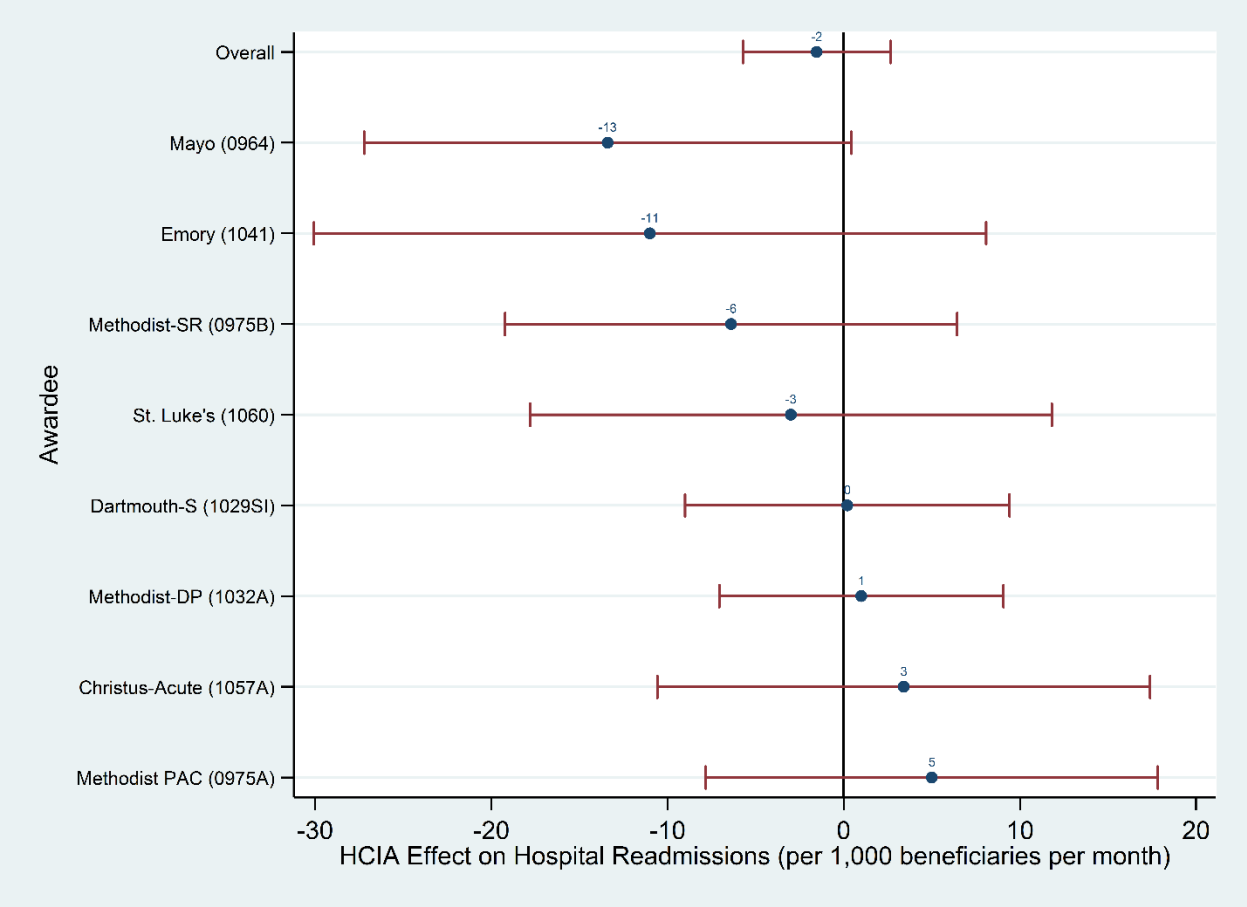


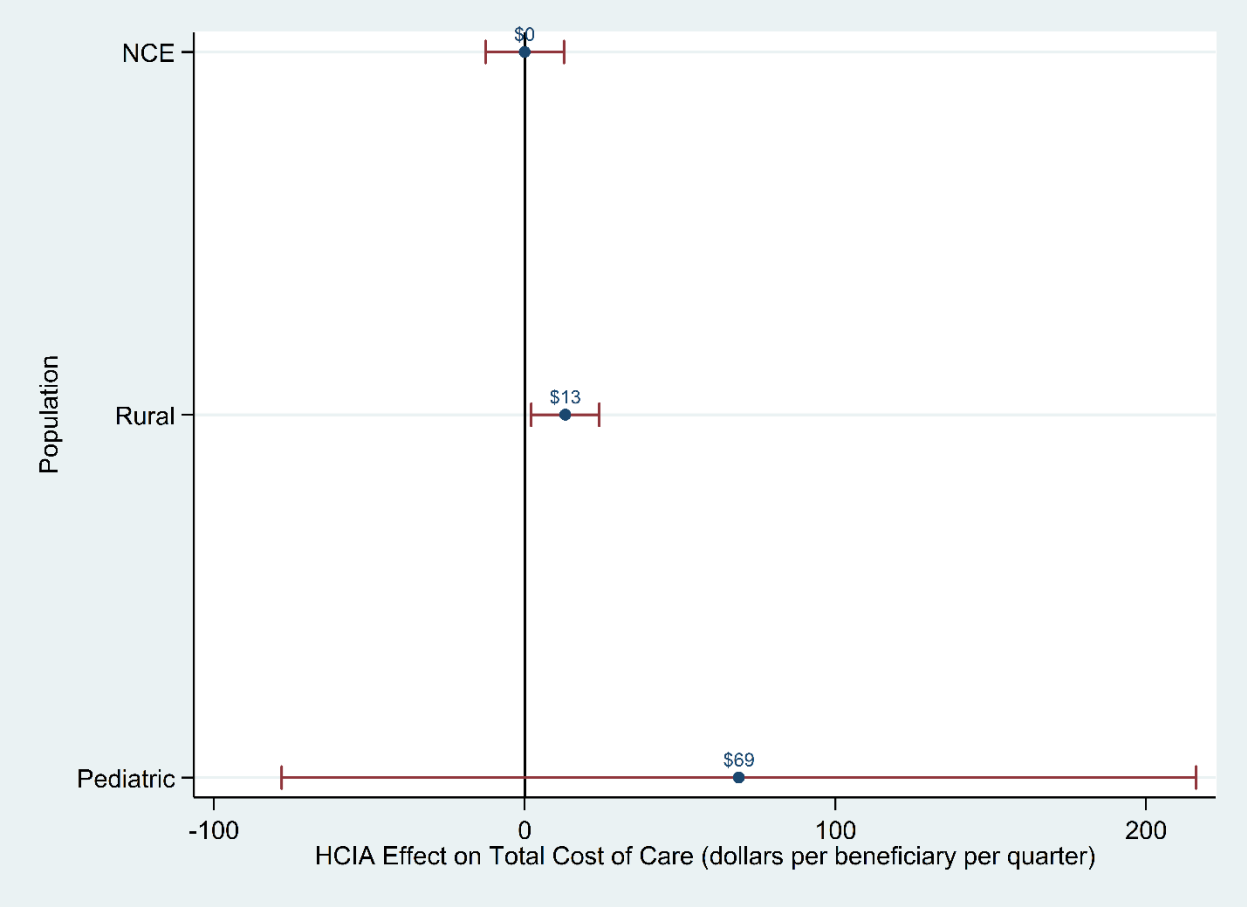
Figure 3-12
Hospital Readmissions: Hospital Setting Innovations



3.2.2 Impact of No-Cost Extension and Rural and Pediatric Interventions on Total Costs of Care

The performance of three subgroups was of special interest to CMMI. We looked at the grand means of the group of awardees that received no-cost extensions (NCE), the group that reported that they serve rural populations, and the group that exclusively served pediatric patients. ***Figure 3-13*** displays the results of this analysis.

Figure 3-13
Impact of Selected Characteristics on Costs of Care



We excluded interventions that served unique populations and those that were explicitly identified by FLEs as being particularly unreliable from our analysis. Because the interventions in the hospital-setting group had effect sizes based on a shorter lookback period, we also excluded them from the analysis.

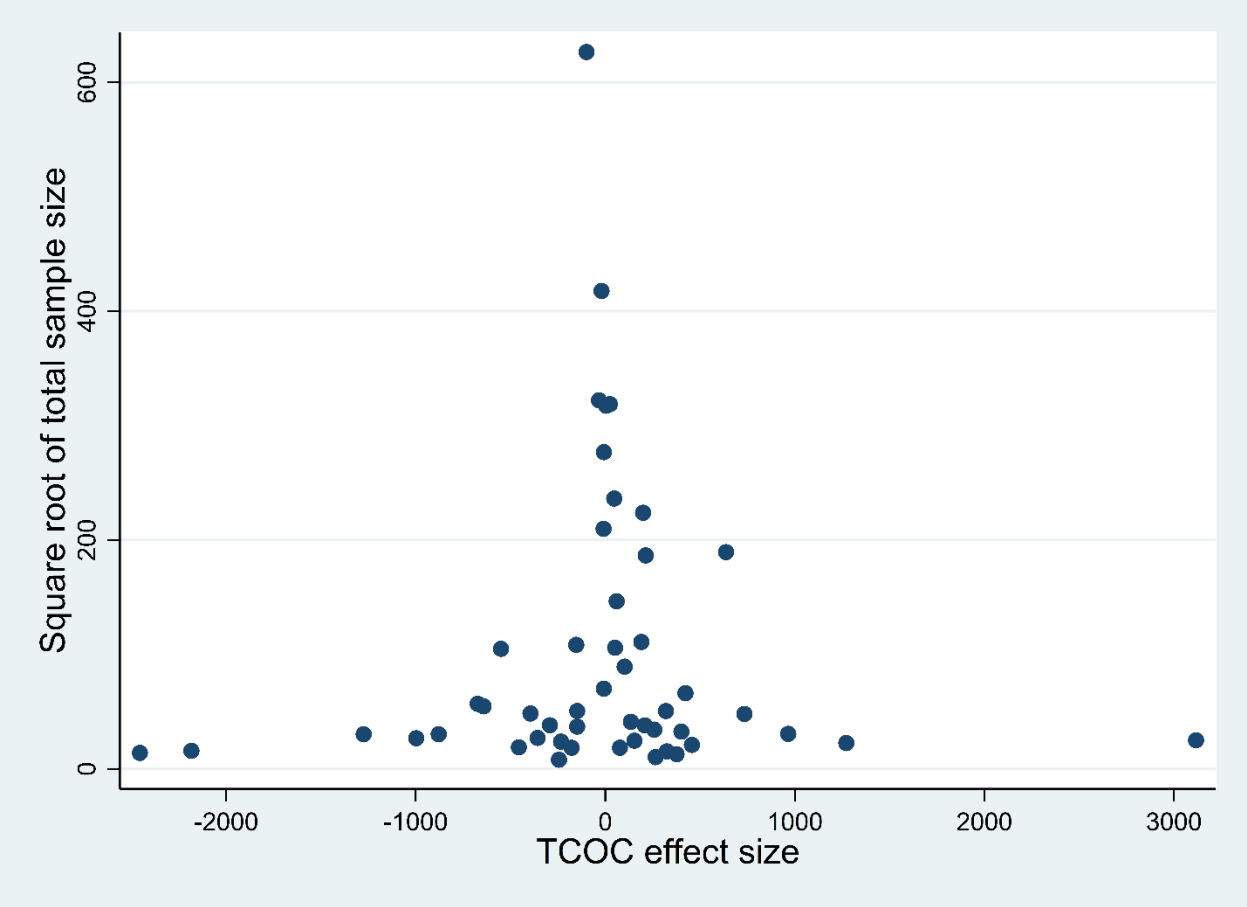
The TCOC grand mean for the interventions that received no-cost extensions was $0 per beneficiary per quarter (SE = $7.66, N = 31). The grand mean for interventions serving rural areas was $13 per beneficiary per quarter (SE = $6.66, N = 40). RTI's project team coded 10 interventions as exclusively serving children (age younger than 18 years). Of those, five had total cost of care effect size data, but one served a unique population and is omitted from the analysis. The overall effect was $69 per beneficiary per quarter (SE = $89.5, N = 4). Thus, the results for interventions receiving no-cost extensions and those exclusively serving pediatric populations would suggest that they did not yield effects of any appreciable magnitude on TCOC. For interventions serving rural areas, significant dissavings were observed at 90 percent confidence.

3.2.3 Sample Size and HCIA Effects

In theory, there should be no relationship between sample size and effect size. In practice, however, an association is not uncommon. Small sample studies can return extreme values and non-reporting of null or negative effects occurs in many literatures. To confirm these results are not subject to untoward bias, we inspected plots of sample size by effect size for symmetry. We now repeat that analysis using additional interventions that now have adequate data. Sample sizes were based on the average of number of quarterly beneficiaries used in DiD analyses during all reported intervention quarters. We computed these quarterly means separately for the innovation and comparison groups, and combined them to derive a total sample size. We then plotted the relationship between the square root of an innovation's sample size and its effect size.

The plot for TCOC among the ambulatory care innovations is shown in ***Figure 3-14***. The relationship resembles an inverted funnel. The five largest innovations all have effects close to zero dollars. The most extreme effects, in both positive and negative directions, are confined to innovations with the smallest sample sizes. Nearly all of the variation in cost of care effects comes from innovations with total samples of less than 3,600 beneficiaries. The Pearson correlation for the data in the plot was $r = 0.046$ ($N = 50$), proving there is little evidence of bias in the FLE-reported findings. This pattern also helps to explain why grand mean effects are so small in our forest plots. Because the large sample innovations have near zero effects and also have the most weight in determining the overall effect size, they tend to draw the grand mean toward zero. Nonetheless, the small correlations in this and ***Figure 3-15*** support the conclusion that the grand mean is close to zero, indicating no effect, regardless of weighting.

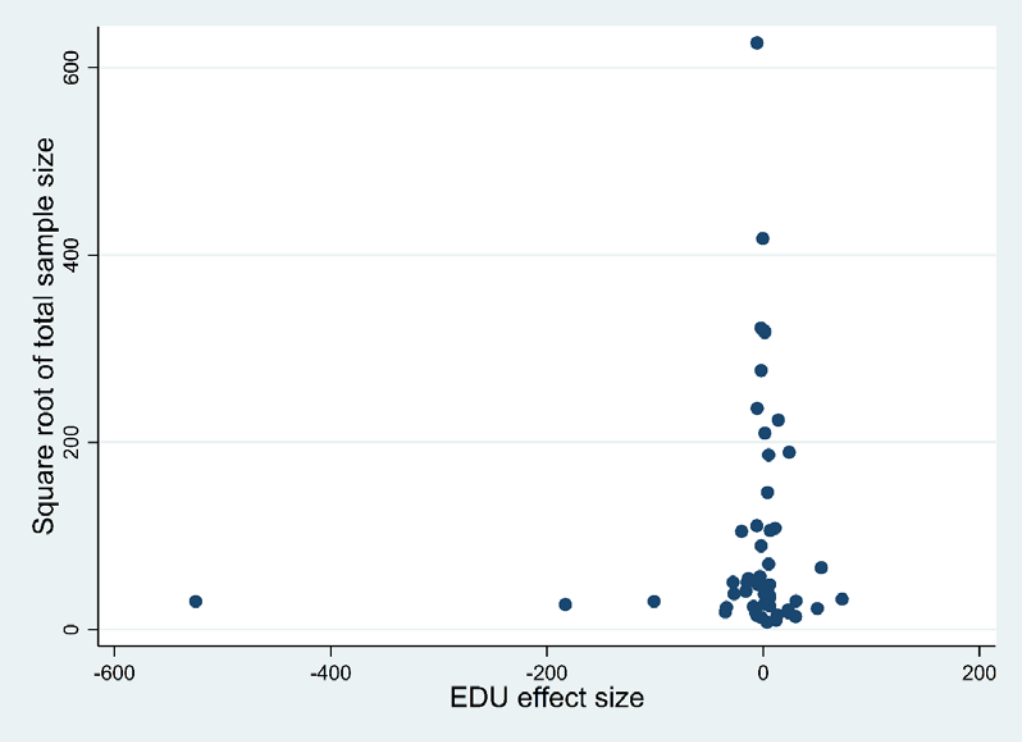
Figure 3-14
Sample Size by Total Cost of Care Effects: Ambulatory Care Innovations



The same funnel pattern is also evident among the smaller number of Ambulatory Care programs that reported effects on ED visits (***Figure 3-15***). Once again, the nine largest studies had almost no impact on visit rates per 1,000 beneficiaries, and most of the effect variation occurred among the smallest innovations. For completeness, the plot includes extreme effect size values. The effect-sample size correlation in this scatterplot was r = 0.095 (N = 50) confirming again little evidence of bias in these results, and that the average impact of innovations on ED visits is near zero.

The funnel plots show a much more normally distributed pattern of results than often occur in literature-based meta-analyses. This suggests that there is little evidence of bias in these results. These results are similar to what we found previously, and we expect that the finding that the largest innovations have impacts close to zero will persist.

Figure 3-15
Sample Size by Emergency Department Visit Effects: Ambulatory Care Innovations



3.3 Impact Effect Heterogeneity

Because of the variety of intervention types, populations, and settings, we expect to see considerable variation in intervention impact. This variation can be seen in the forest plots presented in section 3.2, and can be quantified using statistical tests. In this section we present formal statistical tests of heterogeneity.

We use two measures, Q (Hedges & Olkin, 1985) and I^2 (Higgins & Thompson, 2002) to assess heterogeneity. These are the generally accepted standards for estimating heterogeneity in fixed-effects models, the type of model used to calculate the grand means in section 3.3.1. The Q-test for homogeneity tests the hypothesis that all studies share a common effect size, that is, the variation observed between effect sizes is attributable to sampling error and not to actual differences in intervention effectiveness. A significant p-value is evidence for heterogeneity. The related measure, I^2, estimates the proportion of the total variance (within-intervention variability and between-intervention variability) that is attributable to between-intervention differences. Following convention, I^2 is expressed as a percent. A rule of thumb for interpreting I^2 is as follows: 25 percent indicates low heterogeneity, 50 percent indicates moderate heterogeneity, and 75 percent indicates high heterogeneity (Higgins, Thompson, Deeks, & Altman, 2003). Using Q and I^2, we examined the heterogeneity across interventions for all four core measures for ambulatory care, post-acute care, and hospital-setting interventions. Included in this analysis are interventions for which we had a DiD effect size. Not all interventions provided estimates for all four core measures, thus the number of interventions across the four measures may differ.

We present our findings in ***Table 3-2.*** With respect to total cost of care, the Q-test yields strong evidence that effect sizes vary significantly between interventions ($p < 0.05$) for ambulatory and post-acute interventions. There is also evidence that the effect sizes for hospital setting interventions also vary significantly between interventions ($p < 0.1$). The corresponding I^2 values are also large (between 46 percent and 89 percent, with the 90 percent confidence intervals not crossing 0), indicating that much of the variation is attributable to between-intervention differences. This indicates that it is unlikely that these interventions share the common effect size given by the grand mean.

Table 3-2
Heterogeneity Statistics

Type of Intervention (N)	Q-test statistic (p-value)	I^2 (90% CI)	Interpretation
Total Cost of Care			
Ambulatory (49)	431.15 (p < .001)	88.87% (86.61%, 90.74%)	Strong evidence for heterogeneity
Hospital (8)	13.15 (p = 0.0685)	46.77% (0.39%, 71.55%)	Strong evidence for heterogeneity
Post-Acute (17)	88.13 (p < .001)	81.85% (73.93%, 87.36%)	Strong evidence for heterogeneity
Inpatient Admissions			
Ambulatory (51)	386.1 (p < .001)	87.05% (84.36%, 89.28%)	Strong evidence for heterogeneity
Post-Acute (15)	93.75 (p < .001)	85.07% (78.47%, 89.64%)	Strong evidence for heterogeneity
Hospital Readmissions			
Ambulatory (34)	48.03 (p = 0.044)	31.29% (3.02%, 51.32%)	Strong evidence for heterogeneity
Hospital (8)	4.48 (p = 0.7231)	0% (0%, 40.14%)	Homogeneity
Post-Acute (18)	22 (p = 0.1847)	22.73% (0%, 50.07%)	Homogeneity
ED Use			
Ambulatory (51)	594.5 (p < .001)	91.59% (90.07%, 92.88%)	Strong evidence for heterogeneity
Hospital (8)	8.48 (p = 0.2922)	17.45% (0%, 49.17%)	Homogeneity
Post-Acute (16)	200.36 (p < .001)	92.51% (89.98%, 94.4%)	Strong evidence for heterogeneity

For inpatient admissions, the Q-test provides strong evidence for heterogeneity among both ambulatory and post-acute settings ($p < 0.001$). In each case, the I^2 value also indicates that over 85 percent of the observed variation is likely due to between-intervention differences. It is unlikely that these interventions share the common effect size given by the grand mean.

For hospital readmissions, the Q-test provides sufficient evidence to reject homogeneity for the ambulatory interventions, but does not provide sufficient evidence to reject homogeneity

among the interventions in either the hospital or post-acute setting ($p > 0.1$ in each case). Similarly, the I^2 indicates that the vast majority of the heterogeneity between effects observed in both hospital and post-acute settings is not attributable to between intervention heterogeneity. However, it is important to note that the Q-test (and the I^2 estimate) are sensitive to the number of interventions included in its calculation and it is possible that this or a high degree of within-intervention variance is masking heterogeneity. Even so, the lack of evidence for heterogeneity supports the validity of the grand mean effect and confidence intervals as descriptors of the overall impact of the included interventions on hospital readmissions.

For ED use, the Q-test provides strong evidence for heterogeneity among ambulatory care and post-acute setting interventions ($p < 0.001$) where 92 percent and 93 percent of the variation observed, respectively, is attributable to between-intervention differences. However, there is not strong evidence to reject homogeneity among hospital setting interventions ($p = 0.29$). As before, it is important to remember that Q-test (and the I^2 estimation) are sensitive to the number of interventions included in its calculation. Despite this possibility, at this time the lack of evidence for heterogeneity supports the validity of the grand mean effect and confidence intervals as descriptors of the overall impact of the included interventions on ED usage.

The HCIA programs are ongoing and our findings are preliminary. Nonetheless, the continued presence of heterogeneity among the ambulatory care interventions for total cost of care provides strong evidence that the differences observed are attributable to differences between the interventions and not statistical noise. Because of this, we expand our analysis beyond the quantification of heterogeneity that we have done in this section to meta-regression in section 3.5.3, which uses key intervention features to explain differences in total cost of care for ambulatory setting interventions.

3.4 Meta-Regression Analyses

3.4.1 Key Innovation Features

Once innovation effects were determined for a critical mass of interventions, one of our next objectives was to examine how the magnitude of these effects might be affected by specific features of an innovation—for example, whether cost savings were consistently greater in innovations providing direct services to patients or if utilization rates were lower for innovations affiliated with academic medical centers. The appropriateness of conducting meta-analysis at this stage is supported by the results in the previous section, which suggest that there is substantial unexplained heterogeneity among innovations that might be attributable to such features.

A major limiting factor in meta-analysis is that having a small number of cases makes it difficult to examine more than a few features at a time. Adding more features generally reduces the precision of estimates of the variables already in the model. As we have seen, there are at most 49 ambulatory care programs that have available DiD-based effect estimates suitable for attributing heterogeneity to key innovation features.

To address this limitation, we took our list of key innovation features, divided it into three clusters, and conducted separate analyses for each cluster of features. This permitted us to limit the number of features in any given analysis to no more than eight variables.

Drawn from a variety of sources, the three clusters we selected are summarized below and explicated in ***Table 3-3***.

- Structural Features: These are pre-existing characteristics of the organization implementing the innovation. Included in this group are payer type, academic affiliation, resources, and two measures of previous experience implementing innovations.

- Innovation Components: These features consist of four common types of interventions and indicators for target populations that consist largely of clinically or socially fragile patients.

- Implementation Features: This cluster contains characteristics that were expected to influence the degree to which the planned intervention was successfully implemented. Also included is the summary measure of implementation effectiveness described in section 2.2 of this report.

Another cluster we considered was one for Design Features to characterize patient selection methods, measurement approaches, and statistical analyses. However, the CMMI awardee protocols imposed uniformity in the way that outcomes were measured, and all awardees used similar DiD models to estimate innovation effects. While there were several different approaches to creating comparison groups, our earlier analyses showed that the type of comparison group did not influence the magnitude of cost savings in ambulatory programs (see section 3.2.1). As a result, we did not conduct further analyses of design features.

Table 3-3
Key Innovation Features by Cluster

Structural Features		
Feature	Measurement	Source*
Medicaid/Medicare Advantage payer	Yes/no; payer is Medicaid or Medicare Advantage rather than Medicare FFS	FLE reports
Resource adequacy	Adequacy of site's financial, training, and physical equipment resources	AASF1
Previous demonstration participation	Yes/no; Participating in a CMS shared savings program	AASF1
Had for-profit tax status	Yes/no	Lewin reports
Had an academic affiliation	Yes/no	Lewin reports
Was experienced in implementing similar programs	Not at all /To great extent	AASF2 Item 10d
Innovation Components		
Feature	Measurement	Source
Used health informatics	Yes/no	SQC
Provided behavioral health	Yes/no	SQC
Used telemedicine	Yes/no	SQC
Used community health workers	Yes/no	SQC
Delivered to a clinically fragile population	Yes/no (populations that are clinically complex or at risk for disease progression)	SQC
Delivered to a socially fragile population	Yes/no (populations at risk due to social circumstances or barriers)	SQC
Implementation Features		
Feature	Measurement	Source
Provided direct services	Yes/no	SQC
Was a new innovation	Yes/no; (did not exist or was not piloted pre-HCIA)	SQC
Was a multisite innovation	Yes/no	AASF2
Delivered in a rural location	Yes/no	SQC
Received no-cost extension	Yes/no	CMMI
Experienced barriers to patient recruitment	Yes/no	Lewin reports
Experienced staff turnover challenges	0=Not a challenge, 100=major challenge	AASF2 item 16f
Intervention was implemented effectively	Multi- item scale; 0=lowest effectiveness, 100=highest effectiveness	AASF2; see section 2.2

* FLE = Frontline Evaluator; AASF =Annual Awardee Summary Form; SQC = Structured Qualitative Coding

3.4.2 Meta-Regression Results

The results of the heterogeneity testing in section 3.5.1 indicate that there is substantial effect size heterogeneity among innovations, especially for TCOC and for ambulatory innovations. This variation may in part be generated in part by the kinds of awardee features detailed in the section above. In the field of meta-analysis, the predominant method for analyzing heterogeneity is meta-regression. Meta-regression is similar to ordinary least squares (OLS) regression, the classical method for estimating a linear regression model, in the way in which

regression coefficients are estimated and interpreted, although two important distinctions should be noted in the context of this report. First, the dependent variable in the meta-regression model is the DiD innovation effect rather than the actual value of the outcome itself. With respect to TCOC, our outcome is the estimated innovation impact in terms of quarterly savings or dissavings rather than the level of expenditures. A second difference in meta-regression is that observations are weighted by the precision of estimated effects. We used inverse-variance weighted covariance matrices for our analyses. This means that large studies with small standard errors have greater influence on the regression results than smaller studies whose estimates are less precise.

We focus here on ambulatory setting programs for several reasons. First, there are substantial TCOC differences by setting. Mean expenditures during intervention follow-up periods are dramatically lower for ambulatory interventions ($3,188 per beneficiary per quarter) than for the post-acute care interventions ($13,284). Second, the Cochrane Collaboration guideline is that at least 10 studies are needed to conduct meta-regression (Higgins & Green, 2011). After allowing for missing data, only the ambulatory group meets this criterion in our data. Third, the heterogeneity results in section 3.3 indicate that this is the most likely group to have explainable variation in TCOC. We therefore use ambulatory innovation effects on TCOC as the primary outcome. The utilization measures are associated with these costs, and we explore those relationships in the next section.

For the reasons above, the meta-regressions were limited to ambulatory innovations with comparison groups and regression-adjusted DiD estimates for TCOC effects. In addition to omitting estimates that FLEs had questioned, we also eliminated outlier effects because of the potential bias that can be introduced by only a few outliers in small sample regressions. Outliers were defined as absolute values of more than $1,000 per beneficiary per quarter for TCOC (N = 3), 100 per 1,000 for ED rates (N = 2), and 50 per 1,000 for hospital admissions (N = 1). These are equivalent to 15 percent and 18 percent, respectively, of national utilization patterns and 42 percent of quarterly Medicare FFS expenditures. These exclusions left a pool of 43 ambulatory interventions available for analysis.

All meta-regression models were estimated from weighted covariance matrices in which individual interventions have been weighted by the inverse variance of their TCOC estimates. For each model, we show the unstandardized regression coefficient and its standard error, the zero-order weighted Pearson correlation, and the mean or percent of innovations with the feature in the analysis group. This weighting can alter the prevalence of some features compared to the rates for the entire set of awardees. The standard errors are helpful for assessing how precisely we can determine the cost impact of any particular feature. As shown in the forest plot in ***Figure 3-1***, after eliminating outliers TCOC effects ranged from -$1,000 to $1,000 per beneficiary per quarter with an average value near zero dollars.

The regression results for the six structural innovation features are shown in ***Table 3-4***. Forty ambulatory innovations had complete data and were included in the analysis. All of these measures except for-profit status had point-biserial correlations with TCOC effects that were near zero. Adjusted for the other characteristics in the model, for-profit interventions had savings that averaged $249 more per beneficiary per quarter than nonprofit and government-based programs. The two indicators for previous experience (with either shared savings programs or

programs similar to the intervention) had very weak effects. HCIA impacts were not affected by whether the payer was Medicare FFS, Medicare Advantage, or Medicaid.

The results for the meta-regression based on innovation components is shown in ***Table 3-5***. Interventions with health informatics (-$233) or community health worker (-$157) components had lower costs per beneficiary per quarter relative to a comparison group. Per beneficiary per quarter costs were also lower in interventions serving clinically fragile patients.

Table 3-4
Meta-Regression Results for Structural Features of Innovations (N = 40)

Feature	Costs per beneficiary per quarter (standard error)	Correlation with TCOC	Percent or mean
Medicaid/MA payer	-37 (85)	-0.01	29%
Resource adequacy	8.21 (34.5)	-0.03	11.0
Previous demonstration participation	84 (118)	0.05	57%
For-profit organization	-249* (104)	-0.21	19%
Academic affiliation	-38 (135)	-0.00	9%
Experience implementing programs	1.54 (1.67)	0.07	66.2

* $p < 0.05$; $R^2 = 0.109$

Table 3-5
Meta-Regression Results for Innovation Components (N = 40)

Component/Feature	Costs per beneficiary per quarter (standard error)	Correlation with TCOC	Percent
Health informatics	-233* (108)	-0.23	69%
Behavioral health	-98 (99)	-0.03	16%
Telemedicine	157 (102)	0.11	9%
Community health workers	-157+ (86)	-0.16	36%
Clinically fragile population	-175+ (90)	-0.24	40%
Socially fragile population	119 (214)	-0.01	6%

* $p < 0.05$; + $p < 0.10$; $R^2 = 0.256$

Finally, ***Table 3-6*** presents the model for features relevant to implementation. Two aspects stand out in the list. New innovations generated significantly greater dissavings ($270), as did interventions that experienced challenges with frontline staff turnover or vacancies. A change of one response category on the four-category challenges scale is equivalent to an additional $125 in dissavings. It may be recalled from the path analysis presented in section 2.2 that new programs and interventions experiencing challenges recruiting new staff were somewhat less effectively implemented, albeit not significantly so. Nonetheless, the less effective implementation may contribute to these dissavings findings.

Table 3-6
Meta-Regression Results for Implementation Features (N = 39)

Feature	Costs per beneficiary per quarter (standard error)	Correlation with TCOC	Percent or mean
Direct services provided	-24 (98)	-0.03	69%
New innovation	270** (97)	0.22	19%
Multisite innovation	17 (219)	0.20	95%
Rural location	109 (80)	0.01	72%
Received no-cost extension	60 (69)	-0.19	47%
Reported barriers to patient recruitment	95 (74)	0.01	28%
Staff turnover challenges	4.94** (1.21)	0.37	35.6

** $p < 0.01$; $R^2 = 0.338$

Looking back at all three regressions, we see that the standard errors for binary features ranged in size from around $70 to $135 per beneficiary per quarter. The two features with standard errors near $200 were both instances of prevalences close to 0 percent or 100 percent. In this context, HCIA cost effects would need to be in the $140–$270 range to obtain coefficients that were significantly different from zero. Like any regression analysis, the SEs here are a complex function of the amount of explained variation in the outcome, the variances of the outcome and explanatory variables, the intercorrelations among the explanatory variables, and the sample size.

3.4.3 Meta-Regression Path Model

In this section, we extend meta-regression to create a path analysis model that weaves together several elements of our previous research. This model first links the implementation-related features in the previous section to the FLE-reported summary measure of implementation effectiveness introduced in section 2.2. It then examines the influence of implementation on the magnitude of HCIA effects for utilization and costs. Technical aspects of the estimation methodology are provided in Appendix F.

A common presumption depicted in program logic models is that innovations will reduce utilization, which will in turn reduce costs of care. There is, however, remarkably little evidence to support this contention. In our data, the HCIA measures are innovation effect sizes, not actual utilization rates or total expenditures but the underlying logic is the same. We expect that HCIA utilization effects should be positively associated with effects on TCOC because innovations that reduce utilization relative to a comparison group should also experience a relative reduction in costs. In our last report, we found statistically significant positive correlations between HCIA effects on TCOC and on both hospital admissions and ED visits, but that was based on only 20 ambulatory care innovations with sufficient data. The more recent FLE reports have allowed us to nearly double the number of analyzable innovations. The path analysis framework also enables us simultaneously consider the impact of each type of utilization on cost effects.

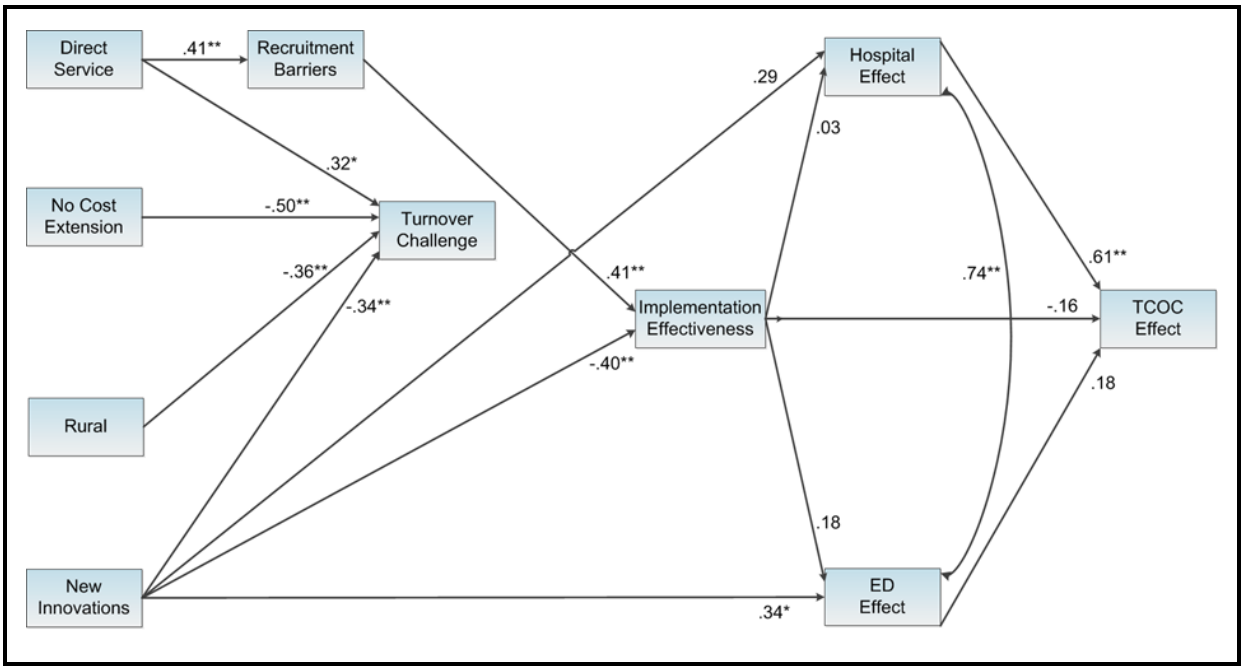
A path analysis assessed relations between potential drivers of implementation effectiveness, utilization, and total cost of care. Significant results are summarized below.

- New interventions and those providing indirect services, receiving no-cost extensions, or based in rural communities experienced fewer turnover challenges than nonrural interventions or interventions that used their award to expand services, provide direct services, or did not receive a no-cost extension.
- Interventions that provided direct services experienced greater recruitment challenges, but counterintuitively, as judged by FLEs, greater recruitment challenges were associated with more effective implementation.
- Effective implementation had minimal direct effect on the total cost of care.
- Although ED visits had little direct effect on total cost of care, it was strongly associated with hospital utilization, which was strongly associated with total cost of care.
- In this model, total cost of care is estimated to fall by $155 per beneficiary per quarter for every decline of 10 beneficiaries per 1,000 in hospital admission rates ($p < 0.000$).

The estimated path model with standardized coefficients is shown in ***Figure 3-16***. As with the implementation effectiveness path model, standardized relations range from -1 to 1, with zero indicating no relationship and -1 and 1 indicating perfect negative and positive relations, respectively. Effects in the model flow from the innovation features at the far left, through implementation effectiveness, to hospital and ED utilization, and ultimately to the TCOC effect size, which is a negative dollar value for innovations exhibiting estimated savings, and a positive value for dissavings. We do not include hospital readmission effects, because they are subsumed under all hospital admissions and are not evaluated for some awardees. Thirty-seven innovations had complete implementation and core outcome data. A simple base model, which restricted the right half of the model to only four paths from implementation effectiveness to utilization and utilization to cost effects, did not fit the data well. In the absence of theoretical

hypotheses, we examined modification indices to suggest features that might be influential for utilization effects. The final version of the model depicted in the figure provided a better, but still marginal fit to the data (RMSEA = 0.128; 90% CI = 0.010–0.211; CFI = 0.906). For clarity, the left side of the model has been trimmed to display only paths with beta coefficients exceeding 0.30 in absolute value. In other words, although other paths are tested in the model, only the substantial and statistically significant relations are displayed.

Figure 3-16
Path Model of Implementation Features, Implementation Effectiveness, Utilization, and TCOC Effects (N = 37 ambulatory setting innovations)



*p < .05; **p < .01

The left side of the model contains relationships among the implementation-related measures. The multisite indicator was dropped from this analysis because its weighted prevalence exceeded 95 percent. Recruitment barriers were reported more often by innovations delivering direct services to patients. All four of the exogenous features affected staff turnover reports. Frontline staff turnover or vacancies were more of a challenge for rural, no-cost extension, and new innovations, and less of challenge for awardees delivering direct services.

The key hypothesized mediator in the model is implementation effectiveness. Only two features had sizable impacts on this variable. New innovations had considerably lower implementation effectiveness summary scores than existing innovations. Innovations reporting barriers to patient recruitment actually had higher implementation effectiveness scores, although it is not clear why this should be unless these were earlier challenges that awardees had successfully surmounted by the time of the AASF administration.

The far right side of the model summarizes the relationships among the effect sizes for the core measures. In our data, there were strong associations among HCIA effects on TCOC and

the effects on both utilization measures (weighted Pearson correlations ranged from 0.63 to 0.76). The correlation between the two utilization outcomes could not be explained by implementation effectiveness or any of the other features in the model. As a result, the model contains an unexplained correlation (represented by the curved arrow) between hospitalization and ED impacts. The paths leading to TCOC show that hospitalization effects have a much larger impact (beta = 0.61) on TCOC effects than ED effects have (beta = 0.18). Using unstandardized coefficients, total costs were estimated to fall by $155 per beneficiary per quarter for every decline of 10 beneficiaries per 1,000 in hospital admission rates ($p < 0.000$), but to decrease by only $22 per quarter for a decline of 10 ED admissions per 1,000 ($p = 0.248$).

The results also highlight the relevance of new innovations. Awardees instituting new innovations had significantly lower implementation effectiveness scores. New innovations were positively associated with both hospitalization and ED effects, which helps to explain why they had significantly higher TCOC dissavings in the regression in the previous section. New innovations are clearly harder to implement and produced less favorable results effects for all three core outcomes compared to existing programs that awardees were primarily seeking to expand. Implementation effectiveness, on the other hand, had only a minor impact on any of the core outcomes. It had a small but insignificant direct effect on TCOC effect sizes (beta = -0.16, indicating a small amount of cost savings), but no notable influence on either hospitalization or ED rates.

3.5 Supplementary Analyses

In addition to the primary analyses reported above, we also conducted several supplementary analyses to determine the extent to which our results might be affected by applying different analysis methods. These analyses and findings are detailed in four appendices.

The supplementary analyses indicated that:

- The method that FLEs chose to construct comparison groups (matching, weighting, or randomization) did not appear to systematically affect the magnitude of HCIA innovation effects on the core outcomes (Appendix I).

- Using a synthetic comparison group in situations when only pre-post treatment data were available resulted in considerably higher estimated cost savings than conventional pre-post analyses (Appendix J).

- A Bayesian fixed-effects meta-analysis yielded nearly identical results with respect to estimated effects and confidence intervals as the more conventional frequentist analysis (Appendix K).

- A time-series analysis of quarterly health care expenditures was consistent with the results of the main meta-analysis, showing only minor differences between innovation and comparison group costs over time (Appendix L).

3.6 Quantitative Findings Summary

3.6.1 Evaluability Assessment

As in previous reports, we report awardee findings at the intervention level. Of the 108 awardees, 22 (20.4 percent) implemented and tested multiple interventions. This increased the number of possible innovations tested from the 108 original awards to 135 interventions. We received usable DiD estimates for at least one core outcome (i.e., total cost of care, hospitalizations, hospital readmissions, and emergency department use) from 68 awardees. This represents a 23.2 percent increase in available data from our last report. It should be noted that these results are based on findings reported by FLEs as achieved during the second year of the award and the HCIA programs are ongoing. These findings must be considered preliminary results, which we expect to change during the third year of awardee funding.

To improve our capacity to identify awardee-related sources of heterogeneity from population- and setting-sources of heterogeneity, we partitioned awardees into the three broad classes: ambulatory care, post-acute care, and hospital-setting. DiD estimates were available for 39 of 97 (40.2 percent) ambulatory care interventions. Most of the post-acute interventions reported a DiD estimate for at least one core measure (18 of 29 interventions or 62.1 percent), and all of those reported for three or more of the measures. Eight of the nine interventions in the hospital setting had DiD estimates.

3.6.2 Comparison group methods

The most common method for creating comparison groups was one-to-many propensity score matching (PSM; N = 32), followed by one-to-one PSM (N = 24), propensity score weighting (PSW; N = 10), and facility- or provider-level PSW (N = 8). Each FLE confirmed the quality of comparison groups created with propensity scores using balance tables, four included propensity score plots, and two reported results from omnibus tests for balance of the variables. Four interventions used randomized control group designs and 20 used another method of comparison. Due to problems with data availability, feasibility, or barriers to forming a comparison group, 36 interventions do not have comparison groups.

An assessment of the methods used to generate comparison group showed that no particular comparison group method produces systematically different effect sizes from those produced by other comparison group construction methods. Moreover, additional tests found comparison groups functionally interchangeable. Therefore, we created synthetic comparison groups using data provided by other awardees for four awardees who provided only pre-post estimates of TCOC.

3.6.3 Meta-analysis

As in earlier reports, the impact of these interventions on TCOC, hospitalizations, hospital readmissions, and ED visits show a range of effects, with some interventions producing positive results, some negative, and most showing no difference in impact relative to their comparators. Across settings and outcomes, the average effect of interventions remains close to zero, indicating no appreciable impact on the core outcomes. Much of this result is likely due to several large studies contributing data to the synthesis. Since meta-analytic results are weighted

by the inverse of their variance, large samples can strongly influence an analysis, and results for these large studies were near zero. As expected from sampling theory, smaller studies are dispersed evenly around the main effect, suggesting that, while large studies likely increased our confidence in the null result, it is unlikely they influenced greatly the overall finding of no significant impact.

Across ambulatory, post-acute, and hospital settings 11 interventions produced significant total cost of care savings ($p < .10$) while 14 produced significant dissavings ($p < .10$). The remaining 43 interventions had no significant impact on total cost of care. An additional 4 interventions, which included unique populations and were not summarized in the above syntheses, did not significantly impact total cost of care.

Total cost of care impact estimates from the interventions were then broken out by whether they received a no-cost extension (N = 31), were serving rural areas (N = 40), or were serving children exclusively (N = 4). The average impact on total cost of care for awardees receiving a no-cost extension was $0.00 (SE = 7.66; NS), for awardees serving rural areas $13 (SE = 6.66; $p <.10$), and for those serving children was $69 (SE = 89.5, NS).

Similar results were documented for ED visits. Of the 75 innovations providing data, 10 produced significant reductions in ED visits, 12 had significant increases in ED visits, while the remaining 53 innovations effects showed no significant impact on ED visits.

Because beneficiaries treated by these hospital-setting awardees were hospitalized around the time of the start of the innovations, admission rates were not assessed in the hospital-setting group. Among the 51 ambulatory and 15 post-acute awardee interventions, significant decreases in hospital admissions were achieved by 12 interventions and significant increases were observed for 12 interventions ($p < .10$).

The final core outcome is readmissions within 30 days of an index hospitalization. These rates are relatively imprecise because their denominators are limited to recent hospitalizations. Across awardees, between 5 and 30 percent of beneficiaries were hospitalized in a given year. Effect sizes were available for 76 innovations. Of these, 4 showed significant decreases in hospital readmissions ($p < .10$). The impact of the remaining 72 interventions did not differ significantly from zero.

3.6.4 Bayesian and Time-Series Analyses

As a complement to the meta-analysis and as a check on the findings produced by that method, we reanalyzed the data using the analogous Bayesian meta-analytic method. The Bayesian approach reproduced the results obtained using the standard (frequentist) methods above. These were then tested to predict the probability of savings beginning with the assumption that each setting began with a 50 percent chance of savings and a 50 percent chance of dissavings. Informed by the data, the probability of savings in each of these settings shifted considerably. For ambulatory setting innovations, the probability of savings shifted to 3 percent (probability of dissavings 93 percent), a zero percent probability of savings was observed (100 percent probability of dissavings) for post-acute setting innovations, while for hospital-setting

interventions, the probability of savings was 64 percent (with the corresponding probability of dissavings equal to 36 percent).

The time-series analysis results likewise replicated the overall findings of no effect when comparing intervention and comparison groups over time, but highlighted the similarity of estimates produced within intervention and comparison groups. Effects within groups were more similar to each other than they were between groups with the intervention group showing consistently higher costs than the comparison group, both before and following intervention implementation. Additional analyses using this approach echoed results from the DiD synthesis: a similar dispersion of effects is observed, no undue influence of sample size on effect size was found, and no significant findings for savings or dissavings across settings emerged.

3.6.5 Explaining Heterogeneity

Two standard meta-analytic measures (Q and I^2) were used to assess if the variability in intervention results was within that expected from sampling error (i.e., the expected variation of results based on samples) or if variation might be due to substantive differences between interventions. If variation in the distribution of results exceeds that expected from sampling, it is possible to test if that excess can be explained by measured differences between interventions. Our analyses found:

- Innovations implemented in ambulatory settings produced results that varied more than expected from sampling for each of the four core outcomes.

- Innovations implemented in post-acute settings produced results that varied more than expected from sampling for TCOC, inpatient admissions, and ED visits. Results for hospital readmissions were within the limits expected from sampling.

- Innovations implemented in hospital settings produced results that varied more than expected from sampling for TCOC, but results were within the range expected from sampling for both the hospital readmissions and ED visits outcomes. Since beneficiaries were hospitalized around the start of the innovation, hospital admissions was not an outcome for these interventions.

Meta-regression is a principal method of testing for sources of excessive variation. However, as with standard multiple regression, a sufficient number of cases are necessary to meet the statistical assumptions of the method. To address this limitation, we tested three models, regressing total cost of care, respectively, on six structural features, six innovation components, and seven implementation features, with the following results:

- After controlling for other structural characteristics, interventions implemented by for-profit organizations produced savings that averaged \$249 (SE = \$104; $p < .05$) more per beneficiary per quarter than nonprofit and government-based programs.

- After controlling for other innovation components, interventions using health informatics produced savings that averaged \$233 per beneficiary per quarter (SE = \$108; $p < .05$) more than those not using health informatics. Savings were also

observed for interventions using community health workers ($157, SE = $86, $p < 0.10$) and those working with clinically fragile populations ($175, SE = $90, $p < 0.10$).

- After controlling for other implementation features, new innovations and those experiencing challenges from staff turnover both showed significant dissavings. New innovations on average cost $270 more per beneficiary per quarter than innovations that were being expanded (SE = $94, $p < .01$) and those challenged by staff turnover averaged $4.94 more per beneficiary per quarter than those not so challenged (SE = $1.21; $p < .01$).

To test the common presumption in many awardee logic models that the innovation would reduce utilization and thus costs, we used path modeling to see if innovations reducing hospital admissions and ED visits relative to a comparison group experienced a relative reduction in costs. We found that reducing hospital admissions was significantly related to lower per beneficiary per quarter costs ($r = .61$, $p < .01$), but no significant relation was observed for ED visits. In dollar terms, these findings indicate that total costs would be expected to fall by $155 per beneficiary per quarter for every decline of 10 beneficiaries per 1,000 in hospital admissions, but to decrease by only $22 per beneficiary per quarter for a decline of 10 ED admissions per 1,000. New innovations had unfavorable impacts on both forms of care utilization and were more difficult to implement effectively. However, in this model, implementation effectiveness was not significantly related to total cost of care.

[This page intentionally left blank]

SECTION 4:
DISCUSSION

One lesson from this project is the importance of collecting information from FLEs in a standardized way. While FLE reports were helpful in highlighting major implementation and process themes, they often did not provide awardee-specific information on the full range of issues awardees faced and the extent to which they experienced them. To fill this gap, we developed a new version of the AASF in 2015 to gather detailed systematic information about FLE perceptions of awardee performance and challenges. In this discussion we relate the themes identified through qualitative coding with evidence submitted by the FLEs to extend our description by proposing plausible connections across findings.

Most innovations were implemented effectively, but adaptation was often instrumental in achieving implementation effectiveness.

For many awardees, attracting enrollees, building partnerships and forging relations among staff, and implementing their innovations was a far greater challenge than was expected. In response to these challenges, however, by the end of the second year, awardees learned that by adjusting their enrollment strategies; taking time to build trust, respect, and appreciation among partners and staff; and adapting their innovations, they were largely able to meet these and other challenges and to effectively implement their innovations.

For the 13 innovations that had HIT as a principal component, and the 77 of 135 interventions that identified HIT as a challenge in implementing their intervention, by the end of the second year, the majority of those challenges were rated modest or small by FLEs. In the second year, innovation HIT systems were refined to better reflect workflow needs and were increasingly integrated with existing HIT systems. With these adaptations, staff increasingly recognized and appreciated the added value of HIT.

Integrating CHWs and other non-licensed staff was a challenge that followed a similar trajectory to HIT. Early in their innovations, approximately half of awardees and innovations using CHWs were challenged by integrating non-licensed staff into their existing staff workflows. Role clarification and delineation of responsibilities helped existing staff accept non-traditional workers. Awardees using CHWs quickly learned the importance of hiring non-licensed staff with the right temperament and training to meet role demands. Combined, these adaptations increased staff appreciation of CHWs for their contributions in improving workflow and connecting with patients. For some awardees using non-licensed staff, a continuing challenge is obtaining reimbursement for services provided by these non-licensed health care workers.

While adaptations have facilitated enrollment and implementation for awardees, and have led to improved workflows, changes in enrollment standards present challenges for drawing conclusions about how innovations achieve cost savings and reduce utilization. Adaptation can take two forms: first are changes that improve the surface structure of delivering innovation core components, second are revisions to the components themselves. While it appears that the core components of most innovations did not change greatly, many features associated with implementing those components required adaptation to meet foreseen and unforeseen challenges.

A strong accounting of adaptations, and how fundamental they were to the innovation, would permit a stronger test of how innovations are successful, which components can and cannot be altered, and what components are essential for improving cost savings and reduced health care utilization.

Prior experience allowed awardees to better anticipate and address challenges, helped them anticipate staffing needs and requirements, and was instrumental in sustainability planning and innovation scaling. Staff with greater experience with change were likely more comfortable adapting to accommodate the innovation. Innovation leaders with prior experience may have better understood what features of the innovation were essential and may recognize opportunities to facilitate implementation without fundamentally changing the innovation. The net result was that awardees who had implemented similar programs at a similar scale were generally able to implement their innovations more rapidly than those with less experience. FLEs reported that experience was associated with better success in delivering innovation components at the intended level of intensity and frequency.

Multiple implications arise from the observation that prior experience is associated with improved implementation experiences. Foremost, perhaps, is that when rapid assessment of impact is expected, organizations with prior experience will likely be better able to implement and demonstrate effectiveness within the period of measurement. Organizations with less experience can be expected to face more unanticipated challenges, likely at every level of implementation—from enrollment to recruitment and staffing and training. On the other hand, if the prior experience was with the innovation being tested, as it was with many of the HCIA awardees, we would expect the DiD effect size for these awardees to be attenuated as some amount of the improvement expected will have occurred prior to baseline measurement. In theory, there is greater opportunity for improvement among awardees implementing a completely new innovation, but among awardees inexperienced with innovation observing that improvement can be expected to take longer to emerge.

Most innovations were implemented effectively, but few direct drivers of effective implementation were identified, and implementation effectiveness did not predict group differences in total cost of care.

To examine the award and innovation features associated with effective implementation we collected data using constructs identified in the literature as associated with implementation effectiveness. Few of these features, however, were found to be systematically associated with effective implementation in this sample. Our multivariate model indicated that implementing innovations in a single site was associated with greater implementation effectiveness, as was implementation planning (developing protocols, timelines, and, in particular, staffing plans) and the extensiveness of staff training.

Awardees also identified existing organizational capacity as an important determinant of rapid and successful health care transformation. Across multiple measures of organizational capacity (e.g., resources; having experienced staff and established partnerships; having a robust HIT, administrative, and technical-support infrastructure), awardees with existing capacity achieved greater success in rapidly implementing their innovations, saw full adoption of the

innovation by staff, and were able to deliver innovation components at the intended level of intensity and frequency and at the prescribed level of quality.

Similar to our discussion of prior experience, if rapid testing of the utility of an innovation for achieving outcomes is paramount, then testing innovations in well-resourced organizations may be warranted. However, results from these tests may not generalize to other, less well-resourced practices and organizations. Challenges of adoption and implementation may not be experienced by a well-resourced organization and when they are, may be less salient to staff or disruptive to implementation. This has implications for scalability that cannot be ignored. If innovations are to be scaled to less-resourced organizations, then they need to be tested in such practices to identify both impediments and facilitators of implementation success, and the period of measurement for detecting success may have to be extended to accommodate unanticipated implementation challenges.

Although we expected effective implementation to be a principal driver of better outcomes (decreased total cost of care, utilization), this hypothesis was not borne out in awardees' second year data. Implementation effectiveness showed small and non-significant relations with total cost of care, emergency department visits, and hospitalization, a result perhaps largely attributable to the overall effectiveness with which awards were implemented. However, many FLEs are reporting staff-perceived improvements in clinical care and workflow management as a result of these innovations, and, independent of their success in reducing costs or utilization, several awardees are planning to sustain all or part of their innovations on the basis of those staff impressions.

More awardees are providing data, but cost and utilization findings remain mixed and mostly null.

Since our last report, the number of awardees with evaluable data has increased by 19 awardees and post-intervention periods have lengthened. The precision of grand mean difference-in-difference estimates has also improved, with confidence intervals compressed below $100 per beneficiary per quarter for the TCOC among ambulatory and post-acute awardees. Because of substantive differences in setting and savings available, we deliberately segregated the largest group of awardees, the ambulatory innovations, from the post-acute care and hospital-setting awardees. Ambulatory innovations are qualitatively different from post-acute and hospital interventions and are associated with substantively lower cost estimates.

Our updated forest plots depict a now familiar pattern: in every setting, the results show a small group of awardees with favorable results (cost savings or utilization reductions), a similarly sized group with unfavorable results, and a majority with estimated effects that were in the vicinity of no effect. Two alternative analysis methods—time-series analysis and Bayesian—confirmed the general pattern of results. The dispersion of results for ambulatory innovations is consistent with a finding of no effect for both cost savings and emergency department visits, with the largest studies finding no effect and the dispersion of results occurring among awardees with the smallest and least reliable cost and utilization estimates. Nonetheless, differences among awardees remain greater than would be expected from sampling error, leaving the door open to identifying possible drivers of innovation success in improving cost and utilization outcomes.

While this finding seems discouraging, several awardees did show significant improvements in one or more outcomes, and the hospital-setting awardees achieved a statistically significant overall reduction of 6 ED visits per 1,000 beneficiaries. Further, these results must be balanced against FLE reports that many of the innovations improved health care delivery (workflow), had a positive impact on patients, and that many clinical staff felt empowered by opportunities offered by the innovation. Although FLEs did not report formal, quantitative assessments of these outcomes, including measures of workflow and staff and patient satisfaction in future evaluations may provide evidence of benefit against which cost and utilization outcomes can be balanced.

If, relative to comparators, innovations improve workflow, increase staff satisfaction and retention, and improve patient outcomes and patient satisfaction without increasing costs or utilization, then null and non-significant findings for cost and utilization might be interpreted quite differently. In particular, if robust measures show significant improvements in patient outcomes in the short term, then the lack of a significant increase in cost or utilization could be recognized as a success. Moreover, with robust evidence of improved health care delivery and improved patient health, econometric projections could be calculated of the reduced costs and utilization expected beyond the period of measurement.

While creating comparison groups was a considerable challenge for FLEs, the different approaches used do not appear to have influenced findings.

One potential source of bias with which we were concerned was the different methods FLEs used to create comparison groups. Constructing these groups was one of the biggest challenges faced by FLEs and five major approaches were adopted. Most FLEs estimated propensity scores, which they used either for matching or weighting. We found no evidence to date that the method of comparison group creation introduced any systematic bias in HCIA effect estimates.

While propensity scores provide samples matched on observable patient characteristics, only the primary care and community FLEs used matching strategies at an aggregate level (e.g., matching at the practice or facility level), and for only a subset (eight total) of their awardees. For the remainder, and perhaps for these as well, other than that they received "usual care," little is known about the characteristics of care received by comparators. In the dynamic world of health care delivery, where many practices are participating in both formal and informal quality improvement initiatives, the impact of innovations cannot be reliably established without knowing what services were received by the counterfactual sample. For example, if both an awardee and their comparator are undertaking similar or equally effective innovations we would expect the DiD score to show no difference, even if the innovation was effective in reducing cost and utilization. Without knowing the care received by enrollees acting as comparators, we are, in many respects, reducing our confidence that the results obtained are attributable to the effectiveness of the innovative practices in reducing costs and utilization. When contrasting results across multiple innovations or implementations, having information on comparison groups' experience of care or service receipt can be a critical confound that can be statistically controlled for using meta-regression.

Preliminary meta-regression results successfully identified some features associated with success, confirming several qualitative themes.

For this report, we performed meta-regression analyses for the first time. These analyses were supported by the results of heterogeneity analyses indicating that most of the variation in effect sizes was not the simply the product of sampling error. We took a conservative approach to including awardees in the meta-analyses, omitting cases in which the core outcome measures were based on unusual subgroups or unusual time frames and situations in which FLEs expressed doubts about the accuracy of their DiD estimates. After accounting for other forms of ineligibility, the number of ambulatory innovations is effectively capped at 43 interventions, a number unlikely to rise much in the remaining project time.

We split our meta-regressions of TCOC into three different clusters of innovation features: structural features, intervention components, and implementation features. Structural features of the innovations explained very little of the variation in cost effects; the only notable impact was found for innovations run by for-profit organizations. It may be that for-profit organizations are better resourced or have more experience with innovation, but these relations are yet to be tested. Three intervention components—health informatics, telemedicine, and community health workers—had TCOC effects exceeding $150 per beneficiary per quarter. Finally, the meta-regression based on implementation features highlighted the deleterious impact on savings of new innovations and staff turnover.

Many of the features and components that emerged in meta-regression were also identified in the thematic and implementation analyses as central to implementation success and are focal issues with which awardees wrestled and on which FLEs reported. This triangulation of the qualitative findings with the preliminary quantitative results provides confidence that awardees and FLEs are focusing on the issues and challenges which are most likely to directly impact effectiveness.

Standard errors for many of the binary innovation features in our meta-regressions were around $100 per beneficiary per quarter, indicating that models were capable of detecting TCOC effects on the order of $200 per beneficiary per quarter. For the current meta-regression models, we estimate that adding five additional innovations to the data (an increase from 39 to 44 innovations for our regression of implementation features) would decrease standard errors from $100 to $87 and increase our ability to detect effective drivers of effectiveness. This, combined with the increased ability to control for multiple confounds while assessing drivers of effectiveness, highlights the value of increasing the number of cases for meta-regression.

It can also be noted that there was considerable dispersion across awardees in the inverse variance weights used in meta-analysis. These weights are largely a function of sample size (i.e., the number of awardees included in the DiD estimate). Regression results tend to be very sensitive to these weights, which we capped to avoid giving undue influence to very large programs. It is also apparent that coefficient precision is sensitive to the prevalence of a characteristic and to correlations among the explanatory variables. All of these factors need to be monitored when conducting meta-regression.

A meta-regression path model confirmed the impact of ED visits and hospital admissions on TCOC.

We constructed a meta-regression path model to link elements of our previous analyses into a single model covering three of HCIA's core outcomes. The results demonstrated that utilization is the main driver of TCOC effects among these interventions, and that hospital admissions have a greater impact on costs than ED visits. New innovations had unfavorable impacts on both forms of care utilization and were more difficult to implement effectively. Other innovation features were largely unimportant for HCIA effects.

Although the results tested in this model are preliminary and the model is only an early test of the analytic approach, in many respects the limited contribution of structural features, intervention components, and implementation features to utilization and TCOC is counterintuitive. There is support in the literature for many of these innovations and for the importance of implementation characteristics in achieving effectiveness. Several of the reasons this model may not be detecting those relations have been discussed above.

Two factors that have not been introduced are what researchers refer to as the attributable fraction (in this case, the proportional improvement in outcome that would be expected from these innovations and their optimal implementation) and the "signal-to-noise-ratio." Summary measures of utilization and TCOC result from multiple factors, only a portion of which (the attributable fraction) might be affected by these innovations. Similarly, to the extent these multiple exogenous factors create noise in the DiD estimates, our ability to detect the signal of other innovation measures (their measured contribution to the effect size) is reduced. To the extent other, perhaps principal, drivers of these outcomes are uncontrolled in the analysis, the ability of meta-regression to detect the role of innovations and innovation features in achieving success on these outcomes may be obscured. If data on these other cost drivers were available, their influence could be statistically controlled using meta-regression and the impact of innovations and implementation features better estimated.

There was a strong correlation between an intervention's effect on hospital admissions and its effect on ED use that was not explained in the path model. These two effect sizes might be causally related to each other, so that preventing ED use also prevents hospitalizations. Other forms of care utilization (physician visits, outpatient visits) also contribute to fee-for-service costs. There are other as yet unexamined innovation features that may also contribute to the model. This initial model serves as a prelude to the more expansive multivariate meta-analyses we will be conducting in the final year of the HCIA project.

Summary

While the first year of awardees' implementation experience could be characterized by both anticipated and unanticipated challenges, the second year saw awardees adapting their innovations to meet and mitigate these challenges. By the end of the second year most innovations were well implemented, and both qualitative and quantitative analyses highlighted several features associated with successful implementation. Prior experience and organizational capacity in particular were associated with greater planning for adoption, the ability to navigate challenges, and rapid adoption of the innovation. There remains a tradeoff between rapid

demonstration of impact in high-resource awardees and generalizability of the innovation in lower-resource settings.

The second year also saw an increase in the awardees providing DiD data and, as in the first year, the results are mixed with most awards showing no innovation effect on cost or utilization. However, FLEs report many awardees experienced improvements in workflow, staff satisfaction, and staff-perceived improvements in patient health and satisfaction. Although formal (quantitative) data are lacking for these outcomes, innovations that are improving health and the experience of delivering health care while maintaining costs might also be considered successful.

The different approaches FLEs used to create comparison groups do not appear to have systematically biased findings, but controlling for patient differences does little to control for services these enrollees received. Modern health care may be characterized by many formal, informal, and local quality improvement initiatives. Knowing what services patients receive and the activities surrounding the delivery of these services might allow better estimation of the innovation effect, especially when comparing results across multiple innovations.

The role of several of the themes identified in the qualitative analysis emerged in the meta-regression as central to reducing second year TCOC estimates among ambulatory patients. This suggests that awardees and FLEs are focusing attention on the issues and challenges that impact implementation effectiveness and are expected to reduce TCOC and utilization. Analysis of the means by which awardees reduced TCOC identified relatively few features associated with success, but did confirm that reduced ED visits and hospital admissions are associated with reduced TCOC. It is possible that other large drivers of TCOC obscure the impact of innovations and implementation features. Measuring and controlling for cost-drivers beyond the innovation being tested would likely improve the ability of meta-regression to detect determinants of favorable outcomes when they are present.

Much of this year's evaluation effort was focused on distilling FLEs' second year reports and data to assemble rich and accurate descriptions of the awardees innovations, their experiences in implementing their evaluations, and obtaining consistent measures of awardee performance on the four core measures. We also began investigating the relations of awardee characteristics and experiences with measures of success in implementing their innovations and performance on the four core measures. In the final year of our evaluation, we will investigate, in greater depth and detail, the interrelations among awardee features and characteristics and how these attributes impact awardee success in reducing health care utilization and TCOC.

[This page intentionally left blank]

REFERENCES

Bonett, D. G., & Price, R. M. (2015). Varying coefficient meta-analysis methods for odds ratios and risk ratios. *Psychological Methods, 20*(3), 394–406. doi: 10.1037/met0000032.

Damschroder, L. J., Aron, D. C., Keith, R. E., Kirsh, S. R., Alexander, J. A., & Lowery, J. C. (2009). Fostering implementation of health services research findings into practice: A consolidated framework for advancing implementation science. *Implementation Science, 4*, 50.

Gavaghan, D. J., Moore, R. A., & McQuay, H. J. (2000). An evaluation of homogeneity tests in meta-analyses in pain using simulations of individual patient data. *Pain, 85*(3):415–424.

Hedges, L. V., & Olkin, I. (1985). Statistical methods for meta-analysis. New York: Academic Press.

Higgins, J. P. T., & Green, S. (Eds.). (2011, March). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0. The Cochrane Collaboration. Available from www.cochrane-handbook.org

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*(11), 1539–1558. doi: 10.1002/sim.1186

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ, 327*(7414), 557–560. doi:10.1136/bmj.327.7414.557

Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods, 6*, 53–60.

Kenny, D. A. (2015). Measuring Model Fit. Retrieved from http://davidakenny.net/cm/fit.htm

Kline, R. B. (2011). Principles and practice of structural equation modeling. Third edition. New York: Guilford Press.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*, 130–149.

Nilsen, P. (2015). Making sense of implementation theories, models and frameworks. *Implementation Science, 10,* 53. doi:10.1186/s13012-015-0242-0

Ragin, C. C. (2000). Fuzzy-set social science. Chicago: University of Chicago Press.

Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41–55.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine, 127,* 757–763.

Schneider, C. Q, & Wagemann, C. (2012). Set-theoretic methods for the social sciences: A guide to qualitative comparative analysis. Cambridge: Cambridge University Press.

Thompson, S. G. & Higgins, J. P. T. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine, 21,* 1559–1573.

Welch, B. L. (1951), On the comparison of several mean values: an alternative approach. *Biometrika, 38,* 330–336.

APPENDIX A:
AASF2 DEVELOPMENT AND RESULTS

A.1 Background

The goals of mixed methods evaluation, in which both qualitative and quantitative information are gathered, are to use the data provided by each method to better inform the findings of the evaluation, and also to provide insights that improve the type or quality of information gathered by each method. The genesis of the Annual Awardee Summary Form (AASF) arose out of this second purpose. In this section we describe the rationale and development of the second AASF.

For our first annual report, qualitative data were abstracted from FLE reports on each HCIA awardee. These data were then rigorously coded and analyzed to identify commonly reported issues affecting innovation implementation. These findings were summarized in our first annual report and formed the basis for 15 lessons learned briefs. While the qualitative data identified many of the issues faced by awardees in implementing their innovations, the FLE reports did not provide consistent evidence of the scope or intensity of those concerns across all awardees. The templates FLEs follow in producing their quarterly and annual reports provide guidance on types of information that should be reported, but allow each FLE considerable leeway in how those findings are reported. In many respects this is appropriate, as it allows each FLE latitude to present those findings most relevant to each awardee's experience in implementing their HCIA award, but when summarizing across awardees, the variation in reporting produced data gaps and may reduce the generalizability of findings.

For example, in their reports several FLEs discussed the challenges associated with implementing health information technologies (HIT), but if implementing those technologies was not a challenge, the lack of a challenge was unlikely to be mentioned. Thus, while the qualitative coding highlighted HIT as an implementation challenge for several awardees, because of the missing data we could not assess the extent to which it was a challenge for all awardees implementing HIT. As another example, an awardee for whom staff training was highly structured and ongoing might receive several paragraphs describing that training, while another awardee, for whom training was just as extensive but which was limited to shadowing an experienced clinician, might only receive a sentence mentioning that training experience.

A.2 Method

To get a more systematic and comprehensive accounting of the various issues and challenges identified through the qualitative document review, in September 2015 a second AASF was developed and administered to FLEs who had conducted site-visits and who were familiar with awardees' implementation activities. Based primarily on the findings from the qualitative review, and informed by the Consolidated Framework for Implementation Research (CFIR; Damschroder et al., 2009), a draft questionnaire was developed and distributed to FLEs for comment. Based on input from the FLEs, obtained in a 1-hour conference call and by e-mail, item stems were clarified, some items were removed and a few added, and additional text introducing each domain was developed. The refined draft AASF survey was distributed to the FLEs and a formal training session was held. FLEs expressed support for the refined version of

the AASF and the survey was finalized and electronically coded for Web administration. FLEs were given 6 weeks to complete the survey and were encouraged to consult with other site visit members in completing the survey. To answer respondent questions or concerns, an additional information session was held during the survey administration period where respondents could call the meta-evaluation team for clarification. No calls were received during that session.

The 2015 AASF consisted of approximately 54 items across 9 domains assessing innovation complexity, planning, process, staff training, organizational leadership, implementation and HIT challenges, and both implementation and innovation effectiveness. Respondents rated each awardee's experience with their innovation using a four-level Likert scale. Response option for most items were "Not at all," To a slight extent," To a moderate extent," and "To a great extent," while the response options for HIT and implementation challenges were rated "Not a challenge," "Small challenge," Modest challenge," or "Major challenge."

Several awardees implemented multiple interventions, which they evaluated separately as part of their award. To accommodate these awardees, we distributed separate AASF surveys for each intervention they implemented and for which they reported outcome data. Thus, FLEs completed 123 surveys for the 108 HCIA awardees.

A.3 Findings

A.3.1 Missing Data Patterns

Two valid missing options were provided. Respondents could select "Not applicable" if an item was not relevant to the innovation being coded (e.g., if HIT was not an innovation component or if the innovation did not involve partners), and they could select "Unable to assess" if they had insufficient information about an awardee to make an assessment. Approximately 58 percent of respondents selected "not applicable" 1 to 5 times throughout the survey, 26 percent of respondents selected it 6 to 12 times, and 5 respondents (4 percent) selected "not applicable" for 15 to 23 of the 54 items. By domain, the "not applicable" responses ranged from 0.6 (assessment planning) to 24.8 HIT challenges (see ***Table A-1***).

Table A-1
Percent Selecting "Not Applicable"

	Complexity	Planning	Process	Training	Leadership	Effectiveness	Challenges	HIT Challenges
Behavioral	0%	0%	2%	0%	2%	4%	7%	23%
Community	2%	0%	2%	0%	1%	3%	8%	46%
Complex	4%	1%	1%	9%	3%	2%	11%	26%
Disease	4%	0%	5%	4%	2%	9%	8%	20%
Hospital	1%	3%	5%	0%	3%	6%	23%	6%
MMSDM	6%	0%	10%	8%	10%	6%	24%	31%
Primary	4%	0%	9%	0%	10%	9%	12%	21%
Overall Average	2.9%	0.6%	4.9%	3.0%	4.5%	5.5%	13.3%	24.8%

"Unable to assess" was selected somewhat more frequently, but 21 percent of respondents did not select "unable to assess" for any of the 54 items. Approximately 53 percent of respondents responded "unable to assess" 1 to 5 times throughout the survey, and 21 percent of respondents responded 6 to 13 times. Five respondents (4 percent) responded "unable to assess" in 18 to 28 of the 54 items. By a considerable margin, "unable to assess" was selected most frequently for items assessing the role and contribution of organizational leadership (range 5–31 percent, overall average 16.3 percent). ***Table A-2*** presents "unable to assess" rates by FLE.

Table A-2
Percent Selecting "Unable to Assess"

	Complexity	Planning	Process	Training	Leadership	Effectiveness	Challenges	HIT Challenges
Behavioral	2%	0%	12%	6%	5%	6%	1%	0%
Community	1%	5%	4%	4%	8%	3%	3%	2%
Complex	1%	3%	4%	1%	27%	8%	2%	5%
Disease	2%	19%	14%	7%	24%	0%	3%	2%
Hospital	5%	3%	4%	10%	31%	11%	6%	5%
MMSDM	3%	7%	0%	7%	10%	10%	2%	4%
Primary	1%	1%	0%	3%	8%	2%	1%	2%
Overall Average	2.1%	5.3%	5.4%	5.3%	16.3%	5.7%	2.4%	2.8%

Caution is recommended when comparing missing data patterns across FLEs and across domains. Because each FLE has a different number of awardees and domains have different numbers of items, the impact of missing data changes as the denominator (the number of awardees times the number of items) changes by domain and FLE. For example, each missing value for a FLE with few awardees (e.g., MMSDM) comprises a greater percentage of the total. If a domain contains four items, each missing MMSDM response increases their missing percent by 2.7 percent. Compare this to Community, a FLE with 26 awardees, where each missing response increases their missing percent by 0.96 percent.

A.3.2 AASF Descriptive results

Table A-3
Innovation Complexity

	Varied by site	To a great extent	To a moderate extent	To a slight extent	Not at all	Not Applicable	Unable to assess
Innovation Complexity							
Involved multiple interconnecting components	(1.6%)	36.6%	26.8%	26.8%	8.9%	0.8%	0.0%
Spanned multiple care settings (e.g., hospital, outpatient, home)	(4.1%)	29.3%	26.0%	17.9%	21.1%	3.3%	2.4%
Involved staff from various groups, departments, or organizational levels providing care or services	(2.4%)	31.7%	27.6%	21.1%	13.8%	4.9%	0.8%
Required formal agreements among organizations	(1.6%)	45.5%	17.1%	13.0%	13.0%	4.1%	7.3%
Required cooperation from distributed independent providers	(3.3%)	35.8%	12.2%	13.8%	28.5%	7.3%	2.4%
Added or significantly changed steps in the service delivery workflow	(5.7%)	30.1%	34.1%	21.1%	8.9%	3.3%	2.4%
Required new health information technology	(8.1%)	28.5%	23.6%	23.6%	20.3%	3.3%	0.8%
Required hiring clinical staff new to the organization	(8.9%)	26.0%	22.0%	22.0%	23.6%	2.4%	4.1%
Required hiring technical, research, or administrative staff new to the organization	(8.9%)	13.0%	32.5%	35.0%	13.8%	1.6%	4.1%
Required changes to existing staffs' roles and responsibilities	(8.9%)	21.1%	34.1%	26.8%	13.0%	1.6%	3.3%
Required training staff for new or additional skills	(3.3%)	47.2%	33.3%	14.6%	4.1%	0.0%	0.8%
Was explicitly developed and designed with intent for future dissemination to other sites	(1.6%)	43.1%	23.6%	11.4%	13.8%	6.5%	1.6%

Table A-4
Implementation Planning

	Varied by site	To a great extent	To a moderate extent	To a slight extent	Not at all	Not Applicable	Unable to assess
Implementation Planning							
Written and available protocols and procedures for innovation delivery	(1.6%)	42.3%	30.9%	18.7%	2.4%	0.0%	5.7%
A detailed timeline with milestones	(0.8%)	39.0%	30.9%	11.4%	2.4%	0.8%	15.4%
A comprehensive staffing plan	(1.6%)	46.3%	30.9%	13.8%	2.4%	0.8%	5.7%
Experience with implementing similar programs at a similar scale	(4.9%)	22.0%	29.3%	22.0%	19.5%	0.0%	7.3%

Table A-5
Implementation Process

	Varied by site	To a great extent	To a moderate extent	To a slight extent	Not at all	Not Applicable	Unable to assess
Implementation Process							
Introduce innovation components in a planned and deliberate way	(1.6%)	58.5%	30.1%	7.3%	0.8%	1.6%	1.6%
Work with all necessary entities within the awardee organization to implement the innovation	(0.8%)	65.9%	27.6%	4.1%	0.8%	1.6%	0.0%
Work with all necessary entities across organizations to implement the innovation	(1.6%)	47.2%	32.5%	5.7%	0.0%	13.8%	0.8%
Execute its self-monitoring plan	(0.8%)	39.0%	30.1%	9.8%	0.8%	0.8%	19.5%
Use a formal improvement framework or change management process (e.g., LEAN, PDSA cycles)	(0.8%)	9.8%	9.8%	16.3%	30.9%	4.1%	29.3%

Table A-6
Staff Training

	Varied by site	To a great extent	To a moderate extent	To a slight extent	Not at all	Not Applicable	Unable to assess
Staff Training							
Front-line staff participate in formal training to support innovation delivery	(4.1%)	56.1%	25.2%	13.0%	1.6%	1.6%	2.4%
Front-line staff receive training through experiential learning (e.g., shadowing or mentoring)	(6.5%)	41.5%	26.8%	13.8%	5.7%	3.3%	8.9%
Front-line staff receive ongoing training throughout the duration of the innovation award	(6.5%)	35.8%	20.3%	27.6%	7.3%	3.3%	5.7%
New (and rotating) staff receive training to accommodate staff turn-over	(6.5%)	39.8%	23.6%	15.4%	3.3%	6.5%	11.4%

Table A-7
Organizational Leadership

	Varied by site	To a great extent	To a moderate extent	To a slight extent	Not at all	Not Applicable	Unable to assess
Organizational leadership							
Attend meetings related to innovation	(5.7)	33.3%	22.0%	15.4%	4.1%	0.8%	24.4%
Act as a liaison to external partners	(5.7)	30.9%	17.9%	11.4%	6.5%	10.6%	22.8%
Closely monitor implementation progress	(6.5)	29.3%	32.5%	13.0%	5.7%	0.0%	19.5%
Proactively resolve problems in response to feedback from staff	(4.9)	18.7%	18.7%	7.3%	12.2%	5.7%	37.4%
Provide in-kind staffing to support the innovation	(4.9)	17.9%	22.0%	20.3%	17.9%	3.3%	18.7%
Provide in-kind resources other than staffing to support the innovation	(7.3)	16.3%	25.2%	17.1%	14.6%	4.9%	22.0%

Table A-8
Implementation Challenges

	Varied by site	Major challenge	Modest challenge	Small challenge	Not a challenge	Not Applicable	Unable to assess
Implementation Challenges							
Enrolling patient participants	(4.9%)	19.5%	29.3%	17.1%	22.8%	10.6%	0.8%
Implementing and/or executing agreements with partners	(0.8%)	19.5%	14.6%	13.8%	30.9%	13.0%	8.1%
Clinician buy-in and engagement	(8.1%)	13.8%	22.8%	34.1%	18.7%	5.7%	4.9%
Staff recruitment	(2.4%)	8.9%	19.5%	30.9%	34.1%	4.9%	1.6%
Management staff turnover or unfilled management roles	(2.4%)	1.6%	7.3%	17.1%	69.9%	3.3%	0.8%
Frontline staff turnover or unfilled frontline staff roles	(4.1%)	9.8%	22.8%	35.0%	29.3%	0.8%	2.4%
Integration of non-licensed staff (e.g., community health workers) into care team	(4.1%)	7.3%	13.8%	8.1%	26.8%	43.1%	0.8%
Competing initiatives or programs	(4.9%)	7.3%	13.8%	16.3%	51.2%	6.5%	4.9%
Level of reimbursement for services	(2.4%)	21.1%	17.1%	10.6%	28.5%	16.3%	6.5%
Changes in federal or state policies, legislation, or regulation affecting implementation	(1.6%)	8.9%	13.0%	14.6%	48.0%	11.4%	4.1%
Workflow redesign	(4.9%)	5.7%	31.7%	26.0%	25.2%	8.9%	2.4%

Table A-9
Health Information Technology Challenges

	Varied by site	Major challenge	Modest challenge	Small challenge	Not a challenge	Not Applicable	Unable to assess
Health Information Technology Challenges							
Selecting or designing HIT to support the innovation	(4.1%)	8.9%	21.1%	22.8%	23.6%	22.0%	1.6%
Building out or installing HIT to support the innovation	(3.3%)	9.8%	27.6%	20.3%	18.7%	21.1%	2.4%
Identifying, hiring, or obtaining vendor support for innovation HIT requirements	(4.1%)	4.9%	10.6%	12.2%	32.5%	30.9%	8.9%
Data standardization across systems	(2.4%)	14.6%	18.7%	15.4%	18.7%	27.6%	4.9%
Interoperability across organizations	(4.9%)	12.2%	18.7%	13.0%	16.3%	38.2%	1.6%
Alignment of HIT with clinical workflow	(4.9%)	8.9%	20.3%	26.8%	17.9%	22.8%	3.3%
Acceptability of the HIT by front line staff	(5.7%)	6.5%	19.5%	24.4%	24.4%	20.3%	4.9%

Table A-10
Implementation Effectiveness

	Varied by site	To a great extent	To a moderate extent	To a slight extent	Not at all	Not Applicable	Unable to assess
Implementation Effectiveness							
Full adoption of innovation by front-line staff	(8.1%)	66.7%	22.8%	4.9%	0.0%	1.6%	4.1%
Full adoption of innovation by external partners	(15.4%)	23.6%	27.6%	9.8%	4.1%	27.6%	7.3%
Rapid adoption of the innovation	(14.6%)	21.1%	45.5%	21.1%	5.7%	0.0%	6.5%
Completion of all tasks needed for full innovation implementation	(9.8%)	56.1%	28.5%	8.1%	2.4%	0.0%	4.9%
Innovation components being delivered as intended and at the prescribed level of quality	(12.2%)	51.2%	32.5%	6.5%	0.0%	0.0%	9.8%
Innovation components being delivered at the intended level of intensity and frequency	(12.2%)	46.3%	31.7%	10.6%	1.6%	0.0%	9.8%

A.4 Summary

In general, and in contrast with the first AASF, site-level FLEs who responded to the second AASF were able to provide ratings for most items, and most items showed moderate variation in response. Respondents had the greatest difficulty providing ratings in the "Organizational Leadership" domain, but were still able to provide ratings in this domain for three-quarters of the interventions funded by CMMI.

Results from the second AASF indicate that most innovations were successfully implemented with generally positive ratings in all implementation domains, and that, while many interventions experienced challenges in implementing their intervention, most challenges were judged to be small or modest challenges. Even HIT, a prominent challenge in our first annual

report, was not a considered by FLEs to be a major challenge for many awardees. Implementation planning was carried out in a structured and organized fashion, with sufficient attention to training for a strong majority of interventions. Undoubtedly, this attention to the mechanics of implementation led to the success awardees experienced in implementing these often complex innovations.

APPENDIX B-1:
COPY OF AASF2 INSTRUMENT

2015 Annual Awardee Summary Form Instructions

The HCIA meta-evaluator will use Frontline Evaluator (FLE) responses on this form to synthesize findings across awardees. Read instructions carefully, consult the "Instructions for Annual Awardee Summary Form 2015" provided to all FLEs prior to completing this form. The awardee assessment provided on this form will not be shared with awardees. Items marked with an "*" are required.

Contact Asha Ayub (aayub@rti.org, (781) 434-1787)) for any questions.

1. Front Line Evaluator Name and HCIA Portfolio

[prepopulated with FLE name]

2. Awardee Name

[prepopulated with Awardee name]

☐ Check this box to confirm this is the Awardee for whom you are reporting on.

If this is not the correct awardee, please check that you have selected the correct link and try again. If you are certain that you have selected the correct link, please contact Asha Ayub for further assistance.

Implementation Sites

The following items ask you to characterize the number of implementation sites used by the awardee.
An implementation site is defined as the organizational unit where innovation components, care, or services are being implemented and monitored by an awardee or an awardee's partners. *Sites serving as comparison or usual treatment comparison sites are not considered implementation sites.*

A site may be at the level of a hospital unit, clinic or practice, hospital system, organization, geographic unit (e.g., county), or other unit defined by administrative/management boundaries. In other words, the implementation site is the organizational unit which tracks patients through their care experience. An awardee with multiple sites may coordinate data collection from independent sites; each site is responsible for implementing the innovation locally, delivering care or services, and providing data on that care experience. **For example,** one awardee disseminated a model for a specific care program to other hospitals. Each hospital is *independently* implementing the model. In this case, each hospital represents an implementation site.

A single implementation site may span multiple care settings. **For example,** a transitional care coordination program implemented by an awardee at one hospital, may involve services at hospital discharge, follow-up in an outpatient setting, and follow-up home visits.

3. Does the innovation involve more than one implementation site? (select one)

☐ Yes
☐ No
☐ Unable to determine

[Yes→skip to check box question on next page.]
Comments_____

4. How many implementation sites does this awardee have? (Enter a numeric value)

_____
Enter "0" if you are unable to determine the precise number of sites and use the comment box below to describe the range or approximate number, for example "less than 5", "not more than 10", etc.
Comment_____

5a. How many implementations sites did you collect implementation experience data from through in-person site visits, telephone interviews, or direct observation?

_____

5b. If you did not visit or collect implementation experience data from all sites within an awardee, how did you select which sites to visit or collect data from? For example, were sites selected based on geography, size, performance characteristics, representativeness, non-representativeness, etc.

_____

For the next items, management team refers to the Innovation Project Director or Principal Investigator and his or her team responsible for oversight of innovation implementation, which may or may not be the same as the patient care team responsible for providing direct care or services associated with the innovation.

6. What model of oversight of the implementation process best describes this awardee? (select one)

- ☐ The management team *at the awardee organization* oversees implementation activities.
- ☐ A management team *at each participating site* oversees implementation activities.
- ☐ Other (please describe) _____

7. Which best describes the management relationship between the awardee organization and implementing sites? (select one)

- ☐ All or most participating sites are under the management of the awardee organization (e.g., different practices that are part of the same health care system).
- ☐ All or most participating sites are external partners to the awardee organization.
- ☐ Other (please describe) _____

8. Please indicate the reporting unit for impact outcomes for this awardee (select all that apply)

Impact outcomes include the core four measures (hospital admission, readmissions, ED visits, total cost of care), along with any other awardee-specific outcomes defined for this awardee.

- ☐ Outcomes are or will be reported at the Awardee level (i.e., data from multiple sites will be aggregated)
- ☐ Outcomes are or will or can be reported at the site level. For this item, please do not consider whether sufficient power exists to estimate impact by site, just whether outcomes can be provided separately by site.
- ☐ Other, please describe_____

On the following pages, please answer the items using information collected through document review, interviews or field observations, surveys, or other primary data collection that you conducted as part of your evaluation.

Read the "Instructions for Annual Awardee Summary Form 2015" for further description of the item questions, response options and for how to complete the item if your response varies by implementation site (i.e., your rating would be "to a great extent" at most sites, but "not at all" at a few sites).

☐ Click here to continue to the next page

9. Innovation Complexity

The awardee's innovation:

	To a great extent	To a moderate extent	To a slight extent	Not at all	Not Applicable	Unable to Assess	Varied by Site
Involved multiple interconnecting components	☐	☐	☐	☐	☐	☐	☐
Spanned multiple care settings (e.g., hospital, outpatient, home)	☐	☐	☐	☐	☐	☐	☐
Involved staff from various groups, departments, or organizational levels providing care or services	☐	☐	☐	☐	☐	☐	☐
Required formal agreements among organizations	☐	☐	☐	☐	☐	☐	☐
Required cooperation from distributed independent providers	☐	☐	☐	☐	☐	☐	☐
Added or significantly changed steps in the service delivery workflow	☐	☐	☐	☐	☐	☐	☐
Required new health information technology	☐	☐	☐	☐	☐	☐	☐
Required hiring clinical staff new to the organization	☐	☐	☐	☐	☐	☐	☐

	To a great extent	To a moderate extent	To a slight extent	Not at all	Not Applicable	Unable to Assess	Varied by Site
Required hiring technical, research, or administrative staff new to the organization	☐	☐	☐	☐	☐	☐	☐
Required changes to existing staffs' roles and responsibilities	☐	☐	☐	☐	☐	☐	☐
Required training staff for new or additional skills	☐	☐	☐	☐	☐	☐	☐
Was explicitly developed and designed with intent for future dissemination to other sites	☐	☐	☐	☐	☐	☐	☐

Comments for this awardee related to any items above _____

10. Implementation Planning

The management team refers to the Innovation Project Director or Principal Investigator and his or her team responsible for oversight of innovation implementation, which may or may not be the same as the patient care team responsible for providing direct care or services associated with the innovation.

To what extent did the management team have:

	To a great extent	To a moderate extent	To a slight extent	Not at all	Not Applicable	Unable to assess	Varied by Site
Written and available protocols and procedures for innovation delivery	☐	☐	☐	☐	☐	☐	☐
A detailed timeline with milestones	☐	☐	☐	☐	☐	☐	☐
A comprehensive staffing plan	☐	☐	☐	☐	☐	☐	☐
Experience with implementing similar programs at a similar scale	☐	☐	☐	☐	☐	☐	☐

Comments for this awardee related to any items above _____

11. Implementation Process

The management team refers to the Innovation Project Director or Principal Investigator and his or her team responsible for oversight of innovation implementation, which may or may not be the same as the patient care team responsible for providing direct care or services associated with the innovation.

To what extent did the management team:

	To a great extent	To a moderate extent	To a slight extent	Not at all	Not Applicable	Unable to assess	Varied by site
Introduce innovation components in a planned and deliberate way	☐	☐	☐	☐	☐	☐	☐
Work with all necessary entities *within* the awardee organization to implement the innovation	☐	☐	☐	☐	☐	☐	☐
Work with all necessary entities *across* organizations to implement the innovation	☐	☐	☐	☐	☐	☐	☐
Execute its self-monitoring plan	☐	☐	☐	☐	☐	☐	☐
Use a formal improvement framework or change management process (e.g., LEAN, PDSA cycles)	☐	☐	☐	☐	☐	☐	☐

Comments for this awardee related to any items above _____

12. Staff Training.

The following items ask about staff training in support of innovation delivery. The staff to consider for responding to this item are awardee or partner staff that had a role in providing or supporting the care or services required to implement and sustain the innovation.

To what extent did:

	To a great extent	To a moderate extent	To a slight extent	Not at all	Not Applicable	Unable to assess	Varied by site
Front-line staff participate in formal training to support innovation delivery	☐	☐	☐	☐	☐	☐	☐
Front-line staff receive training through experiential learning (e.g., shadowing or mentoring)	☐	☐	☐	☐	☐	☐	☐
Front-line staff receive ongoing training throughout the duration of the innovation award	☐	☐	☐	☐	☐	☐	☐
New (and rotating) staff receive training to accommodate staff turn-over	☐	☐	☐	☐	☐	☐	☐

Comments for this awardee related to any items above _____

13. Organizational Leadership.

Leaders include the person(s) to whom the Innovation Project Director/Principal Investigator reports and are individuals with the power to make resource allocation decisions within the organization. This may include the organization CEO or other key senior leaders.

For these items, organization leaders are NOT the PD/PI of the HCIA award or innovation team staff, unless the PI/PD is, in fact, a senior leader in the organization and can make resource allocation decisions within the organization.

To what extent did organization leaders:

	To a great extent	To a moderate extent	To a slight extent	Not at all	Not Applicable	Unable to assess	Varied by Site
Attend meetings related to innovation	☐	☐	☐	☐	☐	☐	☐
Act as a liaison to external partners	☐	☐	☐	☐	☐	☐	☐
Closely monitor implementation progress	☐	☐	☐	☐	☐	☐	☐
Proactively resolve problems in response to feedback from staff	☐	☐	☐	☐	☐	☐	☐
Provide in-kind staffing to support the innovation	☐	☐	☐	☐	☐	☐	☐
Provide in-kind resources other than staffing to support the innovation	☐	☐	☐	☐	☐	☐	☐

Comments for this awardee related to any items above _____

14. Implementation Effectiveness.

The next items ask you to rate how successful the awardee has been in implementing the innovation, which is defined as the extent to which the planned innovation care or services were consistently delivered to the intended target population at the intended level of quality and intensity. *Innovation effectiveness* (e.g., impact outcomes) will be captured in a later section.

To what extent did implementation result in:

	To a great extent	To a moderate extent	To a slight extent	Not at all	Not Applicable	Unable to assess	Varied by Site
Full adoption of innovation by front-line staff	☐	☐	☐	☐	☐	☐	☐
Full adoption of innovation by external partners	☐	☐	☐	☐	☐	☐	☐
Rapid adoption of the innovation	☐	☐	☐	☐	☐	☐	☐
Completion of all tasks needed for full innovation implementation	☐	☐	☐	☐	☐	☐	☐

Comments for this awardee related to any items above _____

To what extent are:

	To a great extent	To a moderate extent	To a slight extent	Not at all	Not Applicable	Unable to assess	Varied by site
Innovation components being delivered as intended and at the prescribed level of quality	☐	☐	☐	☐	☐	☐	☐
Innovation components being delivered at the intended level of intensity and frequency	☐	☐	☐	☐	☐	☐	☐

Comments for this awardee related to any items above _____

16. Challenges.

This item asks about challenges awardees may have faced in implementing or maintaining their innovations. This item asks about non-Health Information Technology (HIT) challenges and the awardee's success in overcoming those challenges. HIT challenges will be assessed in the next item.

Rate the extent of each of the challenges below:

	Major Challenge	Moderate Challenge	Small Challenge	Not a Challenge	Not Applicable	Unable to assess	Varied by site
Enrolling patient participants	☐	☐	☐	☐	☐	☐	☐
Implementing and/or executing agreements with partners	☐	☐	☐	☐	☐	☐	☐
Clinician buy-in and engagement	☐	☐	☐	☐	☐	☐	☐
Staff recruitment	☐	☐	☐	☐	☐	☐	☐
Management staff turnover or unfilled management roles	☐	☐	☐	☐	☐	☐	☐
Frontline staff turnover or unfilled frontline staff roles	☐	☐	☐	☐	☐	☐	☐
Integration of non-licensed staff (e.g., community health workers) into care team	☐	☐	☐	☐	☐	☐	☐
Competing initiatives or programs	☐	☐	☐	☐	☐	☐	☐
Level of reimbursement for services	☐	☐	☐	☐	☐	☐	☐
Changes in federal or state policies, legislation, or regulation affecting implementation	☐	☐	☐	☐	☐	☐	☐
Workflow redesign	☐	☐	☐	☐	☐	☐	☐

Please comment further on the challenges and on the extent to which the awardee was able to overcome each of the challenges you have marked as major or moderate above. _____

18. Health Information Technology (HIT) Challenges.

This item asks about HIT challenges awardees may have faced in implementing or maintaining their innovations and success in overcoming these challenges. If the awardee's innovation does not involve any HIT components, select not applicable for each item.

Rate the extent of each of the health information technology (HIT) challenges below.

	Major challenge	Modest challenge	Small challenge	Not at challenge	Not Applicable	Unable to assess	Varied by site
Selecting or designing HIT to support the innovation	☐	☐	☐	☐	☐	☐	☐
Building out or installing HIT to support the innovation	☐	☐	☐	☐	☐	☐	☐
Identifying, hiring, or obtaining vendor support for innovation HIT requirements	☐	☐	☐	☐	☐	☐	☐
Data standardization across systems	☐	☐	☐	☐	☐	☐	☐
Interoperability across organizations	☐	☐	☐	☐	☐	☐	☐
Alignment of HIT with clinical workflow	☐	☐	☐	☐	☐	☐	☐
Acceptability of the HIT by front line staff	☐	☐	☐	☐	☐	☐	☐

Please comment further on the challenges and on the extent to which the awardee was able to overcome each of the challenges you have marked as major or moderate above. _____

Implementation Evaluation-Other

20. By what approximate date was the innovation considered fully implemented?

- Quarter [Drop down]: Jan-Mar, Apr-Jun, Jul-Sep, Oct-Dec, UNABLE TO ASSESS
- Year: [Drop down]: 2012, 2013, 2014, 2015, UNABLE TO ASSES

The next two items ask you to identify a few factors which you believe uniquely supported or hindered implementation. These can be the factor(s) that stand out in your mind as the most important for innovation implementation success or failure, or unexpected factors that influenced implementation.

21. Was there anything that uniquely supported implementation for this awardee?

☐ Yes
☐ No
If yes, please describe: _____

22. Was there anything that uniquely inhibited implementation for this awardee?

☐ Yes
☐ No
If yes, please describe: _____

23. Based on your knowledge of the processes, activities, and management supporting this innovation, do you believe the innovation could be successfully disseminated for wide-spread adoption and implementation?

☐ Yes
☐ No
If yes, please discuss why, if no, please discuss the limitations to dissemination and wide-spread adoption. _____

Impact Assessment

In this section we will ask for your evaluation of the innovation's impact on the core four outcomes being measured as part of the HCIA evaluation (total cost of care, hospital admissions, 30 day hospital readmissions, and emergency department visits).

We will also ask you about impact on up to 3 additional measures that you are evaluating as part of your evaluation from the categories below. The measures you specify as additional outcomes should be measures that you think have the most direct relevance to the innovation.

> **Health Care Processes and Experience of Care:** *The impact of the intervention or services on clinical processes related to quality or safety, coordination of care, patient experience, timeliness, and efficiency.*
> **Health Outcomes:** *The impact of the intervention or services on health outcomes including mortality, morbidity, health-related quality of life, functional or symptom status, and health-care associated harms (e.g., health-care associated infections, iatrogenic injury or exposures).*
> **Resource Use:** *The impact of the intervention on health care use other than hospital admissions, readmissions and ED visits. For example, use of diagnostic laboratory or imaging tests, medication, outpatient primary care visits, outpatient specialty care visits, outpatient mental health visits, ancillary care, etc.*
> **Health Care Costs:** *The impact of the intervention or services on health care costs other than total cost of care.*

Please specify <u>*up to*</u> 3 ***impact outcomes (***other than the core four) that have the most direct relevance to the awardee's innovation that you are estimating. ***For example***, if you are measuring impact on diabetes control related to an innovation, you would likely specify a measure involving hemoglobin A1C (mean change, % at goal, etc.).

Specify Name of Additional Outcome 1_____

Specify Name of Additional Outcome 2_____

Specify Name of Additional Outcome 3_____

Comments for this awardee related to any items above _____

24. Based on the nature of the innovation and population targeted, *please rate the extent to which the innovation is likely to impact each of the core four measures and the additional outcomes you have specified above.* In other words, does the logic model or theory of change for the innovation suggest a direct impact on these measures and within what timeframe might that be expected?

	Impact within 1 years	Impact within 3 years	Impact beyond 3 years	Unlikely to ever impact this measure	Unable to assess
Total costs of care	☐	☐	☐	☐	☐
Hospital admissions	☐	☐	☐	☐	☐
30-day hospital readmissions	☐	☐	☐	☐	☐
Emergency department visits	☐	☐	☐	☐	☐
Additional Outcome 1	☐	☐	☐	☐	☐
Additional Outcome 2	☐	☐	☐	☐	☐
Additional Outcome 3	☐	☐	☐	☐	☐

Comments for this awardee related to any items above _____

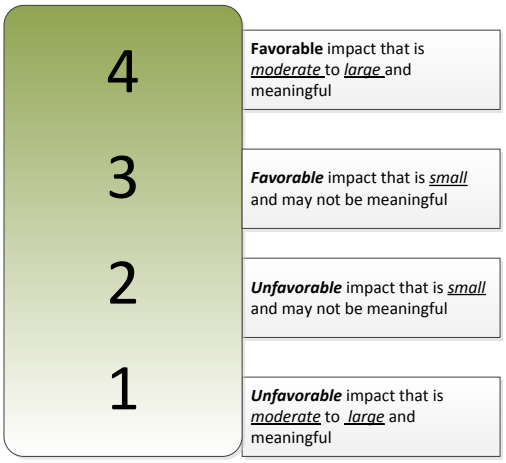
The item below asks you to assess the 1) the magnitude and direction of each impact outcome and 2) the certainty of each impact outcome. You will make these assessments for each of the core four measures and also for the additional outcomes you have specified for this awardee.

Magnitude and direction of impact on each outcome:

Magnitude refers to the effect size of the innovation impact, including whether it is clinically meaningful or meaningful from a policy perspective.

Direction refers to whether the innovation improves the outcomes (i.e., favorable impact) or worsens outcomes (i.e., unfavorable impact).

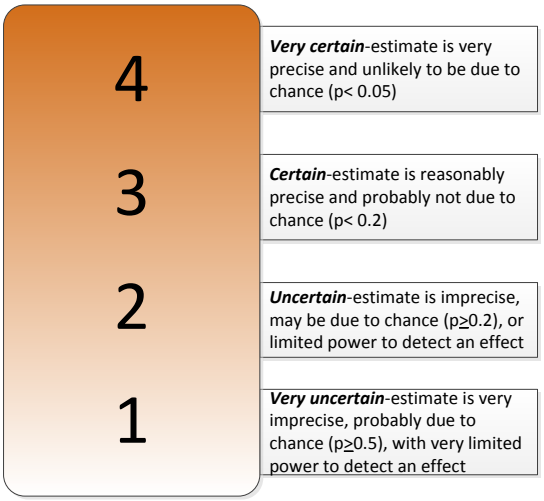
Please select your rating below based on the awardee goals and context (e.g., a 2% improvement on a measure for one awardee may represent a meaningful magnitude, whereas a 2% improvement on a different measure in another awardee may be meaningless).



Certainty of impact for each outcome

The degree to which the impact outcome estimate is precise, whether it could be due to chance, and whether the evaluation was adequately powered to detect an effect.

A precise estimate is one with a narrow confidence interval that has no substantive difference in interpretation across the interval. An imprecise estimate is one with a wide confidence interval for which the lower bound may result in different actions taken as compared to the upper bound.



The degree to which the estimate could be due to chance is assessed through classical statistical significance testing. Estimates associated with a $p < 0.05$ are unlikely to be due to chance, those with $p < 0.2$ are probably not due to chance.

26. Provide your assessment of magnitude and direction and certainty using the drop down selections below. [Each drop down includes: 4, 3, 2, 1, UNABLE TO ASSESS]

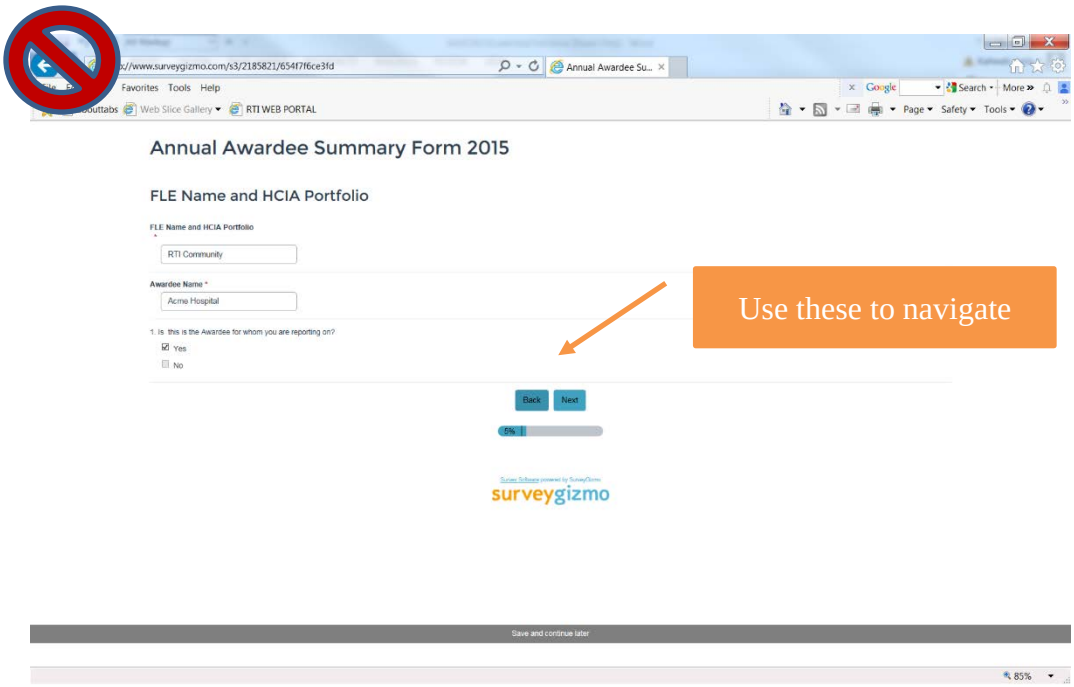
	Magnitude and Direction of Impact	**Certainty of Impact**
Total Cost of Care	[drop down]	[drop down]
Hospital admissions	[drop down]	[drop down]
30 day hospital readmissions	[drop down]	[drop down]
Emergency department visits		
Additional Outcome 1	[drop down]	[drop down]
Additional Outcome 2	[drop down]	[drop down]
Additional Outcome 3	[drop down]	[drop down]

APPENDIX B-2:
AASF2 INSTRUCTIONS

1. Browsers and Navigation

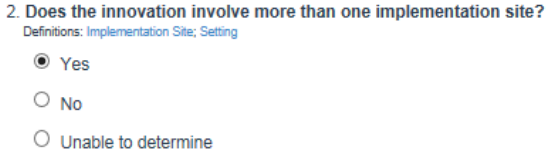
The form is compatible with the latest versions of **Google Chrome, Mozilla Firefox and Microsoft Internet Explorer**. If using the form on a mobile device, use up to date versions of built in mobile browsers such as **Safari**, **Android and IE for Mobile**.

Use the Back and Next buttons within the survey to move among the pages. Do NOT use the browser's back and next buttons.



2. Changing Response Options

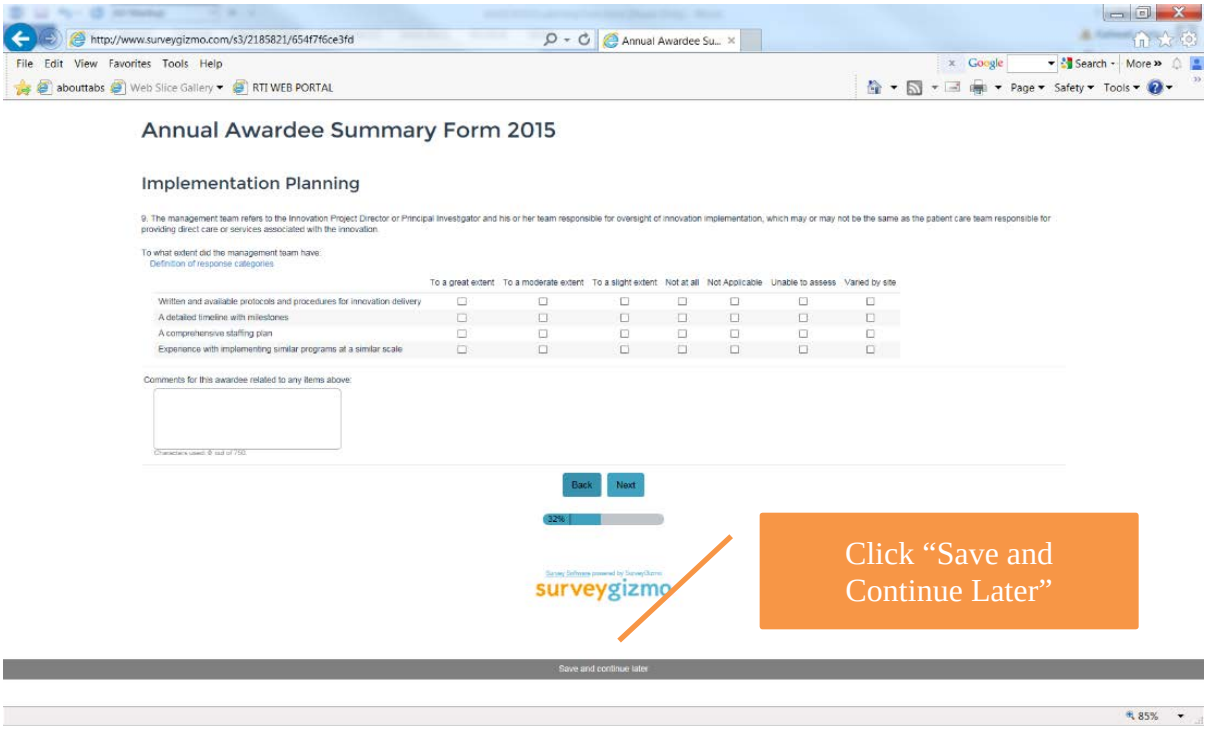
Responses to questions with radio buttons like below can be changed by simply clicking a different button.



Responses to questions with boxes like below can be changed by first clicking the box to uncheck the response, then clicking in a different box.

3. Saving Your Form To Finish at a Later Time

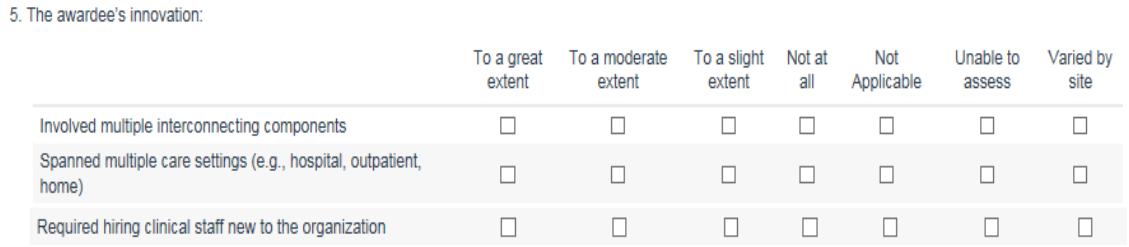
Select the "Save and Continue Later" option at the bottom of any survey page. A box will pop up requesting your email and you will be emailed a link to be able to access the form right where you left off at a later time or date.



4. Response Category Definitions

Nearly all items include an option of "Unable to assess". Please use this option sparingly, and only when you feel you do not have the appropriate data or information to make a reasonable estimate.

Definitions for the response categories that use the scales below are in the tables below. These categories are purposely generic to accommodate the diversity of innovations and items using this scale. For items that use these scales, provide your best overall assessment for the awardee.

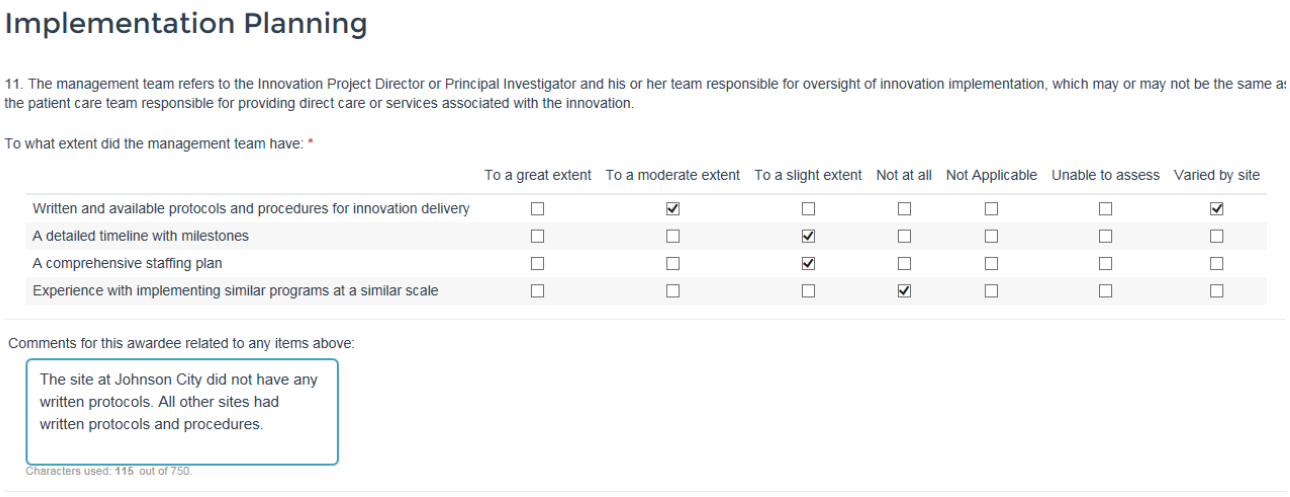


5. The awardee's innovation:

	To a great extent	To a moderate extent	To a slight extent	Not at all	Not Applicable	Unable to assess	Varied by site
Involved multiple interconnecting components	☐	☐	☐	☐	☐	☐	☐
Spanned multiple care settings (e.g., hospital, outpatient, home)	☐	☐	☐	☐	☐	☐	☐
Required hiring clinical staff new to the organization	☐	☐	☐	☐	☐	☐	☐

To a great extent	Data and information collected, compiled, or synthesized about the innovation suggests that the stated characteristic, feature, or description is consistently exhibited to a large degree. The innovation manifests the stated characteristics, feature or description to a large degree; most everyone would agree with characterizing the innovation using this description.
To a moderate extent	Data and information collected, compiled or synthesized about the innovation suggests that the stated characteristic, feature, or description is exhibited to a moderate degree. The innovation manifests the stated characteristic, feature, or description, but may do so somewhat inconsistently across data sources; many would agree with characterizing the innovation using this description but might differ with respect to how strongly they felt the stated characteristic is exhibited by the innovation.
To a slight extent	Data and information collected, compiled or synthesized about the innovation suggests that the stated characteristic, feature, or description is exhibited to a slight degree. The stated characteristic, feature, or description isn't completely absent, but its presence is minimal across data sources. Some would agree with characterizing the innovation using this description but some might differ and feel the innovation doesn't really exhibit the stated characteristic.
Not at all	Data and information collected, compiled or synthesized about the innovation suggests that the stated characteristic, feature, or description is not exhibited at all in the innovation. The stated characteristic, feature, or description is completely absent. Most would agree that the innovation doesn't exhibit the stated characteristic at all. If the stated characteristic does not make any sense for the innovation you are evaluating, select the *not applicable* option instead of this option.
Not applicable	The stated characteristics, feature, or description is not applicable to the innovation being evaluated. Please use this response option sparingly.
Unable to assess	If your data, knowledge, and information collected from the awardee is insufficient for you to make this assessment, select this option.

Varied by site	To be used only as a secondary selection, in addition to your best overall estimate. Use this option only there are major differences across sites. *For example*, if you have several implementation sites, one of which did not exhibit the characteristic at all, and the rest that did to a great extent, an appropriate response for the awardee might be the weighted average "to a moderate extent" (the second strongest rating). Because the sites varied greatly, selecting "varied by site" lets us know the response represents a range of responses across sites. You can provide additional details in the comments box to explain a "varied by site" rating.



Implementation Planning

11. The management team refers to the Innovation Project Director or Principal Investigator and his or her team responsible for oversight of innovation implementation, which may or may not be the same as the patient care team responsible for providing direct care or services associated with the innovation.

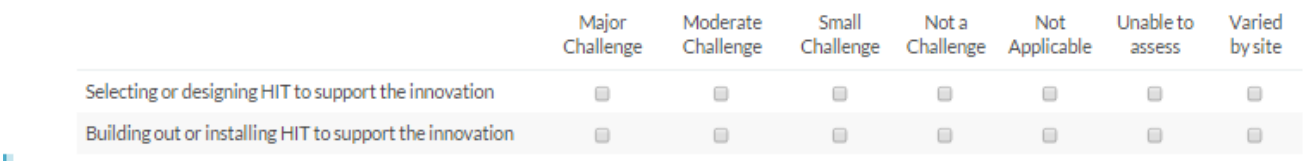
To what extent did the management team have: *

	To a great extent	To a moderate extent	To a slight extent	Not at all	Not Applicable	Unable to assess	Varied by site
Written and available protocols and procedures for innovation delivery	☐	☑	☐	☐	☐	☐	☑
A detailed timeline with milestones	☐	☐	☑	☐	☐	☐	☐
A comprehensive staffing plan	☐	☐	☑	☐	☐	☐	☐
Experience with implementing similar programs at a similar scale	☐	☐	☐	☑	☐	☐	☐

Comments for this awardee related to any items above:

The site at Johnson City did not have any written protocols. All other sites had written protocols and procedures.

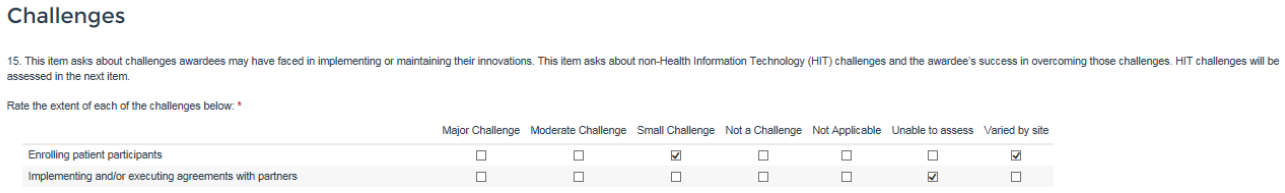
Characters used: 115 out of 750.



	Major Challenge	Moderate Challenge	Small Challenge	Not a Challenge	Not Applicable	Unable to assess	Varied by site
Selecting or designing HIT to support the innovation	☐	☐	☐	☐	☐	☐	☐
Building out or installing HIT to support the innovation	☐	☐	☐	☐	☐	☐	☐

Major Challenge	Data and information collected, compiled, or synthesized about the innovation suggests that the stated challenge was large, impacted the implementation of the innovation in a major way, required significant changes or resources to overcome (if at all). Most everyone would agree with characterizing the stated challenge as a major challenge.
Moderate Challenge	Data and information collected, compiled or synthesized about the innovation suggests that the stated challenge was exhibited to a moderate degree. The challenge had an impact on implementation, but not enough to require significant changes of additional resources to overcome (if at all). Data sources may have been inconsistent with characterizing the stated item as a challenge. Many would agree with characterizing the stated item as a challenge, but might differ with respect to how large of a challenge it was.

Small Challenge	Data and information collected, compiled or synthesized about the innovation suggests that the stated challenge was exhibited to a small degree. The stated challenge wasn't completely absent, but its presence is minimal across data sources. It has minimal to no impact on implementation, and required little if any change or resource to overcome. Some would agree with characterizing the stated item as a challenge, but some might think it was not really a challenge at all.
Not a challenge	Data and information collected, compiled or synthesized about the innovation suggests that the stated item was not a challenge at all. Data sources are consistent in not describing the state item as a challenge. Most would agree that the innovation did not have the stated challenge at all. If the stated challenge does not make any sense for the innovation you are evaluating, select the *not applicable* option instead of this option.
Not applicable	The stated challenge is not applicable to the innovation being evaluated. Please use this response option sparingly.
Unable to assess	If your data, knowledge, and information collected from the awardee is insufficient for you to make this assessment, select this option.
Varied by site	To be used only as a secondary selection, in addition to your best overall estimate. Use this option only there are major differences across sites. *For example*, if you have two implementation sites, a site with a moderate challenge enrolling patient participants and a site that had no challenges at all, an appropriate response for the awardee might be the weighted average "small challenge". If the sites varied, selecting "varied by site" lets us know the response represents a range of responses across sites.



Challenges

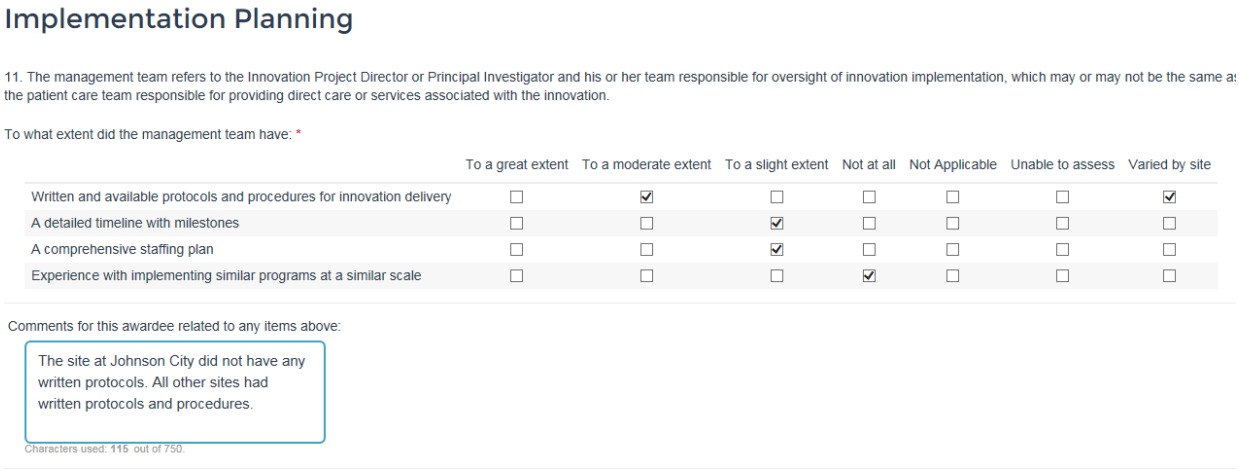
15. This item asks about challenges awardees may have faced in implementing or maintaining their innovations. This item asks about non-Health Information Technology (HIT) challenges and the awardee's success in overcoming those challenges. HIT challenges will be assessed in the next item.

Rate the extent of each of the challenges below: *

	Major Challenge	Moderate Challenge	Small Challenge	Not a Challenge	Not Applicable	Unable to assess	Varied by site
Enrolling patient participants	☐	☐	☑	☐	☐	☐	☑
Implementing and/or executing agreements with partners	☐	☐	☐	☐	☐	☑	☐

5. Variation Among Sites for Multisite Awardees

For multisite awardees base your best overall estimate on the sites for which you have sufficient knowledge and information. Select "Varied by site" in addition to your best overall estimate for the items using the scales above if there are difference among sites. Use this option only there are major differences across sites and should only be used as a secondary selection, in addition to your best overall estimate. ***For example***, if you have several implementation sites, one of which did not exhibit the characteristic at all, and the rest that did to a great extent, an appropriate response for the awardee might be the weighted average "to a moderate extent" (the second strongest rating). Because the sites varied greatly, selecting "varied by site" lets us know the response represents a range of responses across sites. You can provide additional details in the comments box to explain a "varied by site" rating (see screenshot below)



Implementation Planning

11. The management team refers to the Innovation Project Director or Principal Investigator and his or her team responsible for oversight of innovation implementation, which may or may not be the same as the patient care team responsible for providing direct care or services associated with the innovation.

To what extent did the management team have: *

	To a great extent	To a moderate extent	To a slight extent	Not at all	Not Applicable	Unable to assess	Varied by site
Written and available protocols and procedures for innovation delivery	☐	☑	☐	☐	☐	☐	☑
A detailed timeline with milestones	☐	☐	☑	☐	☐	☐	☐
A comprehensive staffing plan	☐	☐	☑	☐	☐	☐	☐
Experience with implementing similar programs at a similar scale	☐	☐	☐	☑	☐	☐	☐

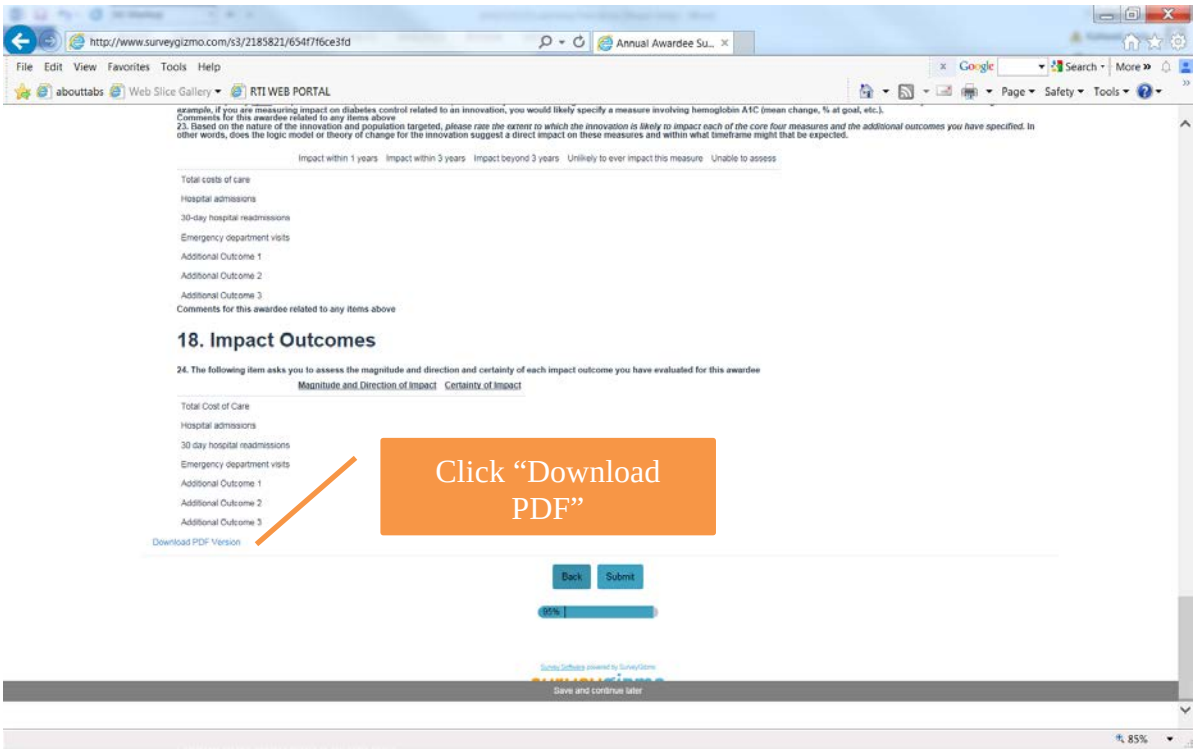
Comments for this awardee related to any items above:

The site at Johnson City did not have any written protocols. All other sites had written protocols and procedures.

Characters used: 115 out of 750.

6. Reviewing Your Responses and Saving a Copy

On the last page of the form, you will be given an option to review your responses before submitting. If you need to change any responses, use the back and next buttons within the survey to navigate to the correct page to make your corrections. Navigate back to the review page, and confirm your corrections before submitting. Once you submit your form, you will not be able to make corrections to it. You save a PDF version of your responses by selecting the “Download PDF” link at the bottom of the review page.



APPENDIX C:
AASF2 FACTOR ANALYSES

Although the survey was not designed to necessarily produce scales, to see if common factors emerged from the results we used exploratory factor analysis to test for commonalities in response within each domain. Because of concerns that our missing data ("unable to assess" and "not applicable") were not missing at random, we first created 20 multiply imputed datasets and developed factors based on those data. Seven factors emerged (see ***Table C-1***). Using three different methods of scoring awardees on factors (the factor score, the item means, and the item sums), we tested the correspondence of the multiply imputed factors against the observed data and confirmed the factor structure identified using multiply imputed results was replicated in the observed data. For ease of interpretation and utility across all analysis frames, we adopted the mean of the observed values as our metric for all analyses. Due to the abundance of missing data for organizational leadership, we omitted this construct from all quantitative analyses.

Table C-1:
Factors Identified in HCIA1 from the Second Annual Awardee Summary Form

Innovation Complexity		**Factor Loading**
Complex Duties	Involved multiple interconnecting components	0.542
	Spanned multiple care settings (e.g., hospital, outpatient, home)	0.503
	Involved staff from various groups, departments, or organizational levels providing care or services	0.742
	Added or significantly changed steps in the service delivery workflow	0.584
	Required new health information technology	0.541
	Required changes to existing staffs' roles and responsibilities	0.519
Complex Partner-ships	Required formal agreements among organizations	0.674
	Required cooperation from distributed independent providers	0.903
	Was explicitly developed and designed with intent for future dissemination to other sites	0.544
Implementation Planning		
	Written and available protocols and procedures for innovation delivery	0.549
	A detailed timeline with milestones	0.917
	A comprehensive staffing plan	0.719
	Experience with implementing similar programs at a similar scale	0.472
Staff Training		
	Front-line staff participate in formal training to support innovation delivery	0.718
	Front-line staff receive training through experiential learning (e.g., shadowing or mentoring)	0.602
	Front-line staff receive ongoing training throughout the duration of the innovation award	0.853
	New (and rotating) staff receive training to accommodate staff turn-over	0.874

Implementation Effectiveness		
	Full adoption of innovation by front-line staff	0.792
	Rapid adoption of the innovation	0.736
	Completion of all tasks needed for full innovation implementation	0.710
	Innovation components being delivered as intended and at the prescribed level of quality	0.795
	Innovation components being delivered at the intended level of intensity and frequency	0.781
Implementation Challenges		
	Staff recruitment	0.672
	Management staff turnover or unfilled management roles	0.392
	Frontline staff turnover or unfilled frontline staff roles	0.880
Health Information Challenges		
	Selecting or designing HIT to support the innovation	0.868
	Building out or installing HIT to support the innovation	0.834
	Identifying, hiring, or obtaining vendor support for innovation HIT requirements	0.752
	Data standardization across systems	0.799
	Interoperability across organizations	0.646
	Alignment of HIT with clinical workflow	0.867
	Acceptability of the HIT by front line staff	0.833

APPENDIX D:
TIME-SERIES PLOTS

D.1 Quarterly Total Cost of Care, Means, and 90% Confidence Intervals by Setting

Figure D-1:
Ambulatory-setting Quarterly Total Cost of Care, Means, and 90% Confidence Intervals

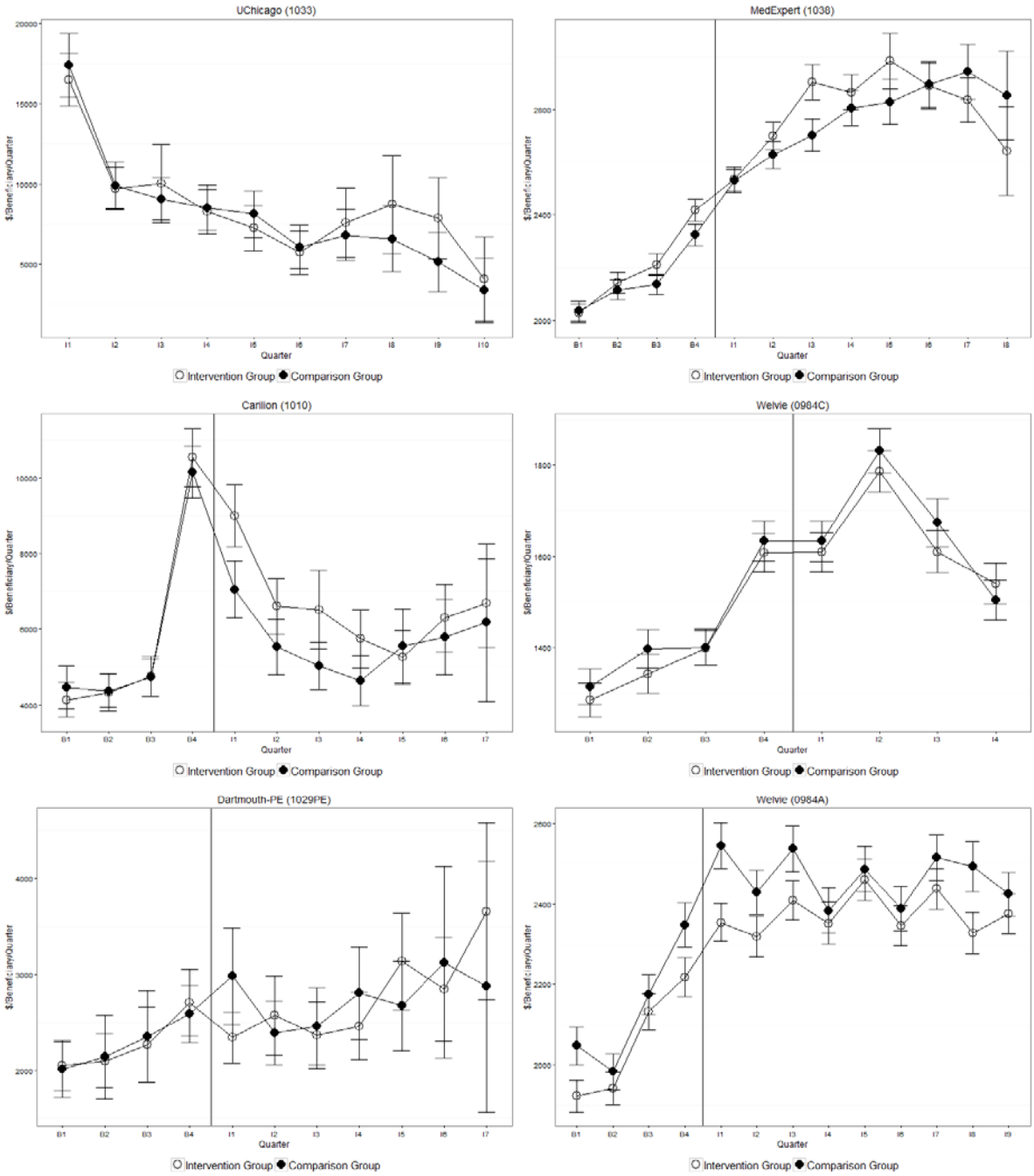


Figure D-1 (Cont.):
Ambulatory-setting Quarterly Total Cost of Care, Means, and 90% Confidence Intervals

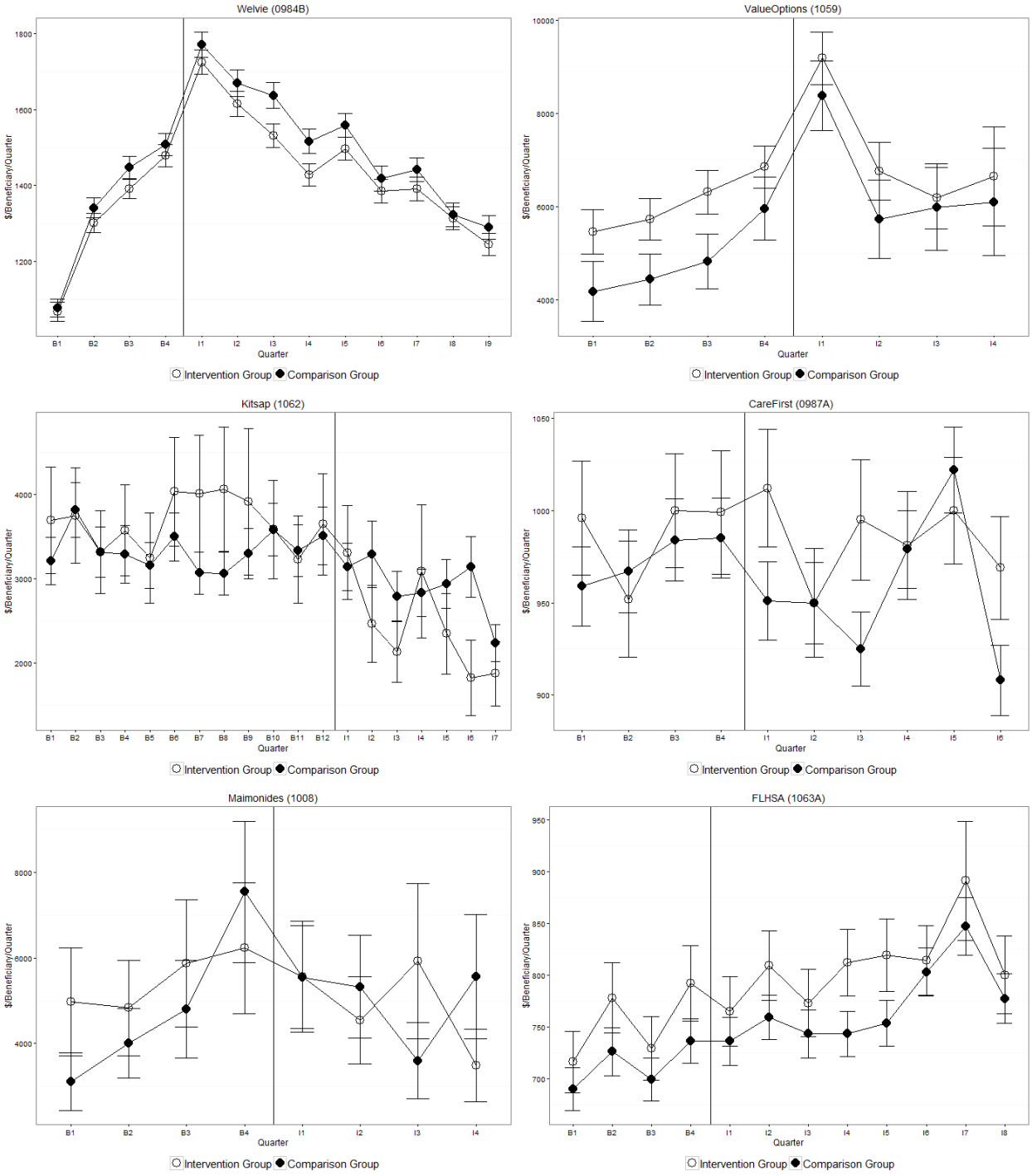


Figure D-1 (Cont.):
Ambulatory-setting Quarterly Total Cost of Care, Means, and 90% Confidence Intervals

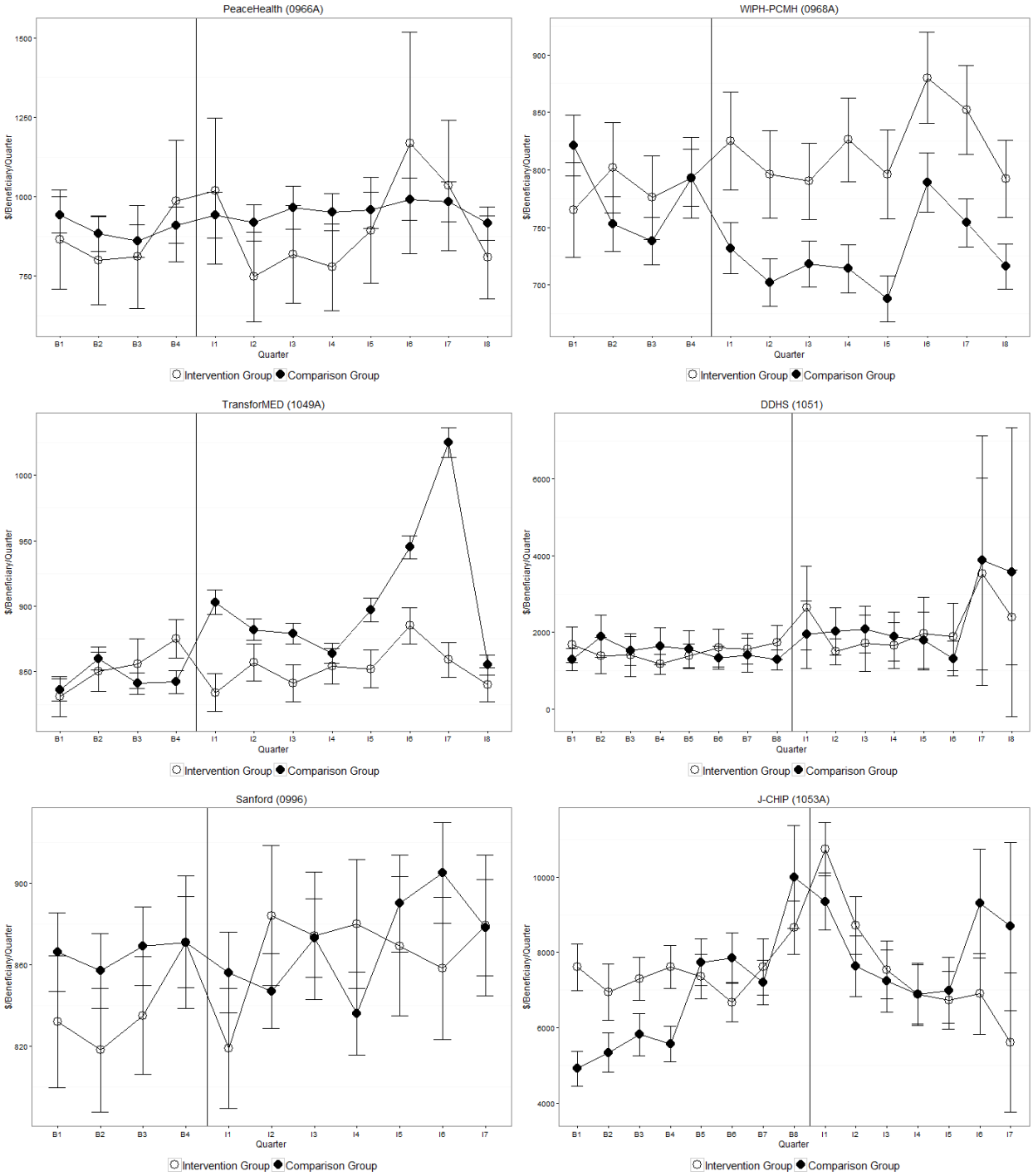


Figure D-1 (Cont.):
Ambulatory-setting Quarterly Total Cost of Care, Means, and 90% Confidence Intervals

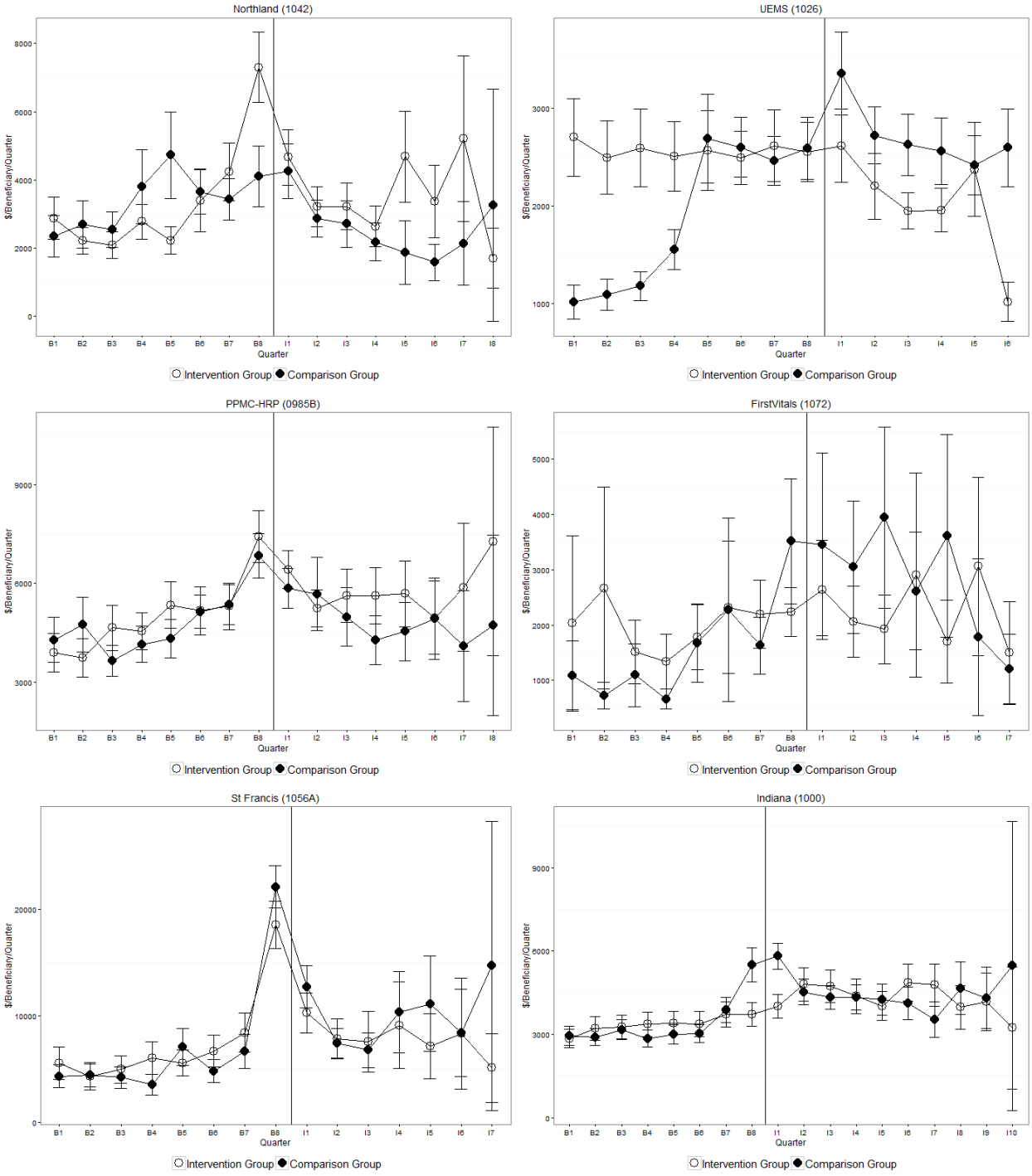


Figure D-1 (Cont.):
Ambulatory-setting Quarterly Total Cost of Care, Means, and 90% Confidence Intervals

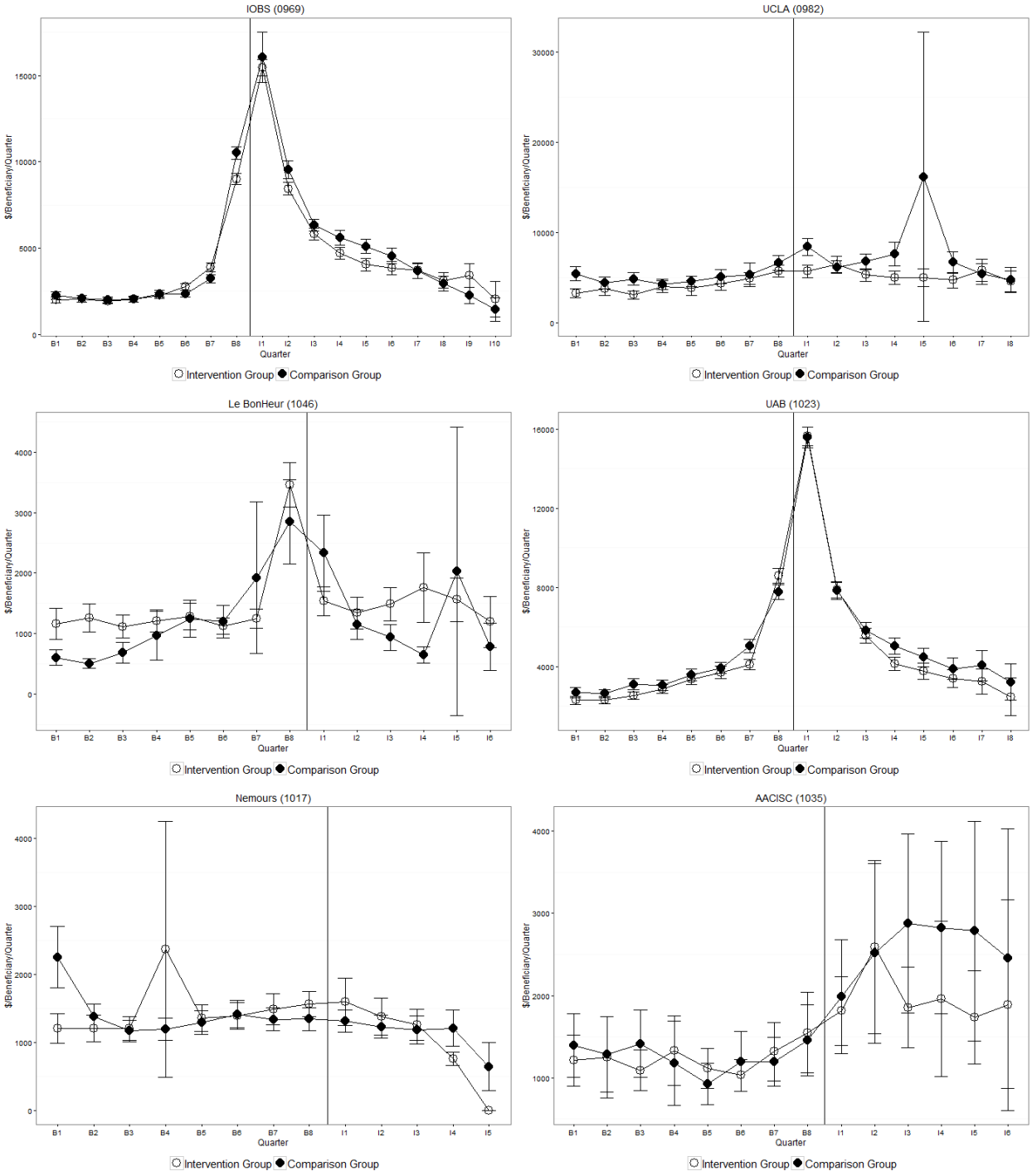


Figure D-1 (Cont.):
Ambulatory-setting Quarterly Total Cost of Care, Means, and 90% Confidence Intervals

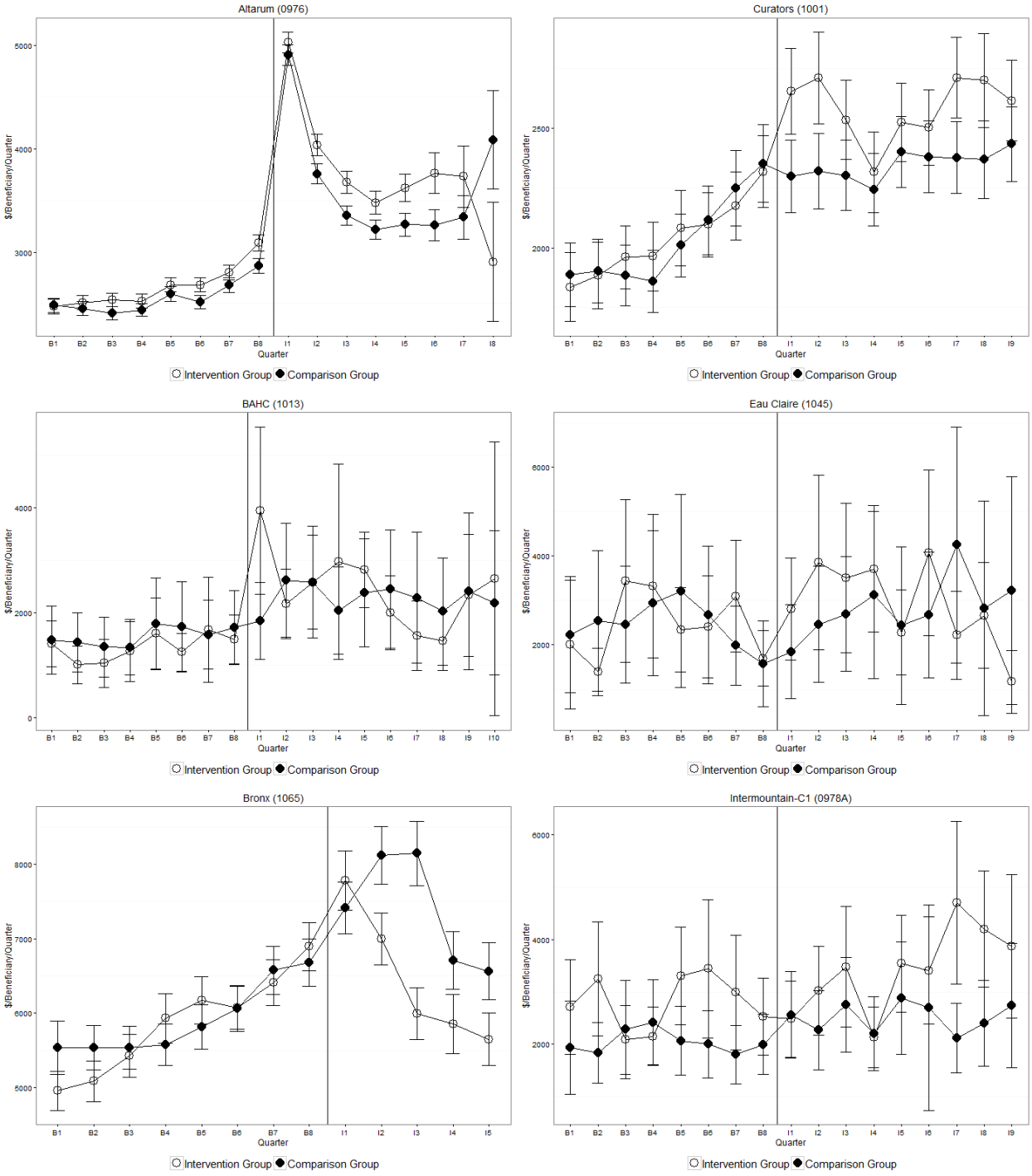


Figure D-1 (Cont.):
Ambulatory-setting Quarterly Total Cost of Care, Means, and 90% Confidence Intervals

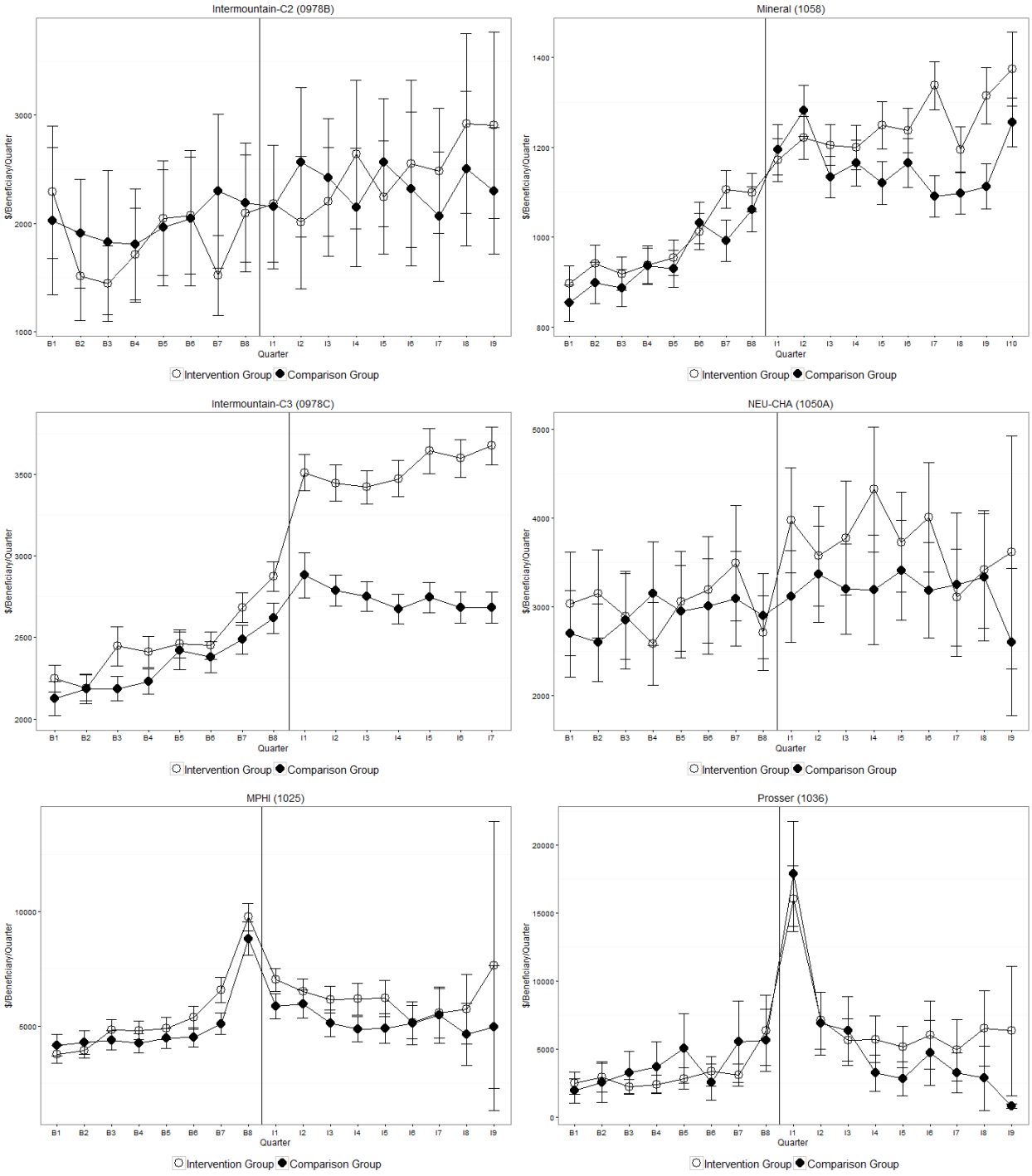


Figure D-1 (Cont.):
Ambulatory-setting Quarterly Total Cost of Care, Means, and 90% Confidence Intervals

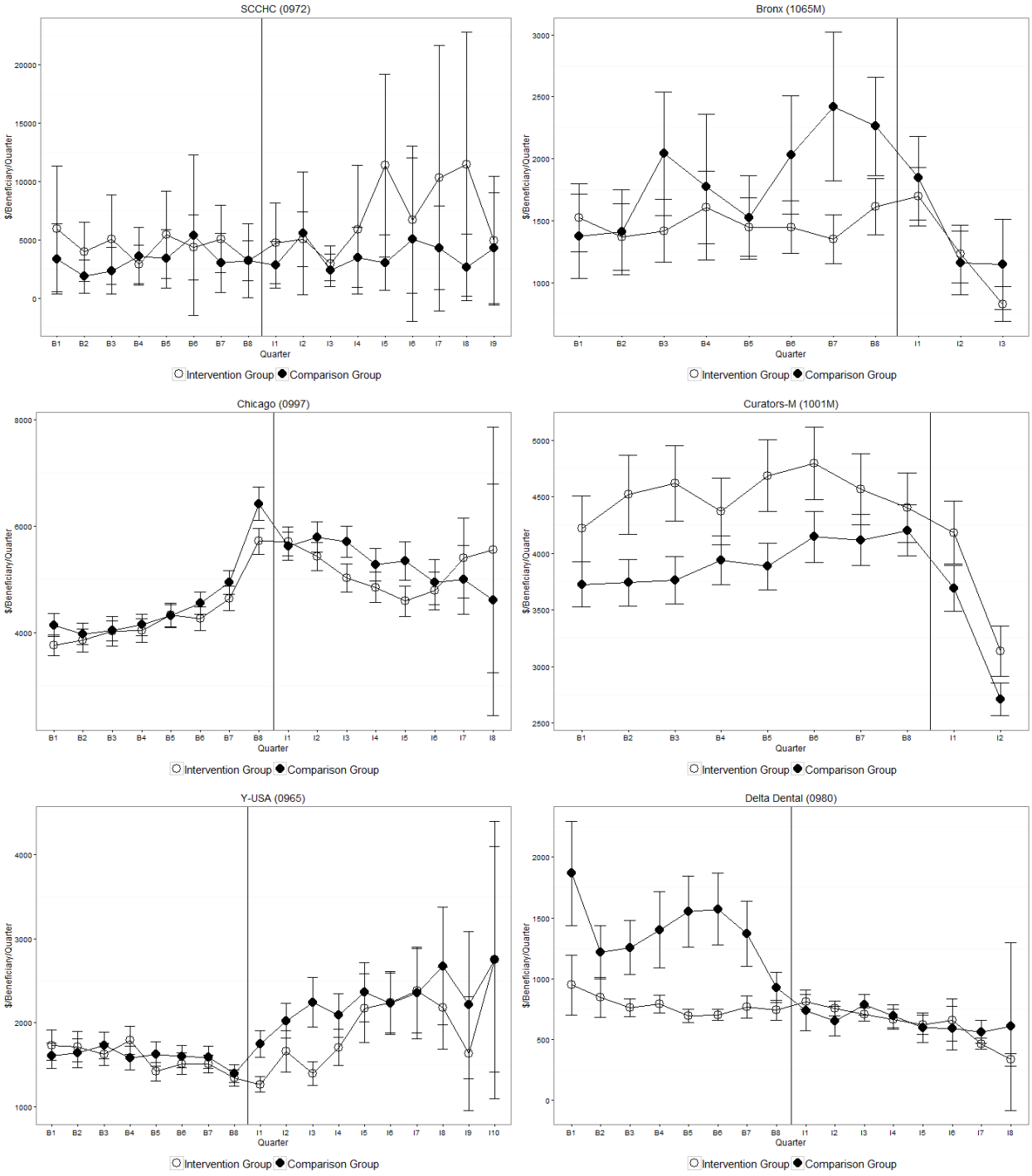


Figure D-1 (Cont.):
Ambulatory-setting Quarterly Total Cost of Care, Means, and 90% Confidence Intervals

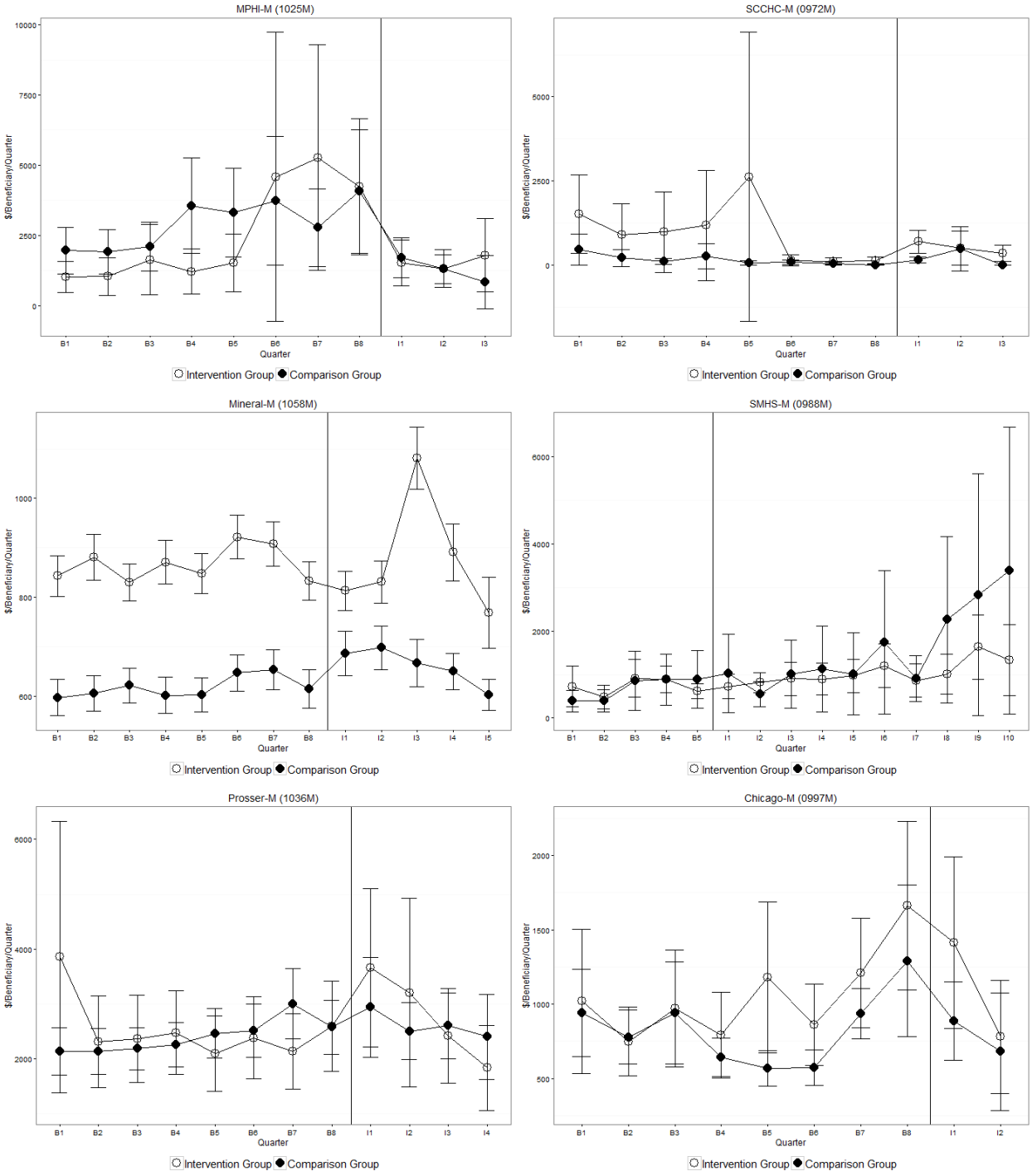


Figure D-2:
Post-Acute-Care Quarterly Total Cost of Care, Means, and 90% Confidence Intervals

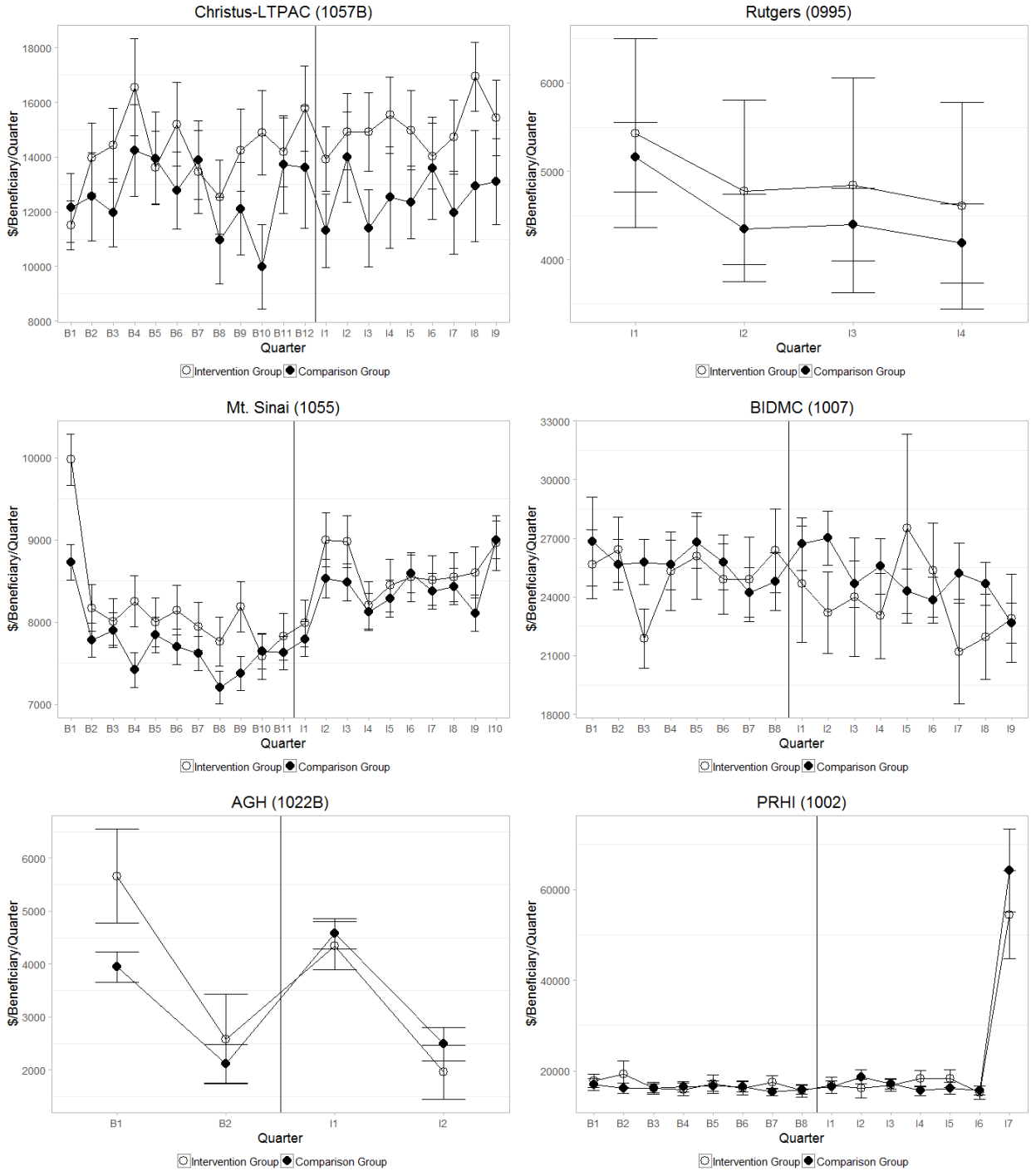


Figure D-2 (Cont.):
Post-Acute-Care Quarterly Total Cost of Care, Means, and 90% Confidence Intervals

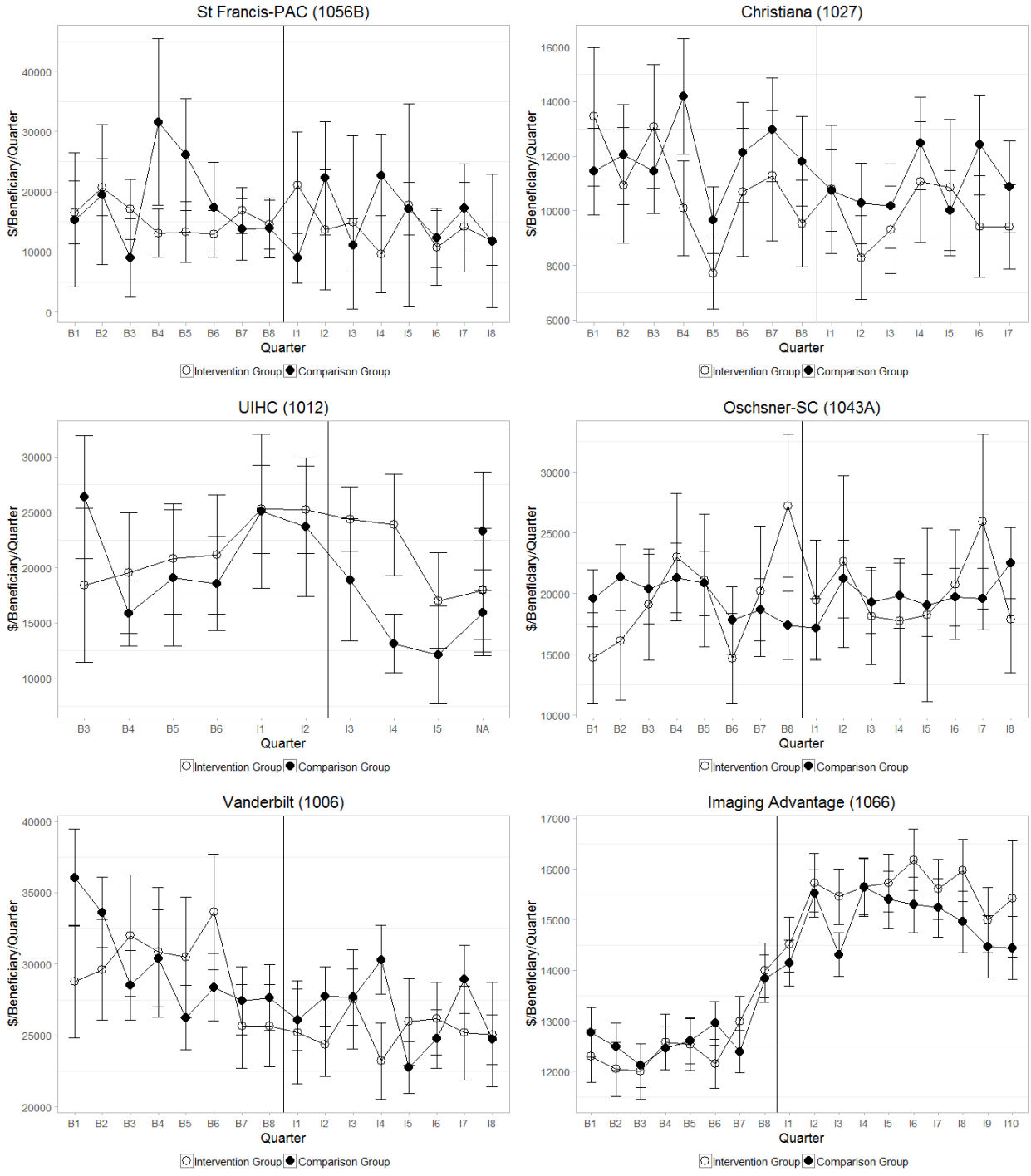


Figure D-2 (Cont.):
Post-Acute-Care Quarterly Total Cost of Care, Means, and 90% Confidence Intervals

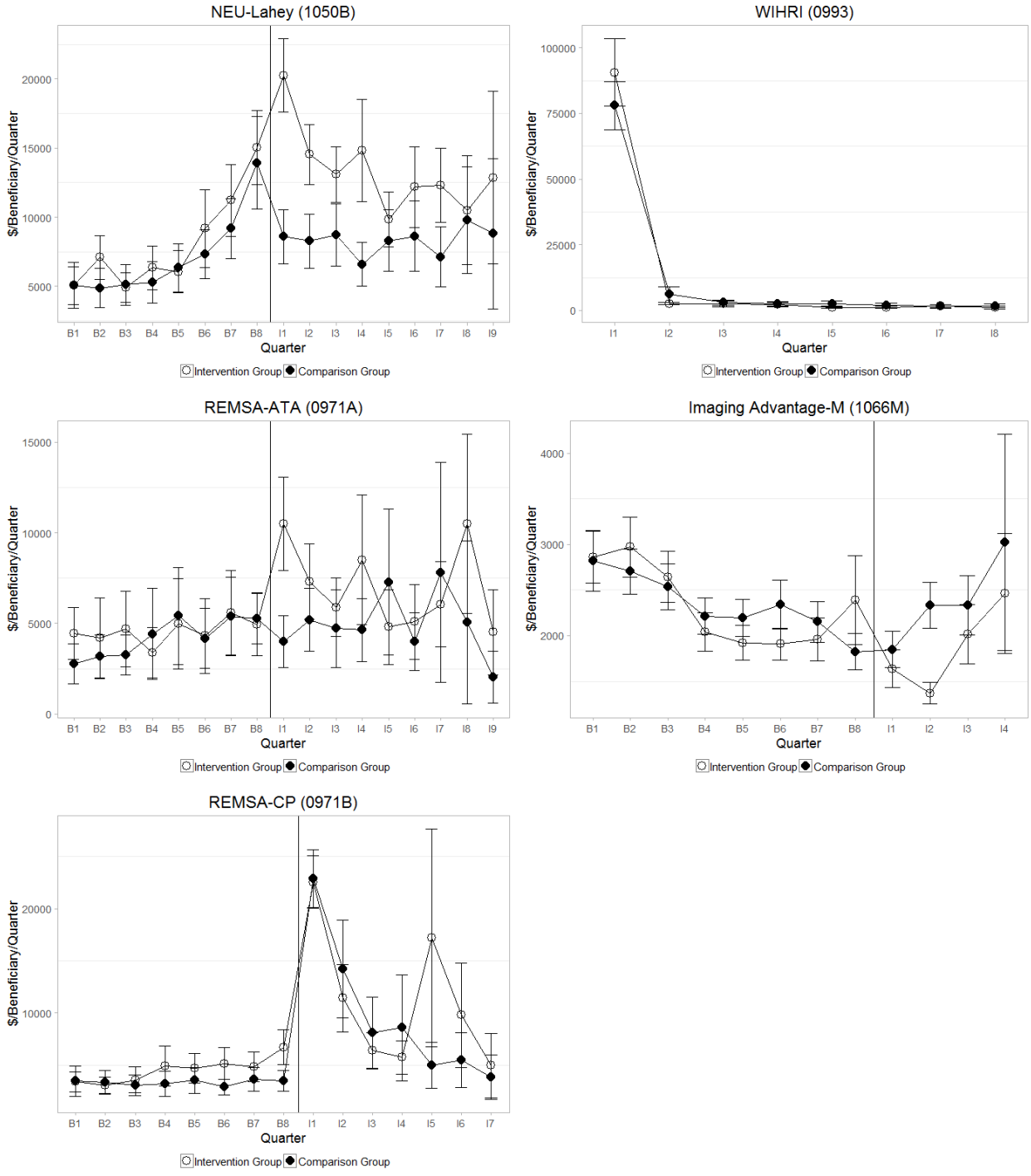


Figure D-3:
Hospital-Setting Quarterly Total Cost of Care, Means, and 90% Confidence Intervals

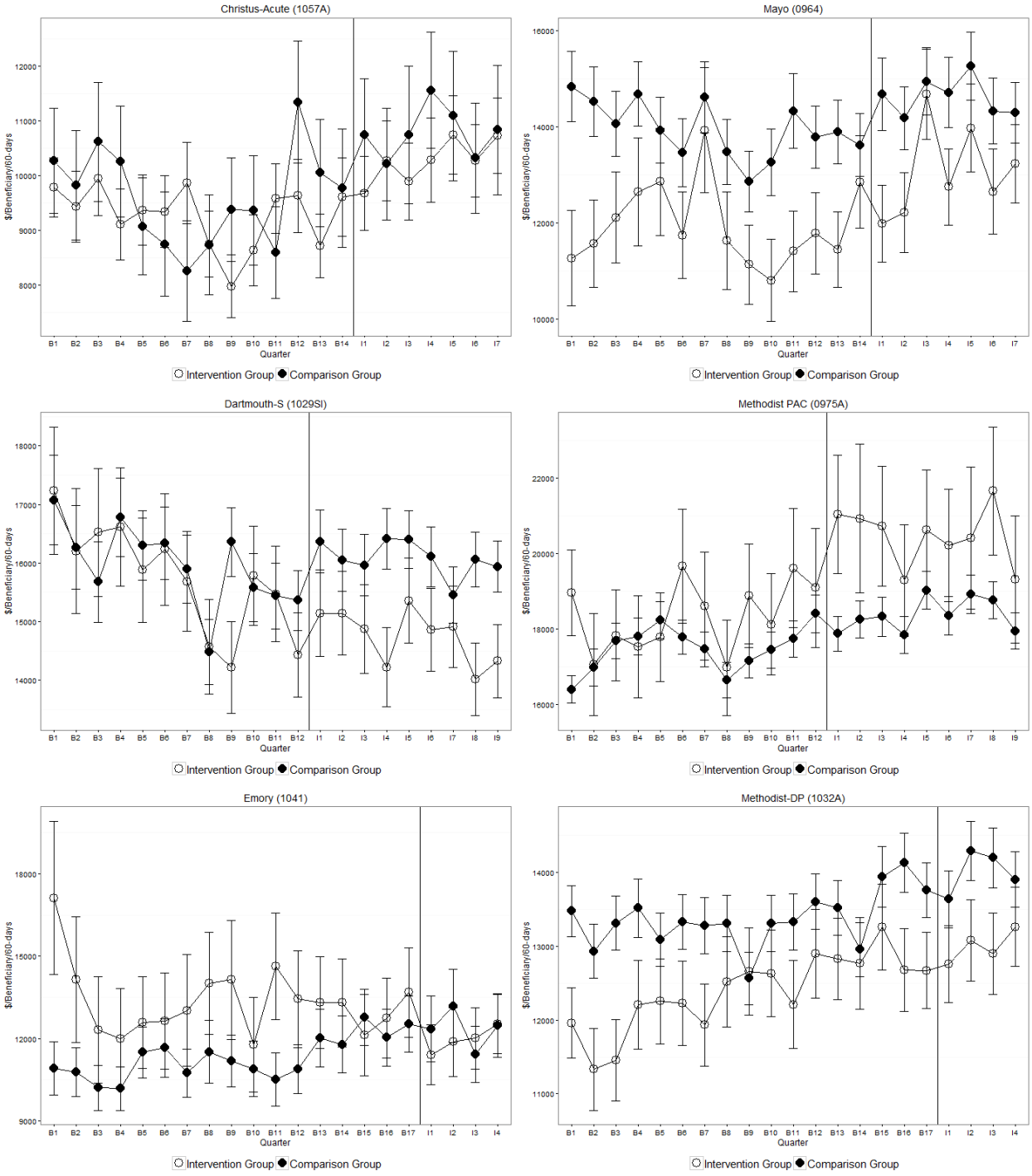
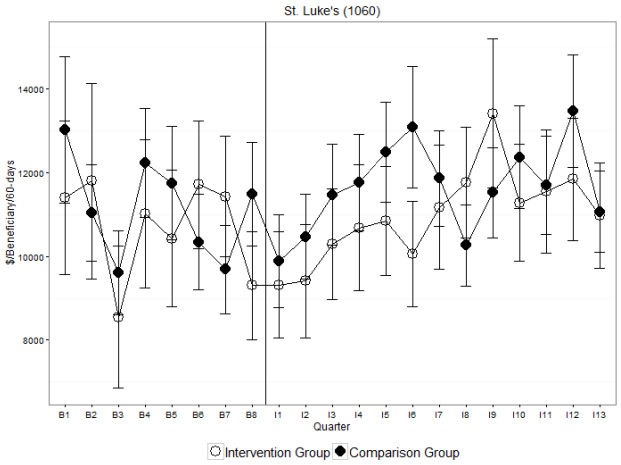


Figure D-3 (Cont.):
Hospital-Setting Quarterly Total Cost of Care, Means, and 90% Confidence Intervals



D.2 Intervention Group and Comparison Group Quarterly Grand Means and Distributional Characteristics

Table D-1:
Intervention Group and Comparison Group Quarterly Grand Means and Distributional Characteristics, Ambulatory Interventions

		Intervention Group			Comparison Group		
Quarter	Number of Estimates	Grand Mean	Lower 90% Confidence Interval	Upper 90% Confidence Interval	Grand Mean	Lower 90% Confidence Interval	Upper 90% Confidence Interval
-12	1*	-	-	-	-	-	-
-11	1*	-	-	-	-	-	-
-10	1*	-	-	-	-	-	-
-9	1*	-	-	-	-	-	-
-8	38	1213.64	1191.61	1235.67	1400.37	1377.57	1423.16
-7	38	1212.96	1191.46	1234.47	1465.58	1441.77	1489.4
-6	38	1222.18	1201.91	1242.46	1279.73	1258.97	1300.5
-5	39	1306.53	1285.47	1327.6	1403.32	1380.57	1426.07
-4	53	935.7	929.81	941.59	1084.18	1075.73	1092.63
-3	53	972.53	966.43	978.63	1114.36	1105.89	1122.82
-2	53	931.29	925.53	937.05	1171.61	1162.52	1180.7
-1	53	365.11	361.36	368.85	1164.29	1155.52	1173.07
0	53	1016.28	1009.85	1022.71	1173.5	1164.6	1182.41
1	53	976.36	970.37	982.35	1148.65	1139.63	1157.66
2	50	963.01	957.06	968.97	1112.17	1103.24	1121.1
3	48	938.87	933.1	944.65	1088.6	1079.77	1097.44
4	43	954.14	947.63	960.65	1064.07	1054.5	1073.64
5	41	990.33	983.82	996.83	1068.83	1059.28	1078.39
6	37	1043.97	1035.89	1052.04	1044.51	1034.55	1054.48
7	29	885.73	879.11	892.35	952.94	942.92	962.96
8	16	1481.75	1458.34	1505.15	1519.19	1496.44	1541.94
9	6	1260.17	1206.37	1313.96	1385.57	1304.45	1466.68

*Quarters with less than three interventions were excluded from the analysis.

Table D-2:
Intervention Group and Comparison Group Quarterly Grand Means and Distributional Characteristics, Post-acute Interventions

		Intervention Group			Comparison Group		
Quarter	Number of Estimates	Grand Mean	Lower 90% Confidence Interval	Upper 90% Confidence Interval	Grand Mean	Lower 90% Confidence Interval	Upper 90% Confidence Interval
-12	1*	-	-	-	-	-	-
-11	2*	-	-	-	-	-	-
-10	2*	-	-	-	-	-	-
-9	2*	-	-	-	-	-	-
-8	13	7170.56	7012.13	7329	6819.91	6637.5	7002.33
-7	13	6906.94	6761.68	7052.19	6999.07	6808.21	7189.92
-6	14	6994.47	6851.43	7137.51	6574.54	6393.38	6755.71
-5	14	6062.58	5930.97	6194.19	5343.13	5185.15	5501.12
-4	14	6021.19	5889.51	6152.87	4793.63	4645.07	4942.2
-3	14	6864.3	6717.52	7011.07	4713.73	4571.72	4855.73
-2	15	6060.41	5935.13	6185.68	5878.09	5717.33	6038.84
-1	15	5500.81	5377.72	5623.9	8094.22	7889.3	8299.15
0	15	5878.17	5757.31	5999.02	5288.05	5139.09	5437.01
1	15	6450.12	6311.54	6588.7	3039.45	2932.58	3146.32
2	14	8472.45	8309.41	8635.5	7722.09	7524.18	7919.99
3	14	9658.36	9467.3	9849.43	9423.94	9196.82	9651.06
4	13	10105.82	9910.63	10301.02	10574.65	10313.26	10836.05
5	12	10419.74	10222.38	10617.09	10615.31	10367.73	10862.89
6	12	9692.64	9498.99	9886.29	10323.75	10070.48	10577.03
7	9	10021.11	9821.77	10220.44	10558.43	10298.58	10818.27
8	6	9285.34	9087.05	9483.62	10137.06	9867.6	10406.53
9	2*	-	-	-	-	-	-

*Quarters with less than three interventions were excluded from the analysis.

Table D-3:
Intervention Group and Comparison Group Quarterly Grand Means and Distributional Characteristics, Hospital Interventions

		Intervention Group			Comparison Group		
Quarter	Number of Estimates	Grand Mean	Lower 90% Confidence Interval	Upper 90% Confidence Interval	Grand Mean	Lower 90% Confidence Interval	Upper 90% Confidence Interval
-17	2*	-	-	-	-	-	-
-16	2*	-	-	-	-	-	-
-15	2*	-	-	-	-	-	-
-14	4	12929.58	12627.2	13231.96	10966.27	10601.36	11331.18
-13	4	12889.34	12596.76	13181.92	11184.63	10804.99	11564.27
-12	7	14400.63	14228.64	14572.62	12473.75	12230.18	12717.33
-11	7	13701.21	13518.98	13883.44	11565.8	11316.21	11815.39
-10	7	13995.25	13810.32	14180.19	11887.35	11632.19	12142.51
-9	7	13819.68	13634.5	14004.86	12122.81	11864.76	12380.85
-8	8	13997.44	13816.75	14178.12	12313.51	12059.44	12567.58
-7	8	13785.51	13610.04	13960.97	11732.04	11482.82	11981.25
-6	8	13677	13502.54	13851.45	11637.25	11396.64	11877.85
-5	8	13725.69	13545.89	13905.48	11926.25	11681.25	12171.26
-4	8	13575.2	13400.92	13749.48	11935.27	11689.62	12180.91
-3	8	14059.99	13876.59	14243.38	12164.18	11917.27	12411.09
-2	8	14200.81	14016.91	14384.7	12000.62	11754.53	12246.71
-1	8	14085.72	13908.22	14263.21	12493.12	12242.35	12743.89
0	8	14230.1	14051.85	14408.35	12309.71	12073.03	12546.39
1	8	14277.16	14094.65	14459.67	12409.83	12158.99	12660.66
2	8	14671.06	14481.24	14860.87	12733.86	12483.37	12984.35
3	8	14468.85	14288	14649.7	12719.76	12477.07	12962.44
4	6	15177.32	14969.59	15385.04	12712.37	12435.57	12989.16
5	6	14771.96	14559.14	14984.77	12538.02	12260.08	12815.96
6	6	14835.07	14626.44	15043.7	12843.28	12557.88	13128.67
7	4	15047.4	14827.5	15267.31	12633.12	12318.22	12948.03
8	4	14879.3	14665.65	15092.96	12957.06	12636.42	13277.71
9	1*	-	-	-	-	-	-
10	1*	-	-	-	-	-	-
11	1*	-	-	-	-	-	-
12	1*	-	-	-	-	-	-

*Quarters with less than three interventions were excluded from the analysis.

D.3 Quarterly Grand Means and Distributional Characteristics

Table D-4:
Quarterly Grand Means and Distributional Characteristics, Ambulatory Interventions

Quarter	Number of Awardees	Median	IQR	Unweighted Mean	Weighted Mean	Lower 90% Confidence Interval	Upper 90% Confidence Interval	I^2	Kurtosis	Skew
-12	1*	-	-	-	-	-	-	-	-	-
-11	1*	-	-	-	-	-	-	-	-	-
-10	1*	-	-	-	-	-	-	-	-	-
-9	1*	-	-	-	-	-	-	-	-	-
-8	38	103.99	793.75	226.6	53.25	-59.32	165.82	79.96%	4.38	0.53
-7	38	-14	870.25	146.96	76.34	-22.22	174.9	73.99%	3.47	0.89
-6	38	41	718.25	167.46	81.01	-12.94	174.96	73.15%	5.13	0.7
-5	39	153	582.5	146.02	149.36	58.8	239.92	66.27%	5.45	0.04
-4	53	25	353	29.92	9.01	-25.78	43.8	69.61%	5.37	-0.08
-3	53	-10	272	74.92	16.31	-18.19	50.81	69.14%	5.14	0.82
-2	53	38	483	166.47	35.63	-0.09	71.35	69.67%	5.38	0.25
-1	53	37	348	-44.63	31.13	-6.49	68.75	72.48%	9.53	-0.6
0	53	61	703	63.8	35.78	-10.99	82.55	79.27%	4.95	-0.68
1	53	72	447	50.42	42.21	1.85	82.57	73.05%	3.32	0
2	50	15.5	765.75	23.78	12.93	-41.2	67.06	86.26%	4.85	-0.26
3	48	35	814.75	65.39	35.46	-12.21	83.13	81.9%	4.79	0.01
4	43	-26	562	-102.5	55.15	2.7	107.6	81.61%	15.73	-1.42
5	41	39	668	69.7	48.6	-10.95	108.15	83.89%	4.04	-0.56
6	37	52	526	103.6	97.98	13.61	182.35	90.15%	13.32	-1.93
7	29	-15	824	389.41	-3.4	-45.76	38.96	54.78%	13.19	3.02
8	16	191.5	1133.75	565.44	80.13	-35	195.26	60.71%	5.93	1.51
9	6	62	1919.5	-516.83	118.76	20.76	216.76	0%	1.52	-0.62

*Quarters with less than three interventions were excluded from the analysis.

Table D-5:
Quarterly Grand Means and Distributional Characteristics, Post-Acute Interventions

Quarter	Number of Awardees	Median	IQR	Unweighted Mean	Weighted Mean	Lower 90% Confidence Interval	Upper 90% Confidence Interval	I^2	Kurtosis	Skew
-12	1*	-	-	-	-	-	-	-	-	-
-11	2*	-	-	-	-	-	-	-	-	-
-10	2*	-	-	-	-	-	-	-	-	-
-9	2*	-	-	-	-	-	-	-	-	-
-8	13	-2	1311	-586.03	145.18	-372.5	662.86	47.05%	4.51	-1.59
-7	13	269	1683	9.03	191.11	-210.05	592.27	27.6%	3.21	-0.96
-6	14	48	1576.41	120.82	41.37	-423.35	506.09	42.4%	4.58	-0.12
-5	14	221.7	1965.22	-1005.4	69.47	-333.51	472.45	32.05%	9.67	-2.74
-4	14	85.5	1412.7	-432.86	87.15	-362.86	537.16	41.5%	9.01	-2.46
-3	14	-134.5	2984.5	448.13	421.99	-275.49	1119.47	72.71%	2.5	0.17
-2	15	674	1790.57	794.51	404.25	19.87	788.63	32.86%	2.53	-0.42
-1	15	469	1607	653.68	401.59	-41.68	844.86	35%	5.89	1.15
0	15	191.63	2691.27	2179.08	887.06	210.35	1563.77	77.27%	3.83	1.5
1	15	-528	3867.48	-775.4	-331.42	-1000.35	337.51	74.77%	3.81	-0.26
2	14	149.13	3521.79	1062.95	580.54	1.77	1159.31	49.3%	2.08	0.69
3	14	-271	5350.95	-72.49	395.67	-700.82	1492.16	75.99%	3.34	-0.24
4	13	1535	2868	2194.76	776.65	159.66	1393.64	19.59%	5.92	1.66
5	12	956.5	1578.82	770.33	280.33	-193.44	754.1	10.47%	2.88	-0.03
6	12	-667.54	4821.29	-662.16	197.38	-750.21	1144.97	52.32%	2.9	-0.27
7	9	357	922	490.23	493.78	-521.32	1508.88	53.09%	2.54	0
8	6	1431.26	1940.44	1686.08	580.53	241.64	919.42	0%	1.81	0.48
9	2*	-	-	-	-	-	-	-	-	-

*Quarters with less than three interventions were excluded from the analysis.

Table D-6:
Quarterly Grand Means and Distributional Characteristics, Hospital-Setting Interventions

Quarter	Number of Awardees	Median	IQR	Unweighted Mean	Weighted Mean	Lower 90% Confidence Interval	Upper 90% Confidence Interval	I^2	Kurtosis	Skew
-17	1*	-	-	-	-	-	-	-	-	-
-16	1*	-	-	-	-	-	-	-	-	-
-15	1*	-	-	-	-	-	-	-	-	-
-14	3	-490.56	2692.39	-746.66	-865.15	-3214.36	1484.06	88.35%	1.5	-0.17
-13	3	-392.79	2019.58	-749.02	-875.25	-2762.27	1011.77	82.03%	1.5	-0.31
-12	5	165.13	1654.86	215.47	178.63	-1216.57	1573.83	81.03%	1.99	0.14
-11	5	-60.93	1232.04	-178.73	-423.7	-1410.17	562.77	57.65%	2.23	0.49
-10	5	293.73	710.37	542.21	310.31	-449.55	1070.17	35.6%	2.36	0.41
-9	5	-159.51	875.45	279.11	-6.2	-1020.92	1008.52	63.09%	2.44	0.61
-8	6	-436.85	1189.93	-114.9	20.58	-729.46	770.62	36.43%	1.97	0.34
-7	6	382.48	1669.84	801.68	585.52	-625.24	1796.28	74.74%	2.44	0.47
-6	6	-635.07	2130.33	-116.57	-287.05	-1259.46	685.36	67.82%	1.97	0.66
-5	6	-319.92	1347.53	-446.37	-525.78	-1422.03	370.47	61.39%	2.13	-0.26
-4	6	-164.49	3321.03	-356.97	-406.21	-1835	1022.58	85.48%	1.28	-0.14
-3	6	-229.04	2005.43	-352.06	-464.51	-1369.98	440.96	62.24%	1.54	-0.05
-2	6	369.04	2475.37	97.37	-56.37	-1214.72	1101.98	78.2%	1.75	-0.39
-1	6	-467.63	1365.93	-368.83	-502.1	-1142.53	138.33	25.57%	2.01	-0.19
0	6	-1000.31	524.37	-554.74	-655.47	-1830.85	519.91	79.25%	3.44	1.19
1	6	-985.32	1067.24	-418.38	-611.24	-1497.35	274.87	60.49%	3.17	1.22
2	6	-561.74	1407.24	-68.45	-168.6	-997.17	659.97	54.44%	2.7	1.04
3	6	-1178.72	1572.38	-832.62	-980.51	-1887.79	-73.23	65.14%	2.22	0.72
4	5	-1045.16	944.43	-546.38	-658.27	-1440.54	124	44.02%	2.59	1.04
5	5	-1247.63	1628.06	-827.13	-799.31	-1887.33	288.71	72.82%	2.08	0.38
6	5	-541.63	583.55	-181	-473.16	-1015.98	69.66	0%	2.7	1.08
7	3	1495.77	2471.81	779.94	675.42	-2067.39	3418.23	91.63%	1.5	-0.48
8	3	1364.09	1754.52	548.36	360.1	-1737.28	2457.48	82.83%	1.5	-0.64
9	1*	-	-	-	-	-	-	-	-	-
10	1*	-	-	-	-	-	-	-	-	-
11	1*	-	-	-	-	-	-	-	-	-
12	1*	-	-	-	-	-	-	-	-	-

*Quarters with less than three interventions were excluded from the analysis.

APPENDIX E:
META-ANALYTIC METHODS

Our impact analyses for the four core outcomes of primary interest for HCIA awardees are limited to innovations in which FLEs conducted beneficiary-level difference-in-difference regression analyses using comparison or control groups, and reported overall effects and variance estimates for the intervention period measured so far. In a few cases, we converted p-values into standard errors. We do not report pre-post analyses without comparison groups or attempt to compute effects based solely on quarterly data means.

When an awardee implemented multiple innovations, we treated each one as a separate innovation as long as it had a distinct target population and unique impact effects; 135 interventions were identified. Some FLEs have begun to provide data for subgroups, such as "high risk" beneficiaries or beneficiaries dually eligible for both Medicare and Medicaid. In this report we include only data from the broader target population, and not for subgroups. For the meta-analysis (section 3.2), we excluded four unique population awardees (end-of-life programs, a program for mothers and infants only) because their outcome data were not comparable to other awardees. For the meta-regressions (section 3.4) we also excluded outlier estimates that could influence the regressions; outliers were defined as absolute values of more than $1,000 per beneficiary for TCOC, 100 per 1,000 beneficiaries for ED rates, and 50 per 1,000 beneficiaries for hospital admissions. When results were given for multiple time periods, we used 90-day outcomes to be consistent with the most common reporting period.

The two key estimates we use in this report—HCIA effects on core outcomes and the standard errors of those effects—were usually abstracted from awardee-specific tables in FLE's eighth quarterly reports. In some instances, results were drawn from second annual reports when impact data were not available in later reports.

We provide results for three broad types of HCIA innovations—post-acute care, ambulatory care, and hospital-setting. Post-acute care innovations are typically transition programs that target patients recently released from hospitals, nursing homes, or skilled nursing facilities. Post-acute care data are generally episode-based with follow-up periods ranging from 30 to 120 days. Ambulatory care innovations generally identify and enroll eligible patients on a rolling basis and then follow them for the reminder of the innovation period. They provide ongoing preventive, primary care, and specialist services in health care facilities or in the patient's home. The hospital-setting group involves a unique set of innovations based in hospitals, long-term care facilities, and intensive care units that are unlike any of the facilities in the other categories. The innovations in the hospital-setting group come exclusively from the Hospital-Setting HCIA Awards. We based our post-acute care and ambulatory care designations on the classifications reported by FLEs. When these classifications were not provided, we based the assignment on FLEs' descriptions of the innovation components and on our qualitative coding of individual programs.

HCIA awardee effects are summarized in the form of forest plots. The FLE-reported DiD effect is represented by a solid dot, and the effect's two-sided 90 percent confidence interval is shown by whiskers extending out horizontally from the estimate of central tendency. In each plot, a vertical line has been drawn at zero that denotes no cost or utilization impact.

For each plot, we calculated the weighted mean effect size across awardees (the grand mean) and the homogeneity statistic. Mean effects are weighted by the reciprocal of the variances for individual awardees. The homogeneity statistic tests the hypothesis that all awardee effects are the same except for sampling error. This statistic needs to be interpreted with caution because it is well known to be underpowered for meta-analysis work (Bonett & Price, 2015; Gavaghan, Moore, & McQuay, 2000). When only a small number of studies are available, the test can frequently produce false negatives by failing to reject the null hypothesis when a considerable amount of effect size heterogeneity is present.

Method for Calculating Summative DiD Estimates from Reported Quarterly DiD Estimates

Three FLEs failed to provide summative DiD estimates for some or all of their interventions for whom they could provide impact estimates for the core measures. In some cases, quarterly DiD estimates for the core measures were reported, and we were able to construct summative estimates. Our method weights each reported quarterly DiD by the number of beneficiaries in the intervention group during that quarter to yield a summative effect size (ES):

$$ES = \sum_{t=1}^{T} \frac{n_t}{N} ES_t,$$

where ES_t is the quarterly DiD effect size for intervention quarter t for $t = 1, \ldots, T$, n_t is the number of beneficiaries in the intervention group during intervention quarter t, and $N = \sum_{t=1}^{T} n_t$. To estimate the standard error associated with ES, $SE(ES)$, we calculated the following:

$$SE(ES) = \sqrt{\sum_{t=1}^{T} \left(\frac{n_t}{N}\right)^2 SE(ES_t)^2},$$

where $SE(ES_t)$ is the standard error reported for ES_t. In most cases, FLEs are following beneficiaries in the intervention and comparison groups over time to evaluate the interventions, and we expect some correlation to be present among the quarterly estimates. It is ideal for the FLEs to report summative DiD estimates that accurately represent this correlation as we can at best only crudely model it in our ad hoc calculations from the quarterly DiD estimates being reported (and do not do so in this report).

APPENDIX F:
META-REGRESSION AND PATH ANALYSIS METHODS

F.1 Meta-Regression Methods

In meta-regression, awardee effect sizes become the dependent variables in the analysis. The explanatory variables in this model are factors that are hypothesized to influence the magnitude of the awardee effects (Thompson & Higgins, 2002). In HCIA, the results for any particular awardee may be a function not only of the effectiveness of a particular treatment component, but also of beneficiary attributes, geographic influences, and other structural and implementation-related features of the innovation.

The meta-regression is based on the following general equation that explicitly accounts for heterogeneity across awardees:

$$Y_i = \alpha + \Sigma_j \beta_j X_{ji} + \Sigma_k \lambda_k Z_{ki} + \mu_i + \varepsilon_i,$$

where

- Y_i = the effect size for the i-th HCIA awardee, based on the most recent quarter of data
- α = an intercept term
- X_{ji} = a set of j characteristics of the i-th awardee's HCIA program
- Z_{ki} = a set of k structural features of the i-th awardee's innovation
- μ_i = unexplained (unobserved) variation in the i-th program from "true" program effect
- ε_i = residual sampling error in the i-th intervention.

The X and Y vectors consist of program-related, structural, and design-related characteristics that may have introduced the heterogeneity into the results and that may systematically elevate or reduce observed effect sizes. The impact of these features are estimated by the associated β_j and λ_k coefficients.

Because estimated outcome effects are considerably more precise for some innovations than for others, we performed weighted regressions with weights equal to the inverse of the TCOC error variance. Large weights were capped at three times the median value to prevent the biggest programs from having undue influence on the results. A weighted covariance matrix was used as input to regression estimation routines.

F.2 Path Analysis Methods

We used path analysis to analyze two models, one for the determinants of implementation effectiveness and a second for the relationships between implementation features and the core outcomes. An extension of regression methods, path analysis is a statistical technique for estimating linear associations among a set of variables arranged in a presumed, hierarchical causal sequence (Kline, 2011). The results of multiple regression equations are displayed in the form of a model that summarizes the key relationships (or paths) in the data. The magnitudes of individual effects are measured by standardized regression (beta) coefficients. These coefficients indicate how many standard deviations (SDs) an outcome would be expected to change in response to a one standard deviation increase in an explanatory variable. A beta value of 0.20, for example, indicates that the outcome is expected to increase 0.20 SDs per SD change in the explanatory variable. With all variables in the path model standardized to their respective metrics, larger betas represent larger relative effects.

Standard output for path models includes modification indexes, which estimate the effect of inserting omitted paths back into the model. We reviewed these indexes for conceptually appropriate changes in model specification.

Several indices can be used to assess the fit of a path model. Fit refers to the degree to which the proposed model reproduces the observed correlations in the data. We made fit assessments based primarily on one absolute fit measure (the Root Mean Square Error of Approximation or RMSEA) and one incremental fit measure (the Comparative Fit Index or CFI). Criteria for good model fit are RMSEA values less than 0.08 and CFI values greater than 0.95 Hooper, Coughlan, & Mullen, 2008). The model results were estimated from weighted covariance matrices using Stata 14.0.

APPENDIX G:
INNOVATION ID INDEX

Intervention Abbreviation	ID	Awardee	FLE
Mayo	0964	Mayo Clinic	Hospital
Y-USA	0965	National Council of Young Men's Christian Associations	Community
PeaceHealth	0966A	PeaceHealth Ketchikan Medical Center	Primary
Cooper	0967	Cooper University Hospital and Camden Coalition of Health Care Providers	Primary
WIPH-PCMH	0968A	Wyoming Institute of Population Health at Cheyenne Regional Medical Center - PCMH Program	Primary
WIPH-RCTP	0968B	Wyoming Institute of Population Health at Cheyenne Regional Medical Center	Primary
IOBS	0969	Innovative Oncology Business Solutions, Inc.	Disease
JHUSON	0970	Johns Hopkins School of Nursing	Complex
REMSA-ATA	0971A	Regional Emergency Medical Services Authority	Community
REMSA-CP	0971B	Regional Emergency Medical Services Authority	Community
SCCHC	0972	South County Community Health Center, Inc.	Community
SCCHC-M	0972M	South County Community Health Center, Inc.	Community
UNM	0973	University of New Mexico Health Sciences Center	Complex
Children's	0974	Children's Hospital and Health System, Inc.	Community
Methodist PAC	0975A	Methodist - Sepsis	Hospital
Methodist-SR	0975B	Methodist - Sepsis	Hospital
Altarum	0976	Altarum Institute	Community
NCH	0977	Research Institute at Nationwide Children's Hospital	Primary
Intermountain-C1	0978A	Intermountain	Community
Intermountain-C2	0978B	Intermountain	Community
Intermountain-C3	0978C	Intermountain	Community
Intermountain-C4	0978D	Intermountain	Community
VUMC	0979A	Vanderbilt University Medical Center (My Health Team)	Disease
VUMC-PAC	0979B	Vanderbilt University Medical Center (My Health Team)	Disease
Delta Dental	0980	Delta Dental Plan of South Dakota	Community
NHCHC	0981	The National Health Care for the Homeless Council	Community
UCLA	0982	Regents of the University of California, Los Angeles	Disease
Vinfen	0983	Vinfen Corporation	Behavioral
Welvie-Oh	0984A	Welvie LLC	MMSDM
Welvie-Oh-MA	0984B	Welvie LLC	MMSDM
Welvie-Tx-MA	0984C	Welvie LLC	MMSDM
PPMC-EDG	0985A	Providence Portland	Complex
PPMC-HRP	0985B	Providence Portland	Complex
PPMC-ND	0985C	Providence Portland	Complex
CLTCEC	0986	California Long-Term Care Education Center	Complex
CareFirst	0987A	CareFirst Blue Cross Blue Shield	Primary
SMHS	0988	Southeast Mental Health Services	Community
SMHS-M	0988M	Southeast Mental Health Services	Community

Intervention Abbreviation	ID	Awardee	FLE
San Juan	0990	Upper San Juan Health Service District	Disease
Felton	0991	Felton Institute (f. Family Service Agency of San Francisco)	Behavioral
HealthLinkNow	0992	HealthLinkNow Inc.	Behavioral
WIHRI	0993	Women & Infants Hospital of Rhode Island	Community
PCCSB	0994	Palliative Care Consultants of Santa Barbara	Complex
Rutgers	0995	Rutgers Center for State Health Policy	Primary
Sanford	0996	Sanford Health	Primary
Chicago	0997	The University of Chicago	Community
Chicago-M	0997M	The University of Chicago	Community
UOA	0998	University of Arkansas for Medical Sciences	Complex
Cleveland	0999	University Hospitals of Cleveland	Primary
Indiana	1000	Trustees of Indiana University	Disease
Curators	1001	The Curators of the University of Missouri	Community
Curators-M	1001M	The Curators of the University of Missouri	Community
PRHI	1002	Pittsburgh Regional Health Initiative	Complex
CHCS	1003	Center for Health Care Services	Behavioral
GWU	1004	The George Washington University	Disease
Sutter-AIM	1005A	Suttercare	Complex
Vanderbilt	1006	Vanderbilt University Medical Center	Complex
BIDMC	1007	Beth Israel Deaconess (BIDMC)	Complex
Maimonides	1008	Maimonides Medical Center	Behavioral
Penn-MM	1009	The Trustees of the University of Pennsylvania	MMSDM
Carilion	1010	Carilion New River Valley Medical Center (IHARP)	MMSDM
URI	1011	University of Rhode Island	Complex
UIHC	1012	University of Iowa	Complex
BAHC	1013	Ben Archer Health Center	Community
SCRF	1014	South Carolina Research Foundation	Complex
NCCCN	1015	North Carolina Community Care Networks, Inc.	Complex
Penn-DS	1016	The Trustees of the University of Pennsylvania	Disease
Nemours	1017	Alfred I. duPont Hospital for Child NCC-W of the Nemours Foundation	Disease
Duke	1018A	Duke University	Disease
MAHEC	1019	Mountain Area Health Education Center, Inc.	Disease
Ford	1020	Henry Ford Health System	Hospital
Joslin	1021	Joslin Diabetes Center, Inc.	Disease
AGH-PAC	1022B	Atlantic General Hospital	Primary
UAB	1023	University of Alabama at Birmingham	Disease
MPHI	1025	Michigan Public Health Institute	Community
MPHI-M	1025M	Michigan Public Health Institute	Community
UEMS	1026	University Emergency Medical Services, Inc.	Complex
Christiana	1027	Christiana Care Health Services, Inc.	Disease
U-Miami	1028	University of Miami	Community

Intervention Abbreviation	ID	Awardee	FLE
Dartmouth-PE	1029PE	Trustees of Dartmouth College - Patient Engagement	MMSDM
Dartmouth-S	1029SI	Trustees of Dartmouth College - Sepsis	Hospital
Courage	1030	Courage Kenny Rehabilitation Institute	Complex
UVA	1031	The Rector and Visitors of the University of Virginia	Disease
Methodist-DP	1032A	Methodist - Delirium	Hospital
UChicago	1033	University of Chicago	Hospital
Finity	1034	Finity Communications, Inc.	Community
AACISC	1035	The Asian Americans for Community Involvement of Santa Clara	Community
Prosser	1036	Prosser Public Hospital District	Community
Prosser-M	1036M	Prosser Public Hospital District	Community
Texas	1037A	University of North Texas Health Science Center	Complex
Texas SNF	1037B	University of North Texas Health Science Center	Complex
MedExpert-FFS	1038	MedExpert	MMSDM
MedExpert-MA	1038MA	MedExpert	MMSDM
HRA	1039	Health Resources in Action, Inc.	Disease
USC	1040	University of Southern California	MMSDM
Emory	1041	Emory University	Hospital
Northland	1042	Northland Healthcare Alliance	Complex
Ochsner-SC	1043A	Ochsner Clinic Foundation	Disease
Houston-p2	1044A	University of Texas Health Science Center at Houston	Complex
Eau Claire	1045	Eau Claire Cooperative Health Centers Inc.	Community
Le BonHeur	1046	Le Bonheur Community Health and Well-Being	Disease
PBGH	1047	Pacific Business Group on Health	Primary
ICSI	1048	Institute for Clinical Systems Improvement	Behavioral
TransforMED	1049A	TransforMED	Primary
NEU-CHA	1050A	Northeastern University (NE-CHA)	Community
NEU-Lahey	1050B	Northeastern University (NEU-Lahey)	Community
DDHS	1051	Developmental Disabilities Health Services PA	Complex
Feinstein	1052	The Feinstein Institute for Medical Research	Behavioral
J-CHIP	1053A	Johns Hopkins University-CHIP	Complex
J-CHIP-PAC	1053B	Johns Hopkins University-CHIP	Complex
LifeLong	1054	LifeLong Medical Care	Complex
Mt. Sinai	1055	Mount Sinai School of Medicine	Hospital
St Francis	1056A	St. Francis Healthcare Foundation of Hawaii	Complex
St Francis-PAC	1056B	St. Francis Healthcare Foundation of Hawaii	Complex
Christus-Acute	1057A	Christus	Hospital
Christus-LTPAC	1057B	Christus	Hospital
Mineral	1058	Mineral Regional Health Center	Community
Mineral-M	1058M	Mineral Regional Health Center	Community
ValueOptions	1059	ValueOptions Inc.	Behavioral
St. Luke's	1060	St. Luke's Regional Medical Center	Hospital

Intervention Abbreviation	ID	Awardee	FLE
Pharm2Pharm	1061	University of Hawaii at Hilo	MMSDM
Kitsap	1062	Kitsap Mental Health Services	Behavioral
FLHSA	1063A	Finger Lakes Health Systems Agency	Primary
Denver	1064A	Denver Health and Hospital Authority	Primary
Bronx	1065	Bronx RHIO, Inc.	Community
Bronx-M	1065M	Bronx RHIO, Inc.	Community
Imaging Advantage	1066	Imaging Advantage LLC	Community
Imaging Advantage-M	1066M	Imaging Advantage LLC	Community
Tennessee	1067	University of Tennessee Health Science Center	MMSDM
CCC	1071	Foundation for California Community Colleges	Primary
FirstVitals	1072	FirstVitals Health and Wellness Inc.	Disease
PSW-FFS	1073	Pharmacy Society of Wisconsin	MMSDM
PSW-MA	1073MA	Pharmacy Society of Wisconsin	MMSDM
MCMCC	1074	Mary's Center for Maternal and Child Care, Inc.	Community
FPHNY	1075	Fund for Public Health in New York, Inc.	Behavioral

APPENDIX H:
DETAILED QCA RESULTS

Appendix Table H-1
QCA Conditions Included in Analyses

Condition	Definition	Source*
Innovation Characteristics		
Direct Innovation	Patients directly participated or received a discrete set of services.	SQC
New Program	A new innovation program.	SQC
Expanding Program	A pre-existing program that the awardee expanded in reach or scope.	SQC
Multiple Sites	Implemented at two or more geographically or organizationally distinct locations.	AASF2, Q4
Disease Focus	Innovation is targeted and provided to a population defined by a single disease or group of closely related diseases.	SQC
Targets Socially Fragile	Targeted patients, either exclusively or partially, who were characterized as being socially fragile or complex OR at risk for disease progression because of social circumstances or barriers (e.g., homeless)	SQC
Targets Clinically Fragile	Targeted patients, either exclusively or partially, who were characterized as clinically fragile or complex or at risk for clinical disease progression. Often, these populations were characterized as such because of multiple and/or advanced health conditions and morbidities, reliance on specialty care and/or health technologies, or coordination that might be required among multiple medical specialists.	SQC
Targets Children	Exclusively targeted children (ages 0–17).	SQC
Targets Adults	Exclusively targeted adults (ages 18+)	SQC
Targets Elders	Exclusively targeted elders (ages 65+)	SQC
Inclusive Population	Participating patients included children and adults.	SQC
Innovation Components		
Staff Training	Staff received formal training, training via experiential learning, ongoing training, and/or training to accommodate staff turnover.	AASF2, Q12
Implementation Process	Introduced innovation components in a planned and deliberate way; worked with all necessary entities within and across organizations; executed its self-monitoring plan; and/or used a formal improvement framework or change management process (e.g., LEAN, PDSA cycles).	AASF2, Q11
Implementation Planning	Had written protocols and procedures, a detailed timeline with milestones, a comprehensive staffing plan, and/or experience implementing similar programs.	AASF2, Q10

Condition	Definition	Source*
Care Coordination	The innovation provided care coordination, case management, or integrated care in that it combined traditionally separate providers and types of services or organized patient care activities between two or more participants to facilitate the appropriate delivery of health care services.	SQC
Behavioral Health	Included a mental and/or behavioral health component, either exclusively or partially.	SQC
HIT	Included an HIT or telemedicine component (e.g., EHRs, decision-support tools, patient portals).	SQC
Innovation Staff		
Uses CHWs	The innovation used a community health worker who is from and/or knows the community/population with which they are working.	SQC
Uses Any Type of Licensed Provider	The innovation used either licensed clinical staff or licensed independent clinical providers.	SQC
Uses Licensed Clinical Staff	The innovation used licensed clinical staff to deliver or provide services OR sought to influence the care delivered by licensed clinical staff (e.g., dental hygienist, occupational therapist, and paramedic).	SQC
Uses Licensed Independent Clinical Providers	The innovation used licensed independent clinical providers to deliver or provide services OR sought to directly influence the care delivered by licensed independent clinical providers (e.g., dentist, nurse practitioner, and MD/DO).	SQC
Uses Non-licensed Support Staff	The innovation used non-licensed clinical support staff OR sought to influence the care delivered by non-licensed clinical support staff. “Non-licensed clinical support staff” refers to staff working and functioning in legitimate roles and jobs that do not have specific licensure requirements (e.g., aides, treatment options counselor).	SQC

* AASF2: Second Annual Awardee Summary Form; Q# refers to the question number from AASF2; SQC: Structured Qualitative Coding

APPENDIX I:
COMPARISON GROUP METHODS

Because the majority of the interventions being implemented do not use randomization to assign beneficiaries to a treatment or control group, FLEs face a major challenge in constructing comparison groups that can serve as a counterfactual to those receiving the intervention. FLEs are using a variety of approaches to construct such comparison groups, and in this section we explore the methods being used across the HCIA interventions.

Most FLEs are using propensity scores (PS) to construct their comparison groups. A propensity score is the probability that an individual receives the intervention conditional on observed covariates (Rosenbaum & Rubin, 1983). The propensity score summarizes the observable characteristics that might affect treatment status into a single probability (Rubin, 1997). Their distribution can be used to create the comparison group. For the HCIA interventions, propensity score matching (PSM) and propensity score weighting (PSW) are the primary ways that propensity scores are being used. PSM involves matching intervention group members to comparators with the closest PS in the comparison pool or to comparators whose PS falls within a permissible range (sometimes called a caliper). PSW uses PSs to differentially weight the members of the comparison and treatment groups.

Table I-1 shows the various methods of comparison group construction being used by the FLEs for all 135 unique interventions. In the table, we distinguish six primary methods for comparison group formation:

1. 1-to-1 PSM (beneficiary level). One-to-one propensity score matching at the beneficiary level is using propensity scores to match one beneficiary in the treatment group to one comparator.

2. 1-to-many PSM (beneficiary level). One-to-many propensity score matching at the beneficiary level is using propensity scores to match one beneficiary in the treatment group to more than one comparator when possible. This includes FLEs that are matching one beneficiary in the treatment group to a variable number of comparators. For example, the FLE for the Community Resources group matches up to three comparators to each treatment group beneficiary, depending on the availability of comparators lying within the specified caliper for each treatment group beneficiary.

3. PSM (higher level). Propensity score matching at a higher aggregation level is using propensity scores to match at the region, provider, physician, facility, clinic, or hospital level. Eight interventions used this method, three used 1-to-1 higher level PSM and five used 1-to-many PSM.

4. PSW (beneficiary level). Propensity score weighting at the beneficiary level.

5. RCT. A few HCIA interventions are randomized controlled trials (RCTs) with a control group. For these, the FLEs are using the randomly selected control group as the counterfactual.

Some comparison group formation methods fall outside of these five designations, and we describe them as "Other" in ***Table I-1*** and in more detail below. Finally, FLEs have not been able to construct a comparison group for 36 interventions.

The constructions of nine other comparison groups have been classified as Other. Three interventions (Mineral and Mineral-Medicaid from the Community Resources Group, PeaceHealth from the Primary Care Redesign group) involve matching at the provider/facility level without the use of propensity scores. The comparison group for URI from the Complex group is being constructed using exact matching at the beneficiary level on gender, age, race/ethnicity, index month, group home status, dual eligibility, and Hierarchical Condition Category (HCC) risk score. A few interventions are being evaluated with comparison groups that are treatment as usual groups (ValueOptions in the Behavioral Health group) or groups that were eligible for services but declined or did not receive services (Children's Hospital in the Community Resources group and SafeMed in the MMSDM group). For CLTCEC in the Complex group and Prosser in the Community group, propensity score adjustments were not used.

The most common comparison group construction method being used is PSM with 1-to-many matching. Most FLEs favor a particular approach, but there are many instances in which a FLE used multiple approaches. For example, the FLE for the Disease Specific evaluation is mostly using 1-to-1 PSM at the beneficiary level whereas the FLE for the Primary Care Redesign evaluation is mostly using 1-to-many PSM and PSM at the facility/provider level. The FLE for the Hospital-setting group is exclusively matching at the provider level where comparison providers are chosen from within the HCIA awardee hospital referral regions and matched on provider characteristics without the use of PSs. These decisions are likely driven partially by FLEs preferences, the availability of comparison beneficiaries, and their perceptions of meaningful differences at the facility/provider level. Evaluations of interventions that span multiple facilities or providers in which service delivery protocols or procedures vary would probably want to use 1-to-1 or 1-to-many upper level PSM to account for facility/provider variation. It is also likely that FLEs with access to a large pool of comparison beneficiaries would be more likely that those with a few comparators to use 1-to-many PSM.

Table I-1
Frequency of Comparison Group Methods Used by FLE

FLE	1-to-1 PSM (beneficiary level)	1-to-many PSM (beneficiary level)	PSM (higher level)	PSW (beneficiary level)	RCT	Other	No comparison group
Behavioral	—	3	—	—	—	1	6
Community	—	25	3	1	—	4	5
Complex	11	—	—	7	1	2	7
Disease Specific	7	—	—	2	—	—	10
Hospital-setting	—	—	—	—	—	11	1
MMSDM	6	—	—	—	3	1	3
Primary Care	—	4	5	—	1	1	4
All FLEs	24	32	8	10	4	20	36

Since propensity score methods, either through matching or weighting, are the most common techniques being employed to create comparison groups, we further investigated the types of diagnostics that FLEs reported using to analyze the quality of their comparison groups created by using propensity score techniques. This is summarized in ***Table I-2***.

Table I-2
Diagnostics Used by FLEs to Assess Quality of Comparison Groups Created with Propensity Scores

FLE	Balance Table	PS Plots	Other techniques
Behavioral	✓	✓	✓
Community	✓	✓	
Complex	✓	✓	
Disease	✓	✓	
MM/SDM	✓		
Primary Care	✓		✓

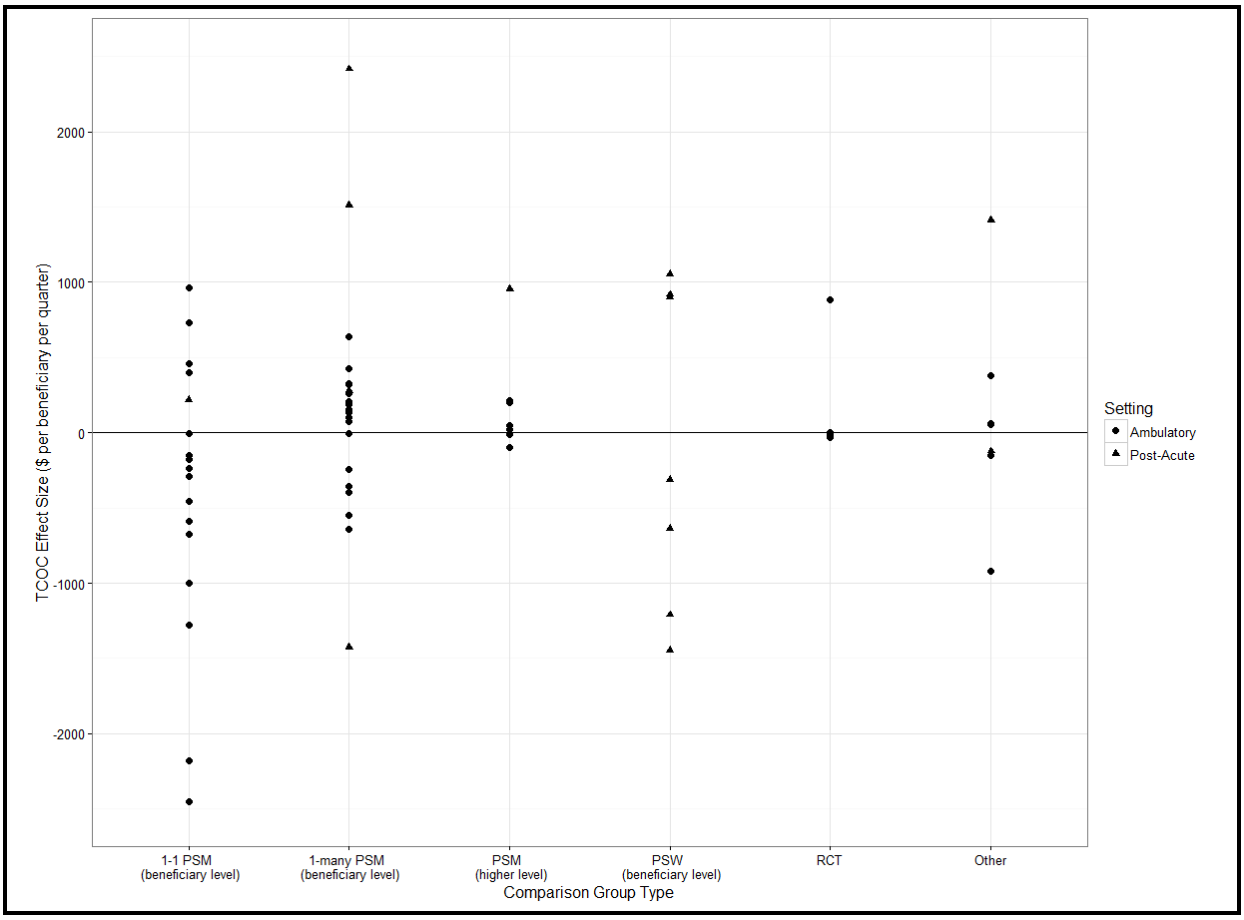
All of the FLEs using propensity score methods reported balance tables. At the minimum, these tables reported on the means and standard deviations of variables in the propensity score models and the standardized mean covariate differences between the treatment and comparison groups before and after the intervention. Ideally, the standardized mean group difference for each variable should be small; less than 0.1 is a common cutoff used in practice. FLEs reported the cutoffs they utilized. Four of the FLEs included plots of the propensity scores to visually show the overlap in propensity scores between the intervention and comparison groups. Strictly

speaking, causal effects should only be tested for treatment group members whose propensity scores overlap with those of comparators. Finally, the evaluator for the Primary Care Redesign group provided the outcomes from an omnibus test for balance on matching variables; the evaluator for the Behavioral Health group also reported outcomes from the omnibus test for balance on matching variables as well as the frequencies of the number of matches made per treatment group beneficiary.

Assessment of Comparison Group Methods and TCOC Effect Sizes

We were also interested in whether any particular comparison group construction method led to systematically larger or smaller TCOC effect sizes. To assess this, we plotted TCOC effect sizes by their comparison group formation method (***Figure I-1***). We considered all interventions with TCOC effect sizes except those that were reported by the FLE to be unreliable (see section 3.1 for further details regarding FLE-reported unreliable estimates). We did not include TCOC effect sizes from the hospital-setting group because the FLE is reporting TCOC with a 60-day lookback rather than a 90-day lookback and hence those effect sizes are not comparable to those from other HCIA groups.

Figure I-1
TCOC Effect Sizes by Comparison Group Construction Method



From visual inspection, all methods had both favorable and unfavorable TCOC effects, with mean effects clustering close to zero. There was no evidence that comparison group method systematically biased results in either direction.

To formalize our analysis of TCOC effect sizes and comparison group construction methods, we also conducted a one-way analysis of means to test if on average the effect sizes yielded by the different methods are the same. A significant test would indicate that at least one method leads to different effect sizes on average that cannot be explained as being the result of statistical noise. The test did not assume the homogeneity of variances (Welch, 1951). The result from this test ($F = 0.22$, $p = 0.95$) supports our conclusion from the visual inspection of ***Figure I-1***, that no particular comparison group method produces systematically different effect sizes from those produced by other comparison group construction methods.

From our analysis of FLE comparison group construction, we found that FLEs are mostly using propensity score methods and using the same method to evaluate most of their awardees. Furthermore, we found no systematic differences in the mean of total cost of care effect sizes across comparison group methodologies. For meta-evaluation this means estimated HCIA effects are unlikely to be due to differences in evaluation methodology but rather due to features of the interventions themselves. We explore these features and their effect on TCOC in section 3.4.

[This page intentionally left blank.]

APPENDIX J:
SYNTHETIC COMPARISON GROUPS

In the FLEs' latest internal reports, there were still four awardees lacking comparison groups for which a FLE (NORC) reported HCIA effects based only on pre-post data. The pre-post approach relies heavily on the assumption that that a baseline trend will persist unchanged throughout the post-intervention period. We have previously shown that pre-post analyses are likely to produce biased estimates of true innovation effects in HCIA. For this report, we developed another method to evaluate the accuracy of pre-post estimates by constructing synthetic comparison groups. The resulting groups are "synthetic" in the sense that they do not represent actual data for any particular awardee but are instead an amalgam of trends across many different comparison groups.

The first step in the process was to determine what comparison group trends looked like for awardees that did have comparisons and how consistent these trends were across awardees. The analysis was limited to ambulatory innovations after removing several awardees with large spikes in their baseline trajectories. We estimated quarterly trends using the following time-series model:

$$Y_t = a + b_1Q_t + b_2POST_t + b_3Q_t*POST_t + e$$

where

- Yt = Total cost of care at quarter t
- a = the intercept term
- Qt = quarter number at corresponding to t
- POSTt = a 0/1 indicator for whether quarter t occurs during the post-intervention period
- t = quarter number ranging from 1 to 8 during the pre-intervention period and from 9 to 18 during the post-intervention period
- e = an error term.

The coefficients in the model estimate a linear time trend (b_1), a shift in the level of the post-intervention series (b_2), and a change in the slope of the trend during the post-intervention period (b_3). ***Figure J-1*** overlays the trajectories for each of the 34 comparison groups included in the analysis. The thick black line shows the estimated mean slope for total costs which was $103 per quarter during pre-intervention and declined to $69 per quarter during the post-intervention period. The trends are largely consistent from one group to another, showing relatively flat expenditure trajectories over time for all but the highest cost innovations. It appears that many comparison groups could be exchanged for one another without greatly affecting estimates of HCIA impacts.

In the next phase of the analysis, we generated a synthetic comparison group for each of the four pre-post only awardees based on the common comparison group trajectories. Cost levels were aligned for each awardee by setting the comparison group's last baseline quarter equal to the observed cost in that quarter for the HCIA group. The estimates for the comparison group slopes were then applied to generate a predicted comparison cost for each quarter of the analysis. Using these predicted values as a synthetic comparison group, a second time-series model was estimated for the combined HCIA and comparison group data. The second model included terms for whether the estimate was for the HCIA or the comparison group and an interaction term identifying HCIA post-intervention estimates. The interaction term provides an estimate of the HCIA innovation's effect on TCOC relative to the expected cost in the absence of the innovation, as estimated by the synthetic comparison group. We used Stata's group function to obtain robust standard errors for the time series. One important limitation of the synthetic approach is that cost variation is artificially suppressed because the predicted comparison group values do not embody the quarterly fluctuations that one would normally see in these trajectories.

Figure J-1
Overlay Plot and Regression Line for 34 Ambulatory Innovation Time Series

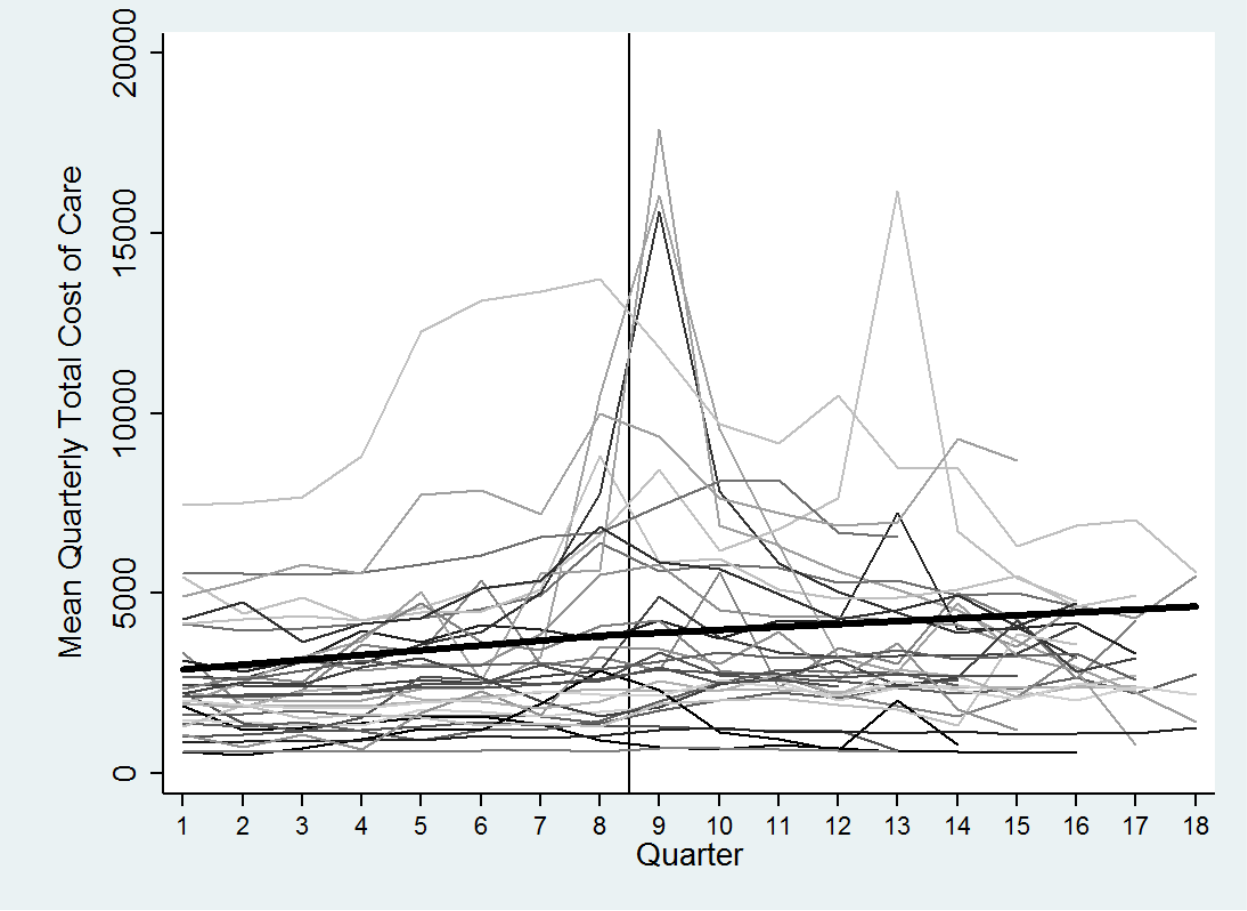


Table J-1 contrasts the pre-post impact on total cost of care reported by the FLE with the estimate derived using the synthetic comparison group for each of the four innovations. Pre-post

estimates were estimated using a beneficiary-level model containing an indicator for the post-intervention period. Beneficiary characteristics were controlled in the model. In the FLEs' pre-post analyses, all four innovations had positive innovation coefficients indicating that total costs were higher on average during the post-intervention period than during baseline. But without a comparison group, we are unable to assess whether this increase is due to the pre-existing trend over time in the outcome.

The synthetic comparison group paints a dramatically different picture. The synthetic-based coefficients differ by more than $500 and by as much as $1,500 from their pre-post counterparts. The two largest FLE dissavings estimates are cut by more than half, and the results for Johns Hopkins School of Nursing shift to significant savings. With one exception, the standard errors for the synthetic comparison estimates are slightly smaller than those for the beneficiary-level pre-post analyses.

Table J-1
Comparison of Pre-Post Only and Synthetic Comparison Group Estimates of Innovation Effects on Quarterly Costs of Care for Four Ambulatory Setting Innovations

Innovation	Pre-Post Only Coefficient ($) (SE)	Synthetic Comparison Coefficient ($) (SE)
Johns Hopkins School of Nursing	253	-1,344*
	(617)	(587)
South Carolina Research Foundation	822	368
	(638)	(502)
Mountain Area Health Education Center, Inc.	308	-468
	(520)	(439)
University of North Texas Health Science Center-A	1,516*	605
	(202)	(676)

We do not know how accurate the estimates based on the synthetic comparisons are. The available time series are comparatively short, and could be susceptible to unusual values for just a few quarters. As noted earlier, the synthetic approach also suppresses the quarterly fluctuations that would be normally be observed. Nonetheless, the coefficient discrepancies make it clear that different methods can yield substantially different results for pre-post data.

As a result of these differences, we decided not to include any pre-post only estimates in the meta-analyses for this report. This is the third method we have employed since the beginning of the project to assess the validity of pre-post effects. All three methods have uncovered potential biases attributable to pre-post data. If FLEs continue to be unable to construct comparison groups in the near future, we will consider creating synthetic comparison groups for them for the project’s final report.

APPENDIX K:
BAYESIAN META-ANALYSIS

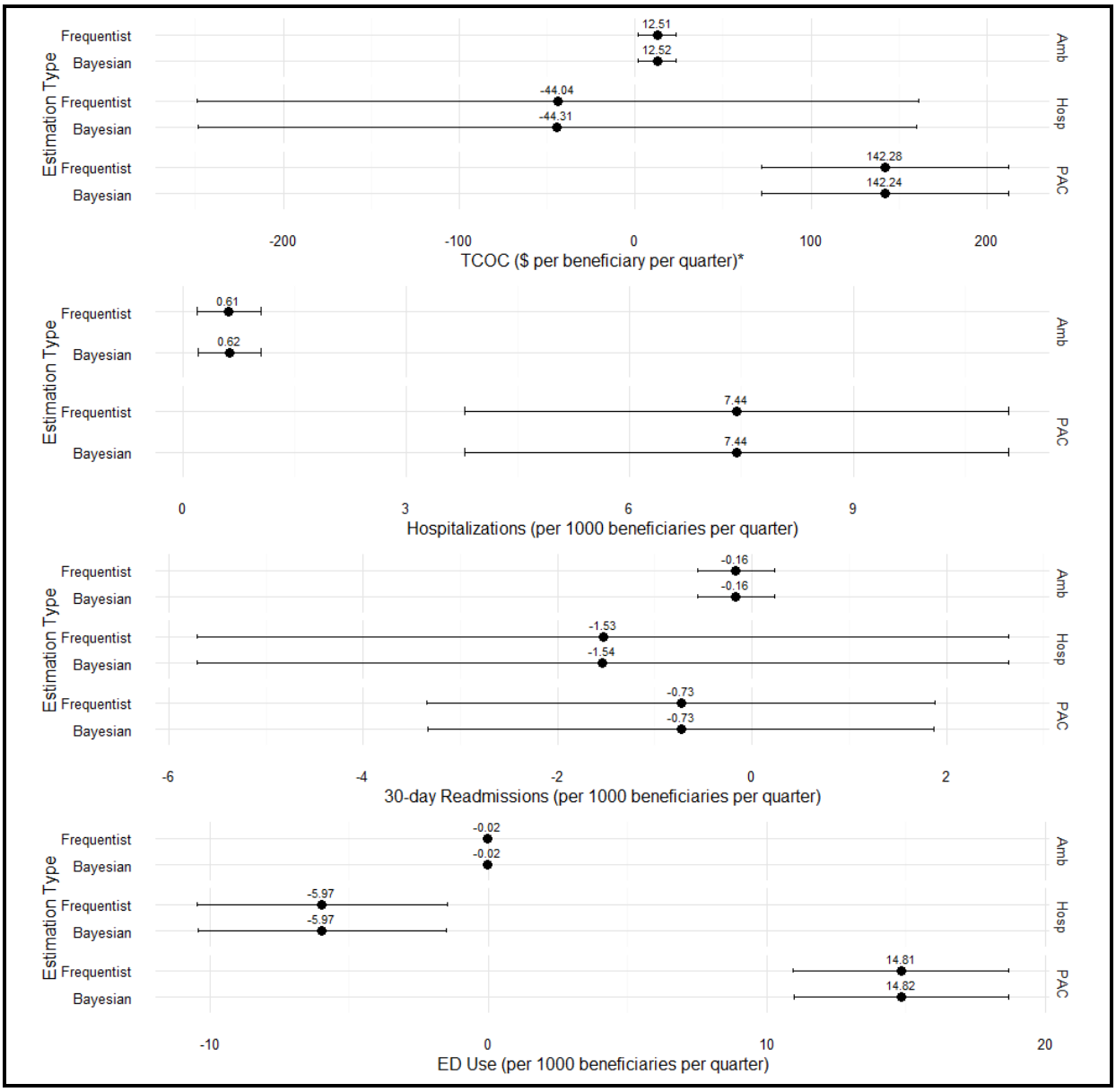
To complement our meta-analytic findings in section 3.2.1, in this section we present a Bayesian fixed-effects meta-analysis. In addition to providing a sensitivity analysis for our grand means findings in section 3.2.1, this analysis yields all the benefits of Bayesian estimation such as having a probabilistic interpretation and lays the foundation for future Bayesian efforts in our meta-evaluation.

Bayesian analyses, in contrast to traditional (or frequentist) analyses, begin with the assumption that parameters are characterized by distributions rather than as fixed unknowns. Before observing the data, the parameters of interest are described by the prior distribution. The prior distribution summarizes all of our initial uncertainty about the parameters. The prior distribution is updated by the data that enters into the model through the likelihood function. The updated distribution for the parameters is called the posterior distribution, or posterior for short. The posterior summarizes all of the uncertainty about the parameters that remains after observing the data. From the posterior, the Bayesian analogues of point estimates can be obtained (usually by taking the mean of the marginal posterior distribution for the parameter of interest). Credible intervals, which are the Bayesian analogues of confidence intervals, can also be obtained from the posterior. One of the most powerful features of Bayesian analysis is the more natural interpretation of the credible intervals. For a 90 percent credible interval, the probability that the parameter lies in the interval is 90 percent. This is in contrast to the interpretation of traditional confidence intervals in which 90 percent confidence indicates the probability that the method used to generate the interval captured the parameter value. This ease of interpretation extends to other probabilistic questions about the parameters of interest which we explore later in this section.

We conducted a Bayesian fixed-effects meta-analysis. This model is the Bayesian analogue of the model used in section 3.2.1. Specifically, we assumed that the DiD effect sizes reported by the FLEs are independent and drawn from distributions with a common mean, the grand mean, and with variances equal to the square of the reported DiD standard errors. We augment this model with a prior distribution for the grand mean that is normal with mean 0 and standard deviation 1,000. This prior is a non-informative prior because it “spreads out” our prior beliefs over a large range of possible grand mean values. We present our findings in ***Figure K-1***, which compares the grand means and 90 percent confidence intervals obtained from the results in section 3.2.1 (labeled “Frequentist”) to the grand means and 90 percent credible intervals from the Bayesian analysis.

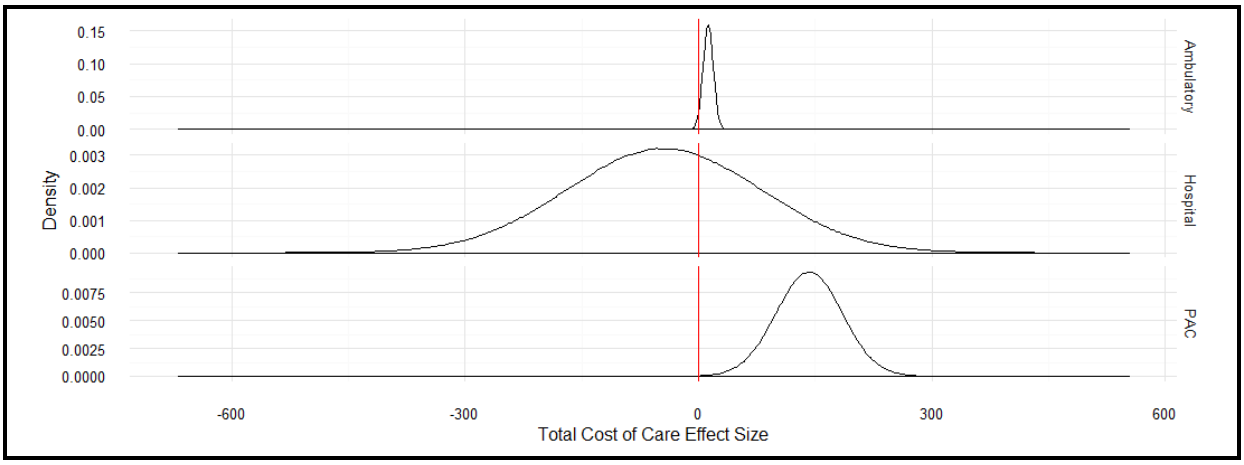
Across the ambulatory, post-acute, and hospital settings and across all core measures, the Bayesian analysis yields results nearly indistinguishable from the frequentist findings. This similarity is what we would expect given the model chosen for the Bayesian analysis and further increases our certainty in the results of our meta-analysis.

Figure K-1
Comparison of Frequentist and Bayesian Grand Means and 90% Confidence/Credible Intervals



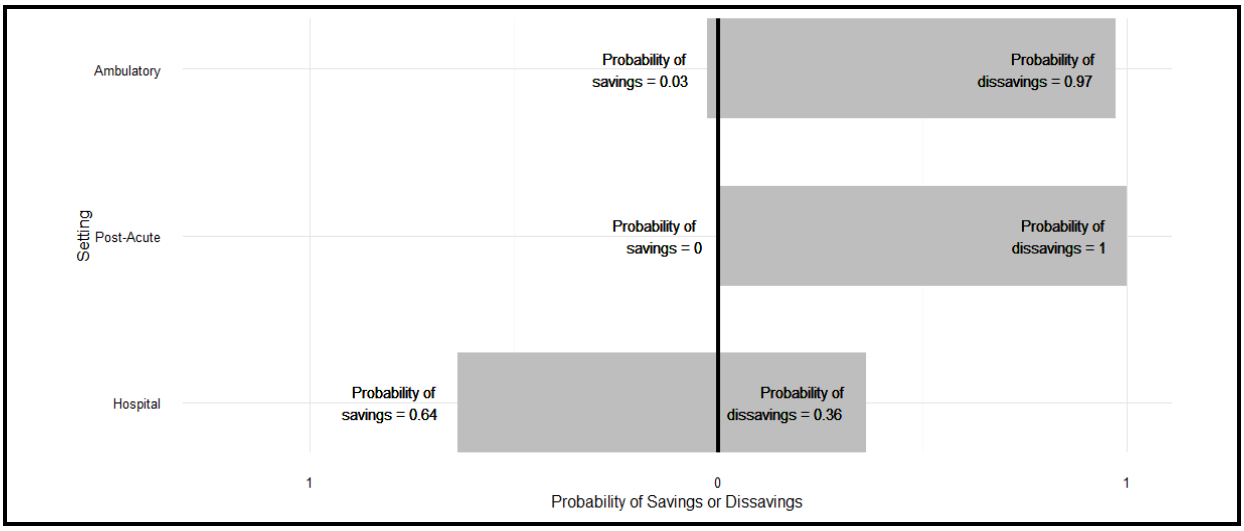
Because Bayesian inference allows for the probabilistic interpretation of findings, we also considered the probability of awardees in each of the settings reducing costs (***Figure K-2 and Figure K-3***), that is we used the posterior distribution for the TCOC grand mean to compute the probability of the TCOC grand mean being greater than 0 and the probability that it is less than 0.

Figure K-2
Posterior distribution of the grand mean for Total Cost of Care



In ***Figure K-2***, possible values of the total cost of care grand mean are given along the horizontal axis. The height of the density curve is highest for values of the total cost of care grand mean that are the most likely and lowest for those that are less likely. The vertical red line demarcates savings (negative values) and dissavings (positive values). The probability of savings and dissavings for each setting are summarized in ***Figure K-3***. In the ambulatory and post-acute settings, the probability that the pooled mean across interventions is less than zero (indicating savings) is small. Because we started with a prior centered at 0, we assume that absent evidence to the contrary, there is a 50 percent chance of savings and 50 percent chance of dissavings. Informed by the data, the probability of savings in each of these settings shifted considerably. In the hospital setting, the probability that the grand mean is less than zero is 0.64 and the probability that it is greater than zero is 0.36. While a 64 percent chance of savings seems substantive, we must also keep in mind that the model started with equally likely chances of savings or dissavings.

Figure K-3
Probability of Savings or Dissavings in Ambulatory, Post-Acute, and Hospital Settings



In the traditional fixed-effects meta-analysis model presented in section 3.2.1, the intervention-level DiD effects were weighted by their uncertainty (their variances). Analogously, in the Bayesian fixed-effects model we have estimated, uncertainty in the intervention-level DiD effects enters the model through the likelihood function and more certain estimates (lower variances) contribute to the grand mean estimates more than less certain estimates. Typically, larger variances are observed in smaller interventions while smaller variances are observed in larger studies. A concern might be that this inverse relationship between intervention size and variance means that the larger interventions are driving the grand mean estimates, however, as reported in section 3.2.3, we found no evidence of sample size bias, giving us more confidence in our results. Finally, as with the frequentist findings, the grand means we report characterize groups of interventions (in our case they characterize interventions that have been classified as ambulatory setting, post-acute setting, or hospital-setting), and should not be taken as evidence for or against any particular HCIA intervention.

We present the Bayesian findings on savings across interventions in each setting to demonstrate the power of Bayesian analysis. If a critical level of savings or dissavings were to be determined to be of practical significance, then Bayesian methods could be used to determine the probability of whether that critical level was achieved or not. For example, if $50 per beneficiary per quarter was determined to be a target level of savings, Bayesian estimation could be used to determine the probability that an intervention achieved that level of savings. In our case of meta-evaluation, we have used Bayesian estimation to determine the probability of savings or dissavings across interventions in the same setting. This type of probabilistic assessment is not possible in traditional (frequentist) estimation, which relies on binary tests of significance (either reject the null hypothesis or fail to reject the null). Because Bayesian techniques are powerful and extremely flexible, we will continue to explore their use to better understand and describe the HCIA interventions. In particular, we will analyze the robustness of our meta-analytic findings under different model specifications (e.g. assuming effects have a t-distribution) and investigate Bayesian meta-regression.

APPENDIX L:
TIME-SERIES ANALYSES

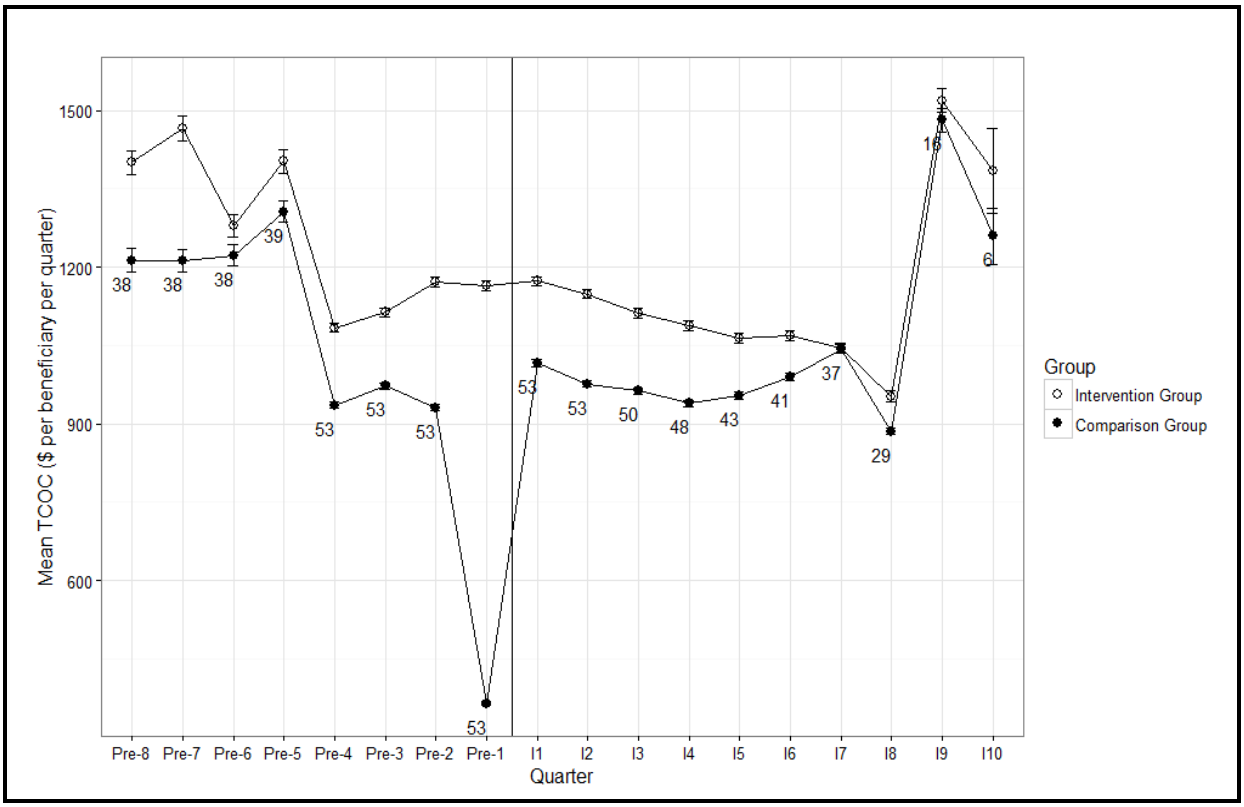
To better understand the trajectories of individual interventions and to assess the stability, change, and impact of outliers on our estimates of TCOC, we conducted analyses using the quarterly means and variance estimates for the treatment and comparison groups of each intervention provided by the FLEs. Using these estimates of the TCOC we plotted the mean and 90 percent confidence intervals associated with each intervention and their comparison group (see Appendix D). These plots highlight the great variability in the trajectories of both intervention and comparison groups and the variation in magnitude of TCOC, both within and across interventions and across ambulatory, post-acute, and hospital settings. Although trajectories of many interventions and their comparators show relatively stable trends in TCOC (both before and after innovation implementation), such stabilities appear to be strongly moderated by sample size (as indicated by the confidence intervals surrounding each TCOC estimate) and deviations from general trends are not uncommon.

Mean Performance on Total Cost of Care Between Awardees and Comparators Over Time

The variability in performance on TCOC by individual interventions makes discerning patterns difficult. Because DiD estimates are derived equally from both awardee and comparator performance, we began our analysis by using meta-analytic techniques to pool TCOC evidence within each quarter across both intervention and comparison groups. This was accomplished by creating grand means (means weighted by their variance) and 90 percent confidence intervals for both the intervention groups and their comparators within each quarter and plotting each group's separate trajectory over time. Interventions included in this analysis were those with quarterly means, standard deviations, and sample sizes supplied to RTI by the FLEs. Additionally, we only included those interventions with data for both the comparison group and intervention group, and we excluded those interventions that served unique populations (unique populations are defined in section 3.1).

Figures L-1 and ***L-2*** display the quarterly weighted average TCOC in dollars for beneficiaries in the intervention group and their comparators for interventions in the ambulatory, post-acute, and hospital settings. This information is presented in tabular form in Appendix D. The average grand means for the intervention group is presented with open points while comparator TCOC is presented with shaded points. Each grand mean is accompanied by its 90 percent confidence interval. The number of estimates contributing to each grand mean is provided below each comparison group estimate. For each plot, we only show quarters in which at least five interventions contributed to the quarterly weighted mean estimate. Quarters preceding the intervention are labeled "Pre-1," "Pre-2," and so on, where "Pre-1" is the first quarter preceding implementation, "Pre-2" is the second quarter preceding implementation, and so on. Quarters during the intervention are denoted by "I1," "I2," and so on, where "I1" is the first quarter after implementation, "I2" is the second quarter after implementation, and so on.

Figure L-1
Weighted TCOC Means for Ambulatory Setting Interventions (90% CIs)



Relative to the instability observed in individual studies, it can be observed that for the summary estimates across interventions, there is considerable stability within groups, especially for those quarters containing data from more awardees. That is, within quarters, estimates among awardees and comparators are relatively similar to each other although comparison estimates are consistently less than those of awardees. For the ambulatory care setting, both treatment and comparison groups show relatively stable trends in costs, with the greatest stability observed during the quarters for which the greatest number of interventions are contributing to our estimates. One notable exception is a very large drop in the average total cost of care for the comparison group that occurs between the second quarter (Pre-2) and the immediate quarter (Pre-1) preceding intervention implementation; a commensurate rise in costs is observed between the immediate quarter preceding implementation (Pre-1) and the first intervention quarter (I1). Since these data have been centered on the awardee's implementation quarter, it is hypothesized that the observed drop is due somehow to methodology, that is, a function of how the comparison groups are identified by the FLEs. Also evident in these findings is the strong influence of sample composition on the observed finding. Estimates further from the implementation quarter (I1) show greater instability and markedly different point estimates than those based on data from more awardees.

For post-acute setting interventions, we observe an increase in costs for both the comparison and intervention groups in the quarters following the onset of the intervention that

appears to stabilize after the seventh intervention quarter at a higher level than in the baseline (***Figure L-2***). Despite the sharp rise in mean TCOC over time, the cost trend for the comparison group tracks closely with the trends for the intervention group.

Figure L-2
Weighted TCOC Means for Post-acute Setting Interventions (90% CIs)

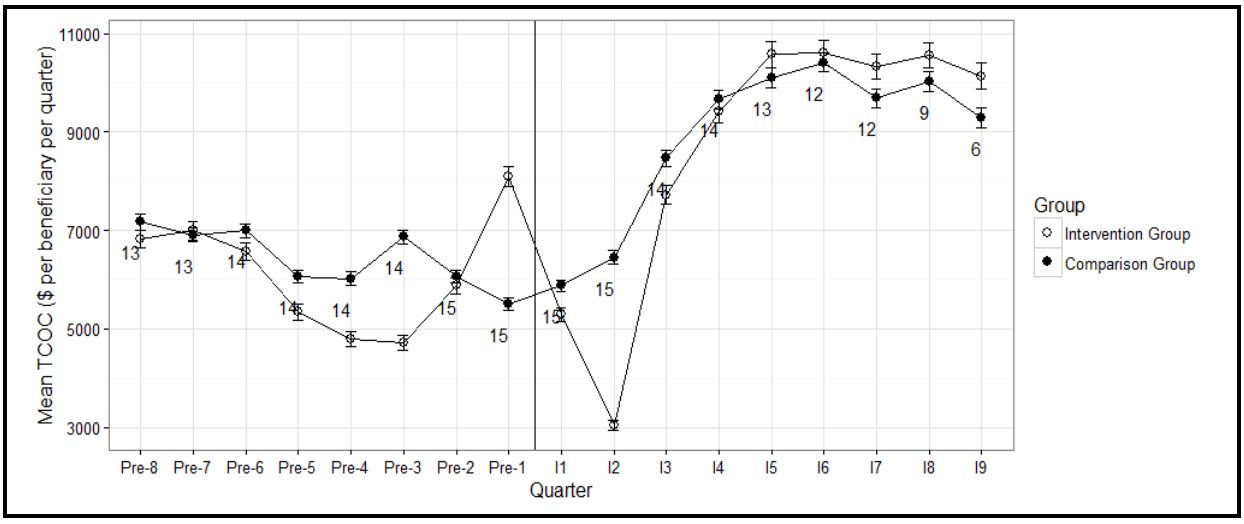
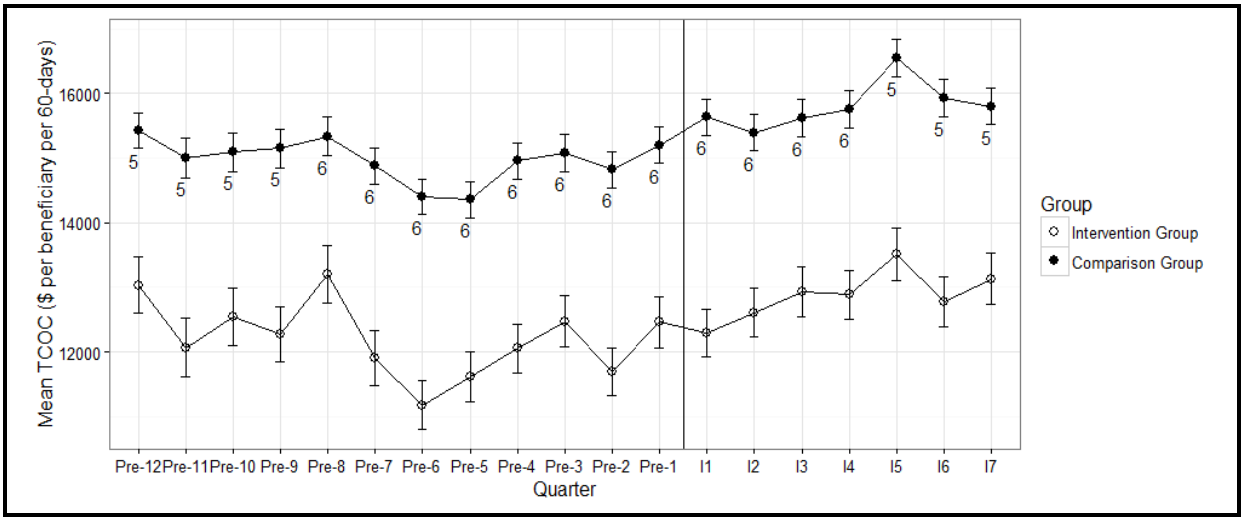


Figure L-3 presents the weighted TCOC means for the intervention group and comparison group in the hospital-setting. In each quarter, mean spending in the comparison group is greater than that of the intervention group. Nonetheless, the shape of the cost trend for the intervention group and comparison group is generally consistent between both groups.

Figure L-3
Weighted TCOC Means for Hospital Setting Interventions (90% CIs)



These descriptive analyses show that in the pre-intervention period comparison group costs are consistently and, with few exceptions, significantly lower than those of awardees. For all but the post-acute awardees, these estimates remain below those of awardees in the post-intervention period. These figures also highlight the critical role awardee composition can play in generating HCIA award findings. Particularly for ambulatory care, the mean estimates for the earliest and most recent quarters are based on fewer awardees than the mean estimates nearer to the implementation quarter. Costs for these fewer awardees (and their comparators) are notably higher than for the entire sample for which we have data.

Mean Difference Between Intervention and Comparison Groups Over Time

We then created quarterly effect sizes estimating the difference in TCOC within each quarter by awardee and each estimate's associated standard error. The difference score is calculated by subtracting the mean of the comparison group from the mean of the intervention group and dividing that result by their pooled standard error. Expressed as a *d*-score (Hedges & Olkin, 1985), positive values represent higher TCOC for the intervention group relative to the comparison group, while negative values indicate TCOC in the quarter were higher for the comparison group. This approach yields effect sizes in terms of dollars per beneficiary per quarter (or per 60-days in the case of the hospital setting interventions).

These quarterly mean scores were then synthesized within and between each quarter using random effects (RE) modeling, which accounts for both differences between intervention effect sizes due to statistical noise and underlying differences between the interventions. Although RE models yield larger standard errors, we believe that this is the appropriate model since within each quarter we are synthesizing estimates from interventions that are similar in setting but not necessarily similar in other ways. In meta-analysis, there is always a concern that large studies reduce the contribution of smaller studies or that the presence of outliers may potentially bias results. To confirm that results were not unduly biased by such considerations, sensitivity analysis was conducted contrasting weighted and unweighted results, and contrasting median and mean estimates (see Appendix D, Tables D-3 to D-6). These analyses confirmed that that large interventions and outliers had minimal impact on quarterly grand means.

The random effects quarterly difference time-series results for ambulatory, post-acute, and hospital-setting interventions are provided in ***Figures L-4, L-5,*** and ***L-6***. In each graph, the difference in TCOC displayed on the vertical axis, the quarter to which the effect size applies is displayed on the horizontal axis (using the same naming convention as before for the quarters preceding the intervention and the quarters during the intervention), and each intervention's difference in TCOC relative to its comparator is represented by a dot whose size is proportional to the number of beneficiaries in the intervention group during that quarter. Effect sizes above the y-axis zero line indicate the intervention group's TCOC was greater than its comparator, while estimates below the line indicate the comparison group's costs were greater. Below each grand mean and 90 percent confidence interval (indicated by the trend line and error bars) is the number of awardees contributing to each grand mean estimate. The figures for ambulatory and post-acute interventions (***Figure L-4*** and ***Figure L-5***, respectively) display two plots: the top plot displays the effect sizes for all interventions while the bottom plot focuses in on those effect sizes closest to the RE grand means.

Figure L-4
Quarterly Mean Difference between Awardee and Comparison Groups: Ambulatory Interventions

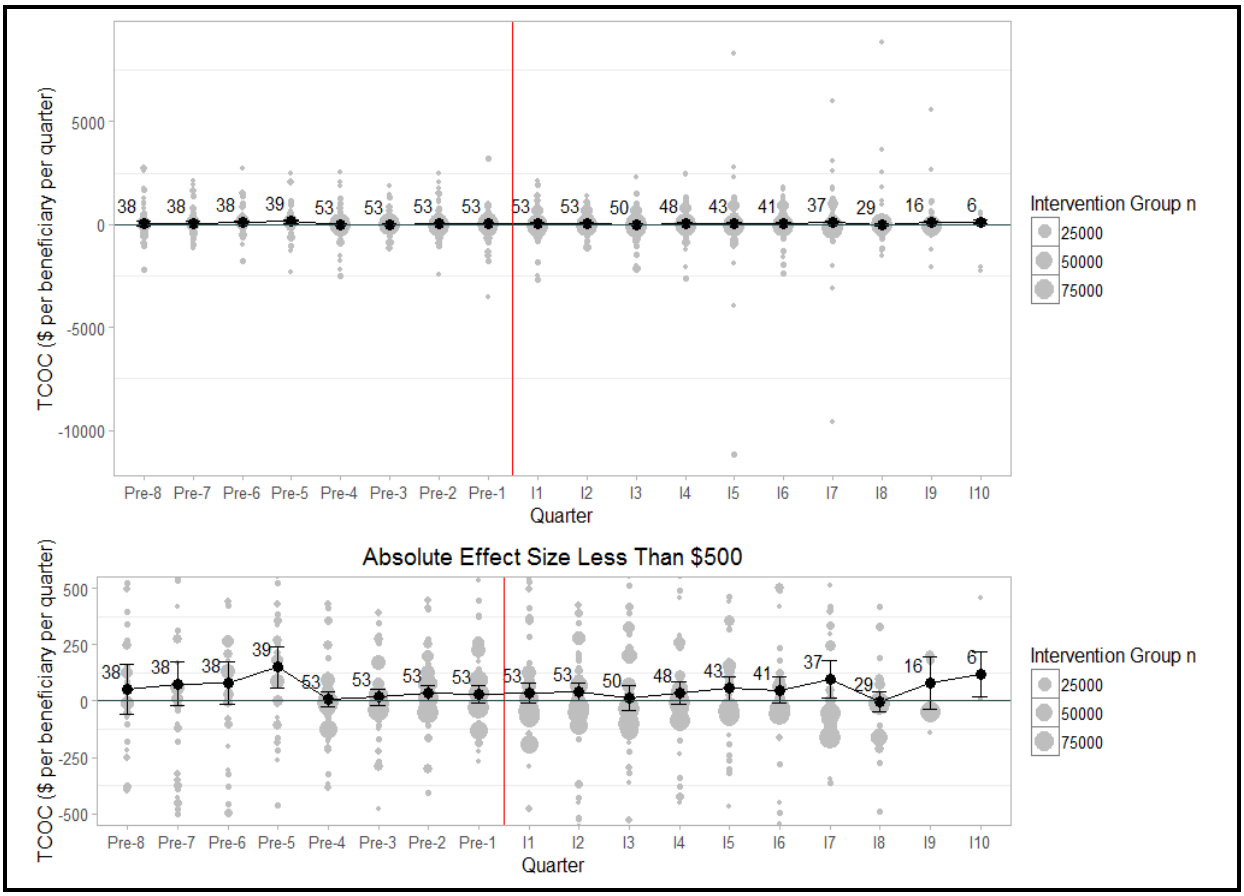


Figure L-5
Quarterly Mean Difference between Awardee and Comparison Groups: Post-acute Interventions

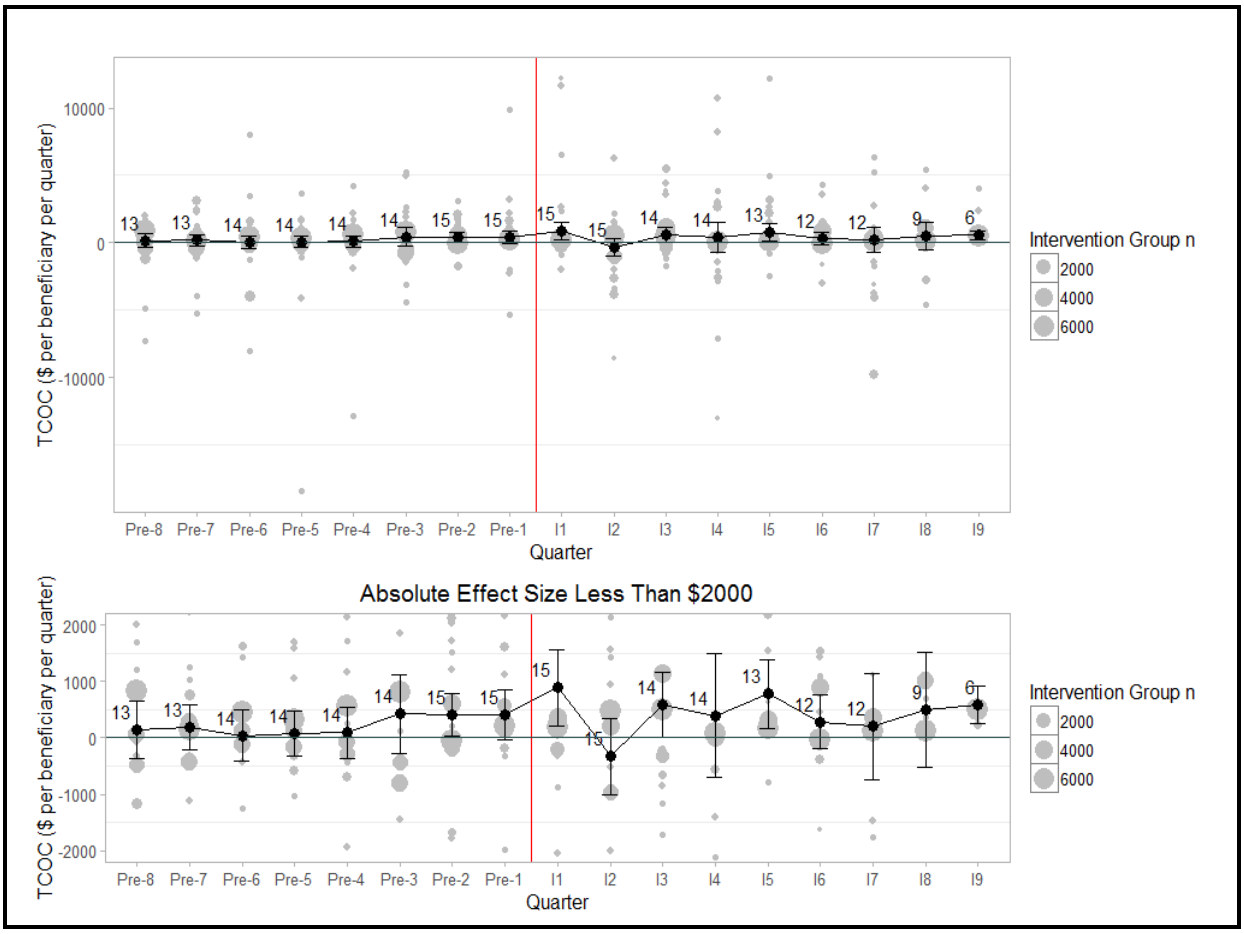
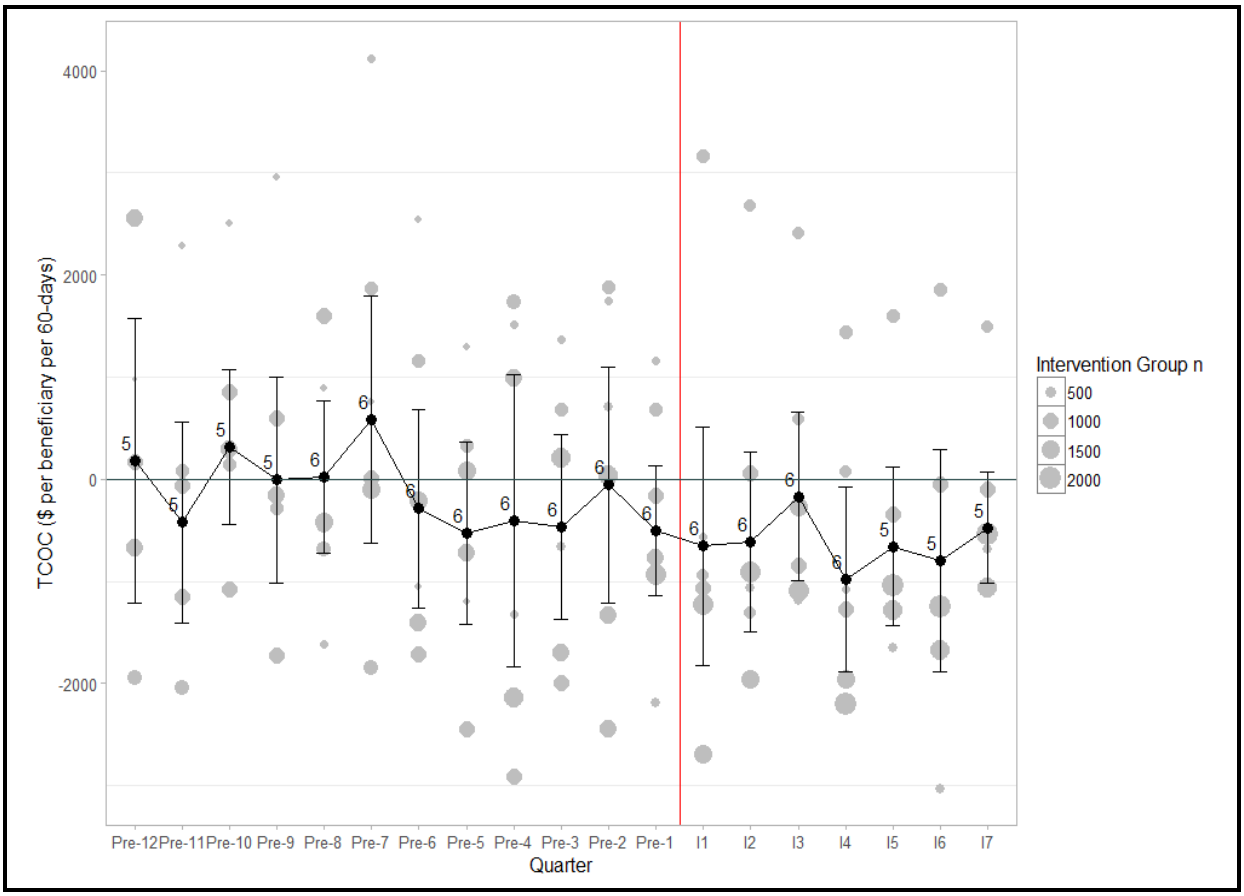


Figure L-6
Quarterly Mean Difference between Awardee and Comparison Groups: Hospital Interventions



A visually striking observation is how the number of interventions included in each quarterly grand mean and the number of enrollees in each estimate affect the grand mean estimate. The highest density of estimates can be observed in the quarters immediately preceding and following innovation implementation. Grand means further from the implementation line are based on fewer estimates and smaller samples. These grand means show greater instability and have larger confidence intervals than those based on the full sample of awardees. Furthermore, as in section 3.2.3, we observe that extreme effects tend to occur for interventions with the smallest sample sizes. Interventions with larger sample sizes tend to have smaller effects and drive the magnitude of the grand means. Combined, these observations highlight how having more studies in a synthesis improves the stability of the effectiveness estimate obtained.

Finally, we do not observe in the quarterly grand means a pattern of savings or dissavings for the ambulatory or hospital settings over time. While a few quarters show significant dissavings (I5, I7, and I10 for ambulatory interventions) or savings (I4 for hospital interventions) at the 90 percent confidence level during the intervention period, no sustained trend is observed. Rather in most quarters, the estimated grand mean is not statistically different from zero. Among

post-acute interventions, nearly half of the intervention quarters (I1, I3, I5, and I9) showed significant dissavings at the 90 percent confidence level; in all intervention quarters with the exception of I2, the estimated quarterly grand means are greater than zero.

Overall, our findings generated from the quarterly unadjusted means for the comparison and intervention groups are consistent with our findings from section 3.2 in which we used summative DiD estimates for meta-analysis. Although the use of summative DiDs may raise concerns that as a single estimate it may be masking underlying effects, our quarterly meta-analysis findings echo rather than contradict our meta-analytic efforts using the summative DiDs. This holds true in terms of the dispersion of effects observed, the influence of sample size on effect size, and savings/dissavings across each of the settings, further supporting our findings throughout this chapter.