MATHEMATICA Policy Research

REPORT

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

Addendum to Third Annual Report

August, 2017

Greg Peterson
Laura Blue
Keith Kranker
Sandi Nelson
Andrew McGuirk
Joe Zickafoose
Andrea Wysocki
Eric Lammers
Randall Blair
Ken Peckham
Sheila Hoag
Boyd Gilman

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

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

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

This page left blank for double-sided copying.

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

Primary Care Redesign Awardee	Impact Evaluation Team	Implementation Evaluation Team
CareFirst Blue Cross Blue Shield	Greg Peterson	Kristin Geonnotti, Lauren Hula
Finger Lakes Health Systems Agency	Randall Blair	Rachel Shapiro, Rebecca Coughlin
Research Institute at Nationwide Children's Hospital/Akron Children's Hospital	Eric Lammers, Keith Kranker	Joe Zickafoose, Brenda Natzke
University Hospitals of Cleveland	Andrea Wysocki	Joe Zickafoose, Brenda Natzke

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

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

Statistics Team
Juan Diego Astudillo
Huihua Lu

Editorial and Production Coordination Team
Mark Ezzo
John Kennedy
Felita Buckner
Alfreda Holmes

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

CONTENTS

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

INTRODUCTION

This report is an addendum to the third annual report of the evaluation of the Health Care Innovation Awards (HCIAs) in Primary Care Redesign (PCR) (Peterson et al. 2017). This addendum builds on that earlier report by providing final evaluation results for the four HCIA-PCR awardees listed in Table 1. We had previously reported preliminary results for two of these four awardees—CareFirst BlueCross BlueShield (CareFirst) and Finger Lakes Health Systems Agency (FLHSA)—in the third annual report. For the other two awardees—Research Institute at Nationwide Children's Hospital (NCH) and University Hospitals Cleveland (UHC)—we have not previously reported impact estimates due to limitations in available Medicaid data.

With this addendum report, the HCIA-PCR evaluation is now complete. In this introduction we summarize our impact findings across the 12 awardees for which we were able to conduct an impact analysis.

I. BACKGROUND

In 2012, the Center for Medicare & Medicaid Innovation (CMMI) awarded cooperative agreements of up to $30 million to organizations that proposed compelling models for improving quality of care, improving health outcomes, and lowering medical spending for beneficiaries of Medicare, Medicaid, and the Children's Health Insurance Program (CHIP). The purpose of these awards—the HCIAs—was to expand the source of innovation in health care delivery, building on ideas generated outside of the Centers for Medicare & Medicare Services. Each organization that received an HCIA proposed its own intervention and target population, leading to substantial variation among awardees in intervention content, who delivered it, who received it, and in what contexts (for example, geographic location or type of health system). CMMI classified 14 of the 107 HCIAs issued in 2012 as programs in PCR.

CMMI selected Mathematica Policy Research to evaluate the 14 HCIA-PCR awardees. In our second annual report (Moreno et al. 2016), we reported on program implementation for all 14 awardees, describing intervention characteristics and providing extensive analysis of the barriers to and facilitators of implementation success. In our third annual report (Peterson et al. 2017), we provided information about the impact of 10 of the HCIA-PCR programs on quality of care, service use, and spending among Medicare fee-for-service (FFS) beneficiaries. The third annual report covered outcomes through June 2015, when the original HCIA funding period ended. However, CMMI granted 6 of the 14 PCR awardees extensions to continue providing program services beyond June 2015, using unspent award funds. Table 1 lists 4 of those 6 awardees, and we present final impact estimates for them in this addendum report. We were unable to estimate program impacts for the other 2 PCR awardees—(1) Cooper University Hospital and Camden Coalition of Health Care Providers and (2) Foundation for California Community Colleges and the Transitions Clinic Network—because of challenges identifying an appropriate comparison group.

Table 1. Awardees with evaluation results in this report

Awardee	Abbreviation	Intervention evaluated	Evaluation population	Program end date	New in this addendum report
CareFirst BlueCross BlueShield	CareFirst	*Practice transformation:* Financial incentives for improving quality of care and reducing Medicare spending; care coordination for high-risk patients; and technical assistance to practices to identify opportunities to improve care and reduce spending	Medicare FFS	December 2015	Chapter updates preliminary implementation and impact findings from the third annual report by adding 6 months of data
Finger Lakes Health Systems Agency	FLHSA	*Practice transformation:* Practice facilitation to convert practices into patient-centered medical homes; care coordination for high-risk patients	Medicare FFS	June 2016	Chapter updates preliminary implementation and impact findings from the third annual report by adding 12 months of data
Research Institute at Nationwide Children's Hospital	NCH	*Transitional care:* Peer support for parents or other caregivers during a child's inpatient admission due to a behavioral health condition, and post-discharge follow-up from a nurse care manager	Medicaid FFS and managed care (pediatric only)	June 2015[a]	Chapter updates the description of the intervention design and implementation from the second annual report, and presents impact estimates for the first time
University Hospitals of Cleveland	UHC	*Practice transformation:* Practice facilitation and financial incentives to improve performance on select quality measures, and a suite of services—including telephone triage services and an after-hours clinic—to reduce unnecessary ED use	Medicaid FFS and managed care (pediatric only)	March 2016	Chapter updates the description of the intervention design and implementation from the second annual report, and presents impact estimates for the first time

[a] NCH received a no-cost-extension through December 2015, but the awardee ended program operations in June 2015 for the program component covered in our impact evaluation.

ED = emergency department; FFS = fee-for-service.

II. SUMMARY OF IMPACT FINDINGS FOR 12 HCIA-PCR AWARDEES

Table 2 summarizes our final impact conclusions for the 12 HCIA-PCR awardees for which we conducted an impact evaluation. This includes the 4 awardees covered in detail in this addendum report (Chapters 1 through 4), as well as 8 additional awardees for which we reported final impact results in the third annual report. We summarize the findings of all 12 awardees in this chapter because this addendum report concludes our evaluation of the HCIA-PCR awardees. As described previously, we were unable to conduct an impact evaluation for the remaining 2 awardees due to lack of a suitable comparison group.

1. Methods and interpretation of results

For each awardee, we estimated impacts in up to four evaluation domains: quality-of-care processes, quality-of-care outcomes, service use, and spending. Within each domain, we selected an outcome or outcomes that the awardee aimed to affect and that we could measure in claims data (either Medicare or Medicaid). For example, the quality-of-care processes domain could include measures of preventive services among people with chronic conditions such as diabetes or ischemic vascular disease, or a measure of appropriate ambulatory follow-up care after a hospital admission. The quality-of-care outcomes domain could include measures such as the number of unplanned hospital readmissions or the number of treatments for dental caries. The service use domain could include measures of hospital admissions or emergency department (ED) visits not ending in admission. Finally, the spending domain could include measures of total Medicare spending or inpatient spending. (We were unable to assess Medicaid spending due to limitations in the claims and managed care data available for evaluation.) Each awardee's chapter in the third annual report or this addendum report outlines the outcomes, time periods, and beneficiary populations that we used to assess an awardee's program impacts, along with our rationale for these decisions. Appendix 1 defines the outcomes and describes their construction in detail.

For all 12 awardees, we estimated impacts relative to a comparison group, typically selected through matching (although for 2 awardees we defined the comparison group as beneficiaries served by any practice that met select criteria within a given geographic region). For most of the awardees, we used a difference-in-differences design—that is, we estimated impacts as the difference in outcomes between the treatment and comparison groups during the intervention period, minus the difference in outcomes between the two groups during a (typically one-year) baseline period, and using regression methods to adjust for differences in other observed demographic and health characteristics. When a difference-in-differences design was not possible, we used a contemporaneous differences model instead, estimating impacts as the regression-adjusted difference in outcomes between the treatment and matched comparison groups during the intervention period. Appendix 2 describes the regression models.

For each awardee, we drew an impact conclusion in each of the four evaluation domains (quality-of-care processes, quality-of-care outcomes, service use, and spending). The impact conclusions were categorized as follows:

1. **Statistically significant favorable effect**—the highest level of evidence

2. **Substantively important favorable effect**—when the point estimate for an impact exceeds a prespecified threshold, defined in each chapter, but is not statistically significant
3. **Substantively important unfavorable effect**—when the point estimate for an impact exceeds the prespecified threshold in the unfavorable direction
4. **No substantively large effect**—when the point estimate for an impact falls short of the prespecified threshold but statistical power to detect an effect of that size is good
5. **Indeterminate effect**—when the point estimate for an impact falls short of the prespecified threshold but statistical power to detect an effect of that size is not good (meaning that the program either had no substantively large effects or it did but our tests failed to detect them)

In addition to the five impact conclusions listed, we drew no conclusion at all in a few cases in which robustness checks or the implementation evidence contradicted results from the primary analysis, making clear interpretation impossible. Appendix 3 describes our decision rules for each possible conclusion.

We could not conclude that a program had a statistically significant unfavorable effect because, in consultation with CMMI, we decided to use one-sided statistical tests, which do not test for evidence of unfavorable effects. We chose to report substantively important findings in addition to statistically significant (favorable) ones because the purpose of the HCIAs was to identify promising interventions that might be retested as part of a future CMMI model, not only programs with definitive evidence.

2. Summary of findings

As shown in Table 2 and described in more detail in the chapters of this report and the third annual report, we found the following:

- **Four awardees showed statistically significant improvements in quality-of-care processes.**

  Three of these awardees—FLHSA, PeaceHealth Ketchikan Medical Center (PeaceHealth), and Sanford Health—aimed to transform the way practices as a whole delivered care; however, the specific interventions varied substantially. For example, FLHSA's intervention hired specialized practice facilitators to work with the 68 participating practices in and around Rochester, New York, helping convert practices to patient-centered medical homes; the program also aimed to identify and manage care for high-risk patients. In contrast, PeaceHealth provided a multipronged intervention that spanned both primary care and transitional care from hospital to home—including care management, transitions management, and expanded population health information technology (IT)—in just two practices in remote island communities of southeastern Alaska. The fourth awardee that showed statistically significant improvements in quality-of-care processes was Atlantic General Hospital (AGH) in rural eastern Maryland, with a transitional care intervention for recently discharged beneficiaries who had an AGH-affiliated primary care provider.

Table 2. Summary of final impact conclusions across 12 HCIA-PCR awardees

Awardee	Intervention type (for component[s] included in the impact evaluation)	Evaluation population	Impact conclusion, by domain			
			Quality-of-care processes	Quality-of-care outcomes	Service use	Spending
AGH	Transitional care	Medicare FFS	**Statistically significant favorable effect**	*Substantively important unfavorable effect*	**Statistically significant favorable effect**	**Statistically significant favorable effect**
CareFirst	Practice transformation	Medicare FFS	No substantively large effect	*Substantively important unfavorable effect*	No substantively large effect	Indeterminate effect
CSHP	Care management for high-risk patients	Medicare FFS	Indeterminate effect	**Statistically significant favorable effect**	Indeterminate effect	Indeterminate effect
Denver Health	Practice transformation	Medicare FFS	No substantively large effect	Indeterminate effect	No conclusion	Indeterminate effect
FLHSA	Practice transformation	Medicare FFS	**Statistically significant favorable effect**	Indeterminate effect	No substantively large effect	Indeterminate effect
NCH	Transitional care	Medicaid FFS and managed care	No substantively large effect	Indeterminate effect	**Substantively important (but not statistically significant) favorable effect**	Not applicable
PBGH	Care management for high-risk patients	Medicare FFS	No conclusion	No conclusion	No conclusion	No conclusion
PeaceHealth	Practice transformation	Medicare FFS	**Statistically significant favorable effect**	No conclusion	No conclusion	No conclusion
Sanford Health	Practice transformation	Medicare FFS	**Statistically significant favorable effect**	No substantively large effect	**Statistically significant favorable effect**	Indeterminate effect
TransforMED	Practice transformation	Medicare FFS	No substantively large effect	Not applicable	**Statistically significant favorable effect**	No substantively large effect
UHC	Practice transformation	Medicaid FFS and managed care	No conclusion	No conclusion	No substantively large effect	Not applicable
WIPH	Practice transformation	Medicare FFS	No conclusion	No conclusion	No conclusion	No conclusion

Source: Impact analyses using Medicare FFS claims data or Medicaid FFS claims and encounter data, as presented in individual report chapters or in the third annual report (Peterson et al. 2017).

Notes: We drew impact conclusions at the domain level. See the text for examples of outcome measures within each domain and for a description of the five possible conclusions we could draw. In some cases, we were unable to draw any conclusions (labeled "No conclusion" in the table) because the impact results conflicted with results from robustness tests or the implementation evidence. A conclusion is "Not applicable" if we did not evaluate program impacts in the domain.

This table does not cover the two HCIA-PCR awardees for which we did not conduct an impact evaluation: (1) Cooper University Hospital and Camden Coalition of Health Care Providers and (2) Foundation for California Community Colleges and the Transitions Clinic Network.

AGH = Atlantic General Hospital; CareFirst = CareFirst Blue Cross Blue Shield; CMMI = Center for Medicare & Medicaid Innovation; CSHP = Rutgers Center for State Health Policy; Denver Health = Denver Health and Hospital Authority; FFS = fee-for-service; FLHSA = Finger Lakes Health Systems Agency; HCIA = Health Care Innovation Award; NCH = Research Institute at Nationwide Children's Hospital; PBGH = Pacific Business Group on Health; PCR = primary care redesign; PeaceHealth = PeaceHealth Ketchikan Medical Center; UHC = University Hospitals of Cleveland; WIPH = Wyoming Institute for Population Health at Cheyenne Regional Medical Center.

Under the AGH program, nurses would contact participants within 72 hours of discharge to monitor their adherence to treatment plans and to provide continuity of care between the hospital and outpatient settings.

The four awardees with improved quality-of-care processes also differed in the particular measures that improved. For example, PeaceHealth and Sanford Health both improved preventive care for people with diabetes. AGH and FLHSA both increased the percentage of people who received timely ambulatory care after being discharged from the hospital.

- **One awardee, the Rutgers Center for State Health Policy (CSHP), measurably improved quality-of-care outcomes.**

  CSHP recruited and enrolled people living in high-poverty areas who frequently used acute care services—typically with two or more hospital admissions in six months. CSHP then provided intensive care management to its participants, connecting them to relevant clinical and social services. We estimate that CSHP reduced the rate of unplanned hospital readmissions within 30 days of discharge by 34 percent. For two other awardees, CareFirst and AGH, we estimated that the programs had *unfavorable* effects on quality-of-care outcomes, driven in both cases by substantively large increases in 30-day unplanned readmissions. For AGH, this might have been because the transitional care nurse identified early a patient's need for admission, moving more admissions within the 30-day window. (We found that the program reduced total admissions over the six months following discharge; see the next bullet.) For CareFirst, the program—with its heavy emphasis on care coordination for high-risk patients—might have diverted attention away from lower-risk patients, and thus inadvertently increased readmissions.

- **Four awardees measurably reduced service use.**

  Two of these awardees delivered practice transformation interventions (Sanford Health and TransforMED), whereas the other two (AGH, described previously, and NCH) provided transitional care services. Sanford Health focused on chronic condition management for people with eight targeted conditions, as well as better integration of behavioral health care into primary care. TransforMED, in contrast, provided new health IT tools to help 90 practices in 15 health systems with population health management and cost reporting. NCH provided peer support (nonclinicians) to the parents or guardians of children hospitalized due to a behavioral health condition, and then provided nurse care management services by telephone following discharge. For both TransforMED and AGH, reduced outpatient ED visits and inpatient admissions drove the declines in service use. For Sanford Health and NCH, reduced outpatient ED visits drove the reduction in service use.

- **Only one awardee, AGH, measurably reduced Medicare FFS spending.**

  AGH reduced total Medicare spending by an estimated 31 percent, or $1,333 per beneficiary per month. For five of the remaining six awardees for which we drew conclusions about program impacts on spending, we did not have good statistical power to detect effects in this domain.

III. CONCLUSION

Overall, the HCIA-PCR evaluation results suggest that primary care redesign can succeed in improving patients' outcomes. However, there is no single path to favorable impacts, nor a single, universal intervention that primary care providers can apply successfully in all environments. Across the HCIA-PCR portfolio, a wide range of intervention types had favorable effects in one or more evaluation domains, and this suggests that a range of interventions could be effective. However, as we describe in the individual awardees' chapters here and in the third annual report, improved outcomes will likely depend on the specific context, content, delivery, and target population of the intervention.

REFERENCES

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

Peterson, Greg, Laura Blue, Lorenzo Moreno, Keith Kranker, Boyd Gilman, Kate Stewart, Sandi Nelson, Kristin Geonnotti, Sheila Hoag, Andrew McGuirk, Ken Peckham with the Implementation Team, Impact Team, Data Processing Team, Surveys Team, and Production Coordination and Editorial Team. "Evaluation of Health Care Innovation Awards (HCIAs): Primary Care Redesign Programs. Third Annual Report to CMS." Princeton, NJ: Mathematica Policy Research, March 2017. Available at https://downloads.cms.gov/files/cmmi/hcia-primarycareredesign-thirdannualrpt.pdf. Accessed April 6, 2017.

CHAPTER 1

CAREFIRST BLUECROSS BLUESHIELD

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

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

CAREFIRST BLUECROSS BLUESHIELD

CHAPTER SUMMARY

Introduction. CareFirst BlueCross BlueShield (CareFirst), the largest private health insurer in the mid-Atlantic region, runs a patient-centered medical home (PCMH) program for its commercial members. In 2012, the Centers for Medicare & Medicaid Services (CMS) awarded CareFirst a $20 million Health Care Innovation Award (HCIA) to extend its PCMH program to Medicare fee-for-service (FFS) beneficiaries in Maryland. The intervention targeted about 35,000 Medicare beneficiaries served by 149 primary care providers (PCPs) in 52 practices. These practices formed 14 medical panels—that is, groups of 5 to 15 PCPs who participated as a performance unit in the intervention. The intervention included three components: care coordination for high-risk patients, financial incentives to panels for improving quality of care while reducing spending, and technical assistance to panels to identify opportunities for improving quality and reducing spending. CareFirst aimed to reduce total Medicare spending by 6 percent in the final year of the planned three-year HCIA period (and by 3 percent in the second year) by reducing patients' need for acute care—such as inpatient admissions and emergency department (ED) visits—and by changing PCPs' referral patterns.

The original award period ran for three years, from June 2012 to June 2015. Due to delays receiving and processing the Medicare claims data needed to implement the intervention, CareFirst started the intervention about a year later than planned. Therefore, the program ran for two years during the original award period. In 2015, CMS granted CareFirst a no-cost extension to spend unused award funds to continue the program for another six months, through December 2015.

Our third annual report (Peterson et al. 2017) described program implementation and impacts on patients' outcomes during the original three-year award period. We found that, after the initial delay, the program was generally implemented as planned and it engaged PCPs as planned in coordinating care for high-risk patients. However, we did not find any favorable program impacts in any of the four study domains: quality-of-care processes, quality-of-care outcomes, service use, and spending.

Objectives. In this addendum report, we update the impact and implementation findings from our third annual report to include the six-month extension period through December 2015. Specifically, we (1) describe the program CareFirst implemented during the six-month extension; (2) assess program impacts on patients' service use and Medicare spending during the full intervention period, including the six-month extension; and (3) use both implementation and impact findings to identify possible explanations for the observed impact results in the evaluation domains of service use and spending. We do not update estimates in the quality-of-care processes or outcomes domains because, with the switch from *International Classification of Diseases,* 9th edition (ICD9) to ICD10 reporting in October 2015, we could not construct outcomes for the extension period that were comparable to those in the initial award period.

Methods. To update findings on program implementation, we analyzed CareFirst data on program staffing, care coordination services, technical assistance, and incentive payments during the six-month extension. To estimate program impacts, we compared outcomes for Medicare FFS beneficiaries served by the 14 treatment panels with outcomes for Medicare beneficiaries served by 42 matched comparison panels participating in CareFirst's commercial PCMH program (which does not serve Medicare FFS beneficiaries), adjusting for observed differences in outcomes for the two groups during a one-year baseline period.

Implementation. CareFirst continued to deliver its intervention largely as intended during the six-month extension. Nurses hired by the program collaborated with PCPs to coordinate care for 1,000 to 1,400 high-risk Medicare beneficiaries each month. CareFirst calculated and paid financial rewards to panels as planned based on their cost and quality performance in 2015. Of the 14 panels, 13 earned rewards ranging from $48,770 to $434,197. CareFirst also continued to provide regular technical assistance to panels, with program consultants meeting with each panel about 11 times per quarter.

Impacts on patients' outcomes. The program did not measurably reduce service use (outpatient ED visits and inpatient admissions) or spending during the award period, including the additional six months. Specifically, the estimates in these domains were neither statistically significant nor larger than the prespecified substantive thresholds. In the service use domain, our statistical power to detect effects was good (82 percent), so the program likely did not have substantively large effects. In the spending domain, our power was modest (65 percent), so it is possible that the program had substantive effects but our tests failed to detect them. However, given the lack of impact on service use, we think it is more likely that the program did not substantively reduce spending either.

Conclusion. The findings in this addendum report are fully consistent with those in the third annual report. We found that, during the additional six months, CareFirst continued to implement all elements of the intervention generally as planned. Thus, the program as implemented presents a strong test of the model's core design. However, as in the third annual report, we found no measurable program impacts on service use or spending. The lack of measurable program effects might be due to (1) challenges identifying clinically unstable patients, (2) challenges adapting care coordination strategies from commercial to Medicare populations, and (3) limitations in the intervention design itself. For example, the program design relies on care coordination services for a small subset of high-risk Medicare beneficiaries to drive the overall reduction in spending for panels' Medicare patients. The impacts for this subset of beneficiaries would have to very large—and perhaps unrealistically so—to achieve the planned panel-wide reduction in spending.

In conclusion, we found no evidence of favorable impacts on service use, spending, or quality of care that would support expanding the current version of the CareFirst program to Medicare beneficiaries more broadly. Further revisions to the model could make the program more successful for Medicare beneficiaries.

Summary of intervention and impact results for CareFirst

Intervention description			
Awardee description		Largest commercial health insurer in the mid-Atlantic region	
Award amount ($ millions)		$20.0	
Award extended beyond June 2015?		Yes (6 months)	
Location		Maryland, statewide (urban and suburban)	
Target population		About 35,000 Medicare FFS beneficiaries (excluding those also enrolled in Medicaid) served by 149 PCPs in 52 primary care practices grouped into 14 medical panels	
Interventions		Extended a PCMH program developed for commercial members to Medicare FFS beneficiaries. The program included: • Care coordination, in which 44 HCIA-funded nurses worked with PCPs to develop and implement care plans for high-risk beneficiaries • Financial incentives to (1) reward panels that reduced total spending while meeting quality targets and (2) pay PCPs to participate in care coordination • Technical assistance to panels to identify opportunities to generate savings though changes in referrals	
Metrics of intervention delivered		• Implemented care plans for 3,491 Medicare FFS beneficiaries (almost 10 percent of panels' Medicare beneficiaries) • Care plans active for 260 days, on average, with roughly weekly nurse contact • Rewards from $48,700 to $434,197 to panels with spending below target in 2015	
Impact evaluation methods			
Core design		Difference-in-differences model with matched comparison group	
Treatment group	Definition	Medicare FFS beneficiaries attributed to 14 treatment panels (excluding those also enrolled in Medicaid)	
	# of beneficiaries during primary test period[a]	35,526 to 38,926	
Comparison group definition		Medicare FFS beneficiaries attributed to 42 matched comparison panels participating in CareFirst's commercial PCMH program (excluding those also enrolled in Medicaid)	
Impact results: Quality-of-care processes domain			
Ambulatory care visit within 14 days of discharge (% of beneficiaries/quarter)		Comparison mean[b]	66.0%
		Impact estimate (% difference)	+0.3 pp (+0.4%)
Received recommended lipid test, for patients with IVD (% of beneficiaries/year)		Comparison mean[b]	80.0%
		Impact estimate (% difference)	-0.8 pp (-1.0%)
Received all four recommended diabetes processes of care (% of beneficiaries/year)		Comparison mean[b]	48.5%
		Impact estimate (% difference)	-2.8 pp (-5.7%)
Combined impact estimate[c]		(-2.1%)	
Impact conclusion[d]		No substantively large effect	
Impact results: Quality-of-care outcomes domain			
30-day unplanned hospital readmissions (#/1,000 beneficiaries/quarter)		Comparison mean[b]	8.6
		Impact estimate (% difference)	+1.3 (+16.3%)
Inpatient admissions for ACSC conditions (#/1,000 beneficiaries/quarter)		Comparison mean[b]	11.2
		Impact estimate (% difference)	+0.4 (+3.7%)
Combined impact estimate[c]		(+10.0%)	
Impact conclusion[d]		**Substantively large unfavorable effect**	
Impact results: Service use domain			
All-cause inpatient admissions (#/1,000 beneficiaries /quarter)		Comparison mean[b]	70.1
		Impact estimate (% difference)	+1.4 (+2.0%)
Outpatient ED visits (#/1,000 beneficiaries/quarter)		Comparison mean[b]	85.5
		Impact estimate (% difference)	-2.5 (-3.0%)
Combined impact estimate[c]		(-0.2%)	
Impact conclusion[d]		No substantively large effect	
Impact results: Spending domain			
Medicare Part A and B spending ($/beneficiary/month)		Comparison mean[b]	$1,007
		Impact estimate (% difference)	-$1 (-0.1%)
Impact conclusion[d]		Indeterminate effect	

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

Note: See the CareFirst chapter for details on the intervention, impact methods, and impact results. The measures for the service use and spending domains cover the full intervention period, including the six-month extension. Due to limitations in the data available, the measures in quality-of-care processes and outcomes domains do not cover the extension period, and are the same as reported in the third annual report (Peterson et al. 2017).

[a] Number of beneficiaries in the full treatment group across the quarters in the primary test period—that is, the period over which we test for program impacts.

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

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

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

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

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

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

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

I. INTRODUCTION

CareFirst BlueCross BlueShield, the largest commercial insurer in the mid-Atlantic region, runs a patient-centered medical home (PCMH) program for its commercial members. In 2012, the Centers for Medicare & Medicaid Services (CMS) awarded CareFirst a $20 million Health Care Innovation Award (HCIA) to extend its PCMH program to Medicare beneficiaries. The original award period ran for three years, from June 2012 to June 2015. However, CareFirst started its intervention only in August 2013—13 months later than expected—due to challenges accessing and processing Medicare claims data essential to its intervention. In 2015, CMS granted CareFirst an extension so that it could use unspent award funding to continue the intervention for another six months through December 2015. CareFirst could also use funding for another six months (through June 2016) to calculate and pay financial rewards to participating practices based on their performance in 2015.

Our third annual report (Peterson et al. 2017) described the CareFirst intervention and estimated its impacts on outcomes for Medicare beneficiaries during the original three-year award period. This report updates those findings in three ways. First, we describe the intervention that CareFirst delivered in the final six months of operations under the no-cost extension (July through December 2015). Second, we update the impact estimates to include these final six months for outcomes in the study domains of service use and spending. Including these six months is important because CareFirst expected its impacts on service use and spending to increase over time. We do not update outcomes in the other two study domains—quality-of-care processes and quality-of-care outcomes—due to the shift from *International Classification of Diseases,* 9th edition (ICD9) to ICD10 diagnosis coding on claims that occurred in October 2015, during the six-month extension period. As explained in Section IV, that shift meant we could not construct outcome variables for the final six months of operations that were comparable to outcomes during the original award period. Third, using the longest outcome period available for each domain, we present final conclusions about the program's impacts in all four study domains.

This report is structured as follows. Section II reviews the design of the CareFirst program for Medicare beneficiaries. Section III briefly summarizes earlier findings about the implementation of the model during the original award, and then describes implementation during the no-cost extension period. Section IV describes our methods for, and results from, estimating program impacts on patients' outcomes over the full intervention period, including the six-month program extension. Section V draws conclusions by synthesizing the impact and implementation findings.

II. PROGRAM OVERVIEW

CareFirst built its program for Medicare beneficiaries from its PCMH program for commercial members, which began in 2011. In the commercial program, CareFirst grouped primary care providers (PCPs)—physicians, nurses, and physician assistants—into medical panels. These panels include 5 to 15 PCPs and are the performance unit for the program. CareFirst selected this size so that panels would be large enough for reliable performance measures but small enough for an individual PCP to see his or her contribution to the panel's

performance. Panels can consist of small practices that join together to become units of 5 to 15 people (called virtual panels) or can be larger practices (or a subset of them) that are already the right size.

In 2013, CareFirst selected 14 of the 450 panels in the commercial program to participate in its expansion to Medicare beneficiaries. These 14 panels included 149 PCPs in 52 primary care practices. CareFirst selected panels in Maryland that (1) represented a range of practice sizes and practice ownership, (2) had participated in the commercial program since it began in 2011, and (3) had performed well on cost and quality measures in the commercial program. The target population for the intervention was the roughly 35,000 Medicare fee-for-service (FFS) beneficiaries that the panels served, excluding beneficiaries also enrolled in Medicaid.

The intervention had three components:

1. **Care coordination.** CareFirst hired 44 registered nurses (called local care coordinators [LCCs]) to work with the panels' PCPs to coordinate care for high-risk Medicare beneficiaries with multiple chronic conditions. Using risk scores based on diagnoses in Medicare claims, CareFirst grouped beneficiaries into risk bands and encouraged PCPs and nurses to select beneficiaries in the top bands for the care coordination services. In collaboration with the beneficiaries' PCPs, the LCCs developed and implemented care plans designed to bring the selected beneficiaries' chronic conditions under control. Care plans typically described a regimen of medications, specialty care, diet, exercise, and response to early warning signs.
2. **Financial incentives.** CareFirst paid rewards (called outcome incentive awards) to panels that kept the total cost of care for their Medicare FFS patients below a prespecified target, with the size of the reward scaled to the panel's performance on measures of quality-of-care processes and health outcomes. CareFirst also paid PCPs directly for coordinating care for high-risk patients ($200 for each new care plan or $100 for updating existing plans).
3. **Technical assistance.** CareFirst hired five program consultants who analyzed cost and quality performance data for each panel and met with panels regularly to identify opportunities to improve care processes or refer Medicare beneficiaries to more cost-effective providers or care settings.

CareFirst expected that these intervention components would reduce the need for hospitalizations and post-acute care among high-risk Medicare beneficiaries and encourage PCPs to refer patients to cost-effective providers and settings of care. The reductions in acute care and changes in use of specialty care were expected, in turn, to reduce total Medicare spending.

CareFirst administrators expected that the program's impacts on patients' outcomes would grow over time. They expected little or no impact in the first year of program operations as the panels began changing care for their patients, including coordinating care for high-risk patients. They expected half of the full program impacts by the second year of operations. Then, by the third year, the program would have achieved its full anticipated impacts—reducing hospitalizations by 7.5 percent and total Part A and B spending by 6 percent.

III. PROGRAM IMPLEMENTATION

This section briefly reports our earlier findings (Peterson et al. 2017) about the extent to which CareFirst delivered the intervention as planned during the original award period. Then, we describe the intervention that CareFirst delivered in the final six months of operations in more detail.

A. Implementation during the original three-year award period

The CareFirst program began 13 months later than anticipated because it took time for CareFirst and CMS to develop a process through which CareFirst could access and analyze Medicare claims data needed to support the intervention. However, after that delay, CareFirst ramped up its intervention quickly and delivered it largely as intended. Specifically, CareFirst hired 44 registered nurses who worked with the panels' PCPs to develop care plans for 3,276 high-risk Medicare beneficiaries by June 2015. CareFirst hired more LCCs than the 27 it initially expected because it was able to use unspent funds from the first award year. These LCCs ramped up care coordination quickly, so that CareFirst was able to surpass its original target (set before the 13-month delay) of 1,350 care plans by December 2014. The vast majority (91 percent) of care plans went to beneficiaries in CareFirst's top two illness bands (among five bands). On average, beneficiaries received care coordination services for 260 days, during which LCCs contacted beneficiaries about three times per month, almost always by telephone. CareFirst calculated and paid outcome incentive awards in the two performance years (2013 and 2014) covered during the original award period. Of the 14 panels, 5 earned rewards in 2013 (with rewards ranging from $8,000 to $16,000) and 12 panels earned rewards in 2014 (ranging from $3,000 to $494,000). Program consultants met with panels, on average, about 10 times per quarter, well above CareFirst's initial target of once per quarter.

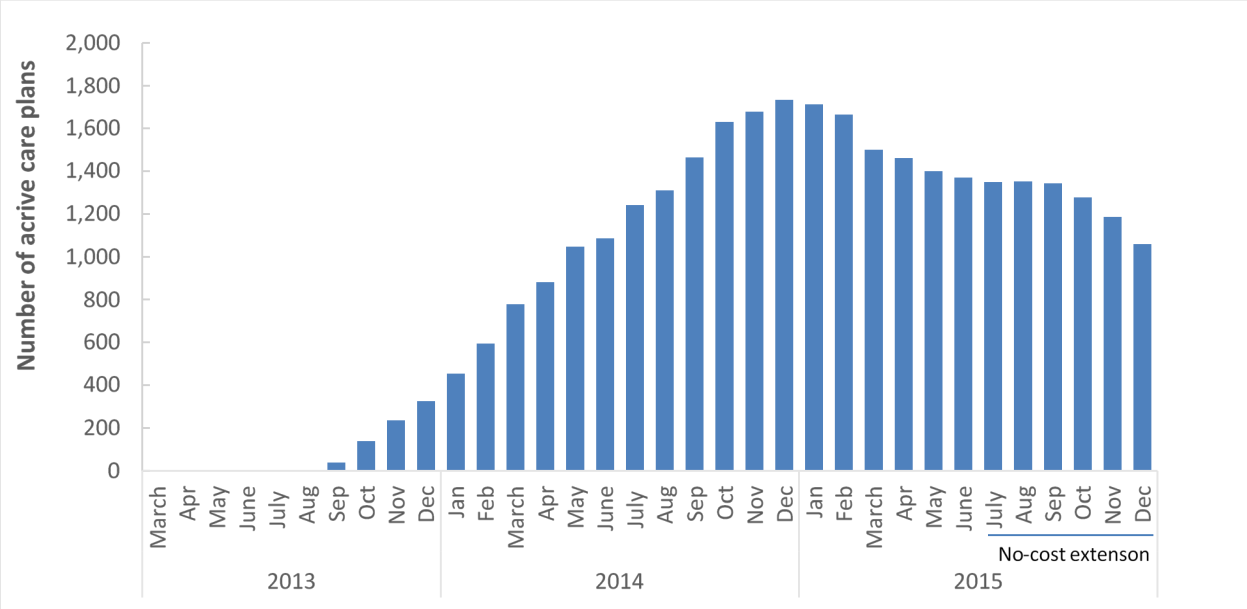
Although CareFirst implemented the intervention largely as planned after the 13-month delay, two key barriers might have limited success of the care coordination component of the intervention. First, CareFirst learned that the process for identifying the highest-risk beneficiaries for care coordination services could have been limited by the extent to which LCCs and PCPs relied on illness burden scores. Some beneficiaries could have had have high illness burden scores but were not actually clinically unstable; rather, their high illness burden scores reflected a recent hospitalization for an acute, nonchronic event, making them less likely to benefit from a care plan targeting complex, chronic conditions. Over time, some PCPs and LCCs refined their process for identifying candidates for care plans, relying on clinical judgement to supplement illness burden scores. Second, although the program was modeled on an existing program targeting commercial members, the clinical complexity of Medicare beneficiaries relative to the commercial patients made implementing the care coordination component challenging.

B. Implementation during the six-month extension period

All 14 of the intervention panels continued to participate in the intervention during the six-month extension (July to December 2015). During those months, CareFirst and the panels continued to deliver the intervention largely as intended, although the number of staff and the reach of care coordination services declined modestly. The following implementation metrics, broken out by the intervention's three components, support this conclusion.

Care coordination. The panels developed and implemented care plans for a total of 3,491 Medicare FFS beneficiaries by December 2015, which included 215 new care plans during the final six months of program operations. During the no-cost extension, CareFirst decided not to open new care plans for beneficiaries unless they were expected to receive a substantial benefit during the short time remaining in the award. This resulted in a modest decline in the number of active care plans, especially in November and December 2015 (Figure III.1), as CareFirst closed some previously opened plans. The program's capacity to create care plans also declined as the number of LCCs dropped from 44 to 30 during the no-cost extension.

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



Sources: Analysis of CareFirst's HCIA quarterly reports, December 2012 through December 2015.
HCIA = Health Care Innovation Award.

Financial incentives. In August 2016, CareFirst paid the third round of outcome incentive awards to panels based on CareFirst's assessment of their performance in calendar year 2015 (the final year of program operations, including the six-month extension). Of the 14 panels participating in the intervention, 13 received awards with payments ranging from $48,770 to $434,197. The mean award in 2015 was $170,144, or $15,987 per participating PCP. CareFirst reported that the panels should have received higher payments based on the savings that CareFirst calculated the panels had generated. However, CareFirst had to cap the incentive payments to remain within the award limits approved by CMS.

Technical assistance. Program consultants met with panels, on average, 11 times per quarter during the six-month no-cost extension period, ending December 2015. This was well above the initial target of once per quarter and a slight increase over rates in the first half of 2015. During these meetings, consultants reviewed and discussed the cost and quality data for the panels' beneficiaries. Program consultants focused increasingly over time on developing strategies to improve panel referral patterns to more cost-effective specialists and settings of

care. The increase in meeting frequency occurred despite the number of program consultants decreasing from five to three during the no-cost extension period.

IV. PROGRAM IMPACTS ON PATIENTS' OUTCOMES

This section presents the methods and results for estimating the impacts of CareFirst's program on patients' outcomes during the full intervention period, from August 2013 to December 2015. We first briefly review the methods for estimating impacts (Section IV.A). We next describe the quantitative impact estimates, their plausibility given implementation findings, and our conclusions about program impacts (Section IV.B).

We have updated the impact estimates only for outcomes in the service use and spending domains. For these domains, we added two quarters of data to the eight quarters included in the third annual report. We were unable to extend the outcome period for the other two study domains: quality-of-care processes and quality-of-care outcomes. The measures in those two domains rely on diagnosis codes in claims and all Medicare providers switched from ICD9 to ICD10 diagnoses at about the same time that CareFirst's extension began. Thus, measures in these domains in the two quarters that correspond to the extension period are not exactly comparable to measures in the previous period. This limits our ability to estimate impacts for the no-cost extension period because we cannot assume that changes over time in measure values represent changes in patients' outcomes—an assumption that underlies our estimation strategy. Although we do not present new impact estimates for the quality-of-care processes and outcomes domains, in Section IV.B.6 we present our impact conclusions from the third annual report for these two domains. We do this so that Section IV.B.6 presents the evaluation's final impact conclusions for all four study domains.

A. Methods

1. Overview

We estimated the impact of CareFirst's HCIA program on patients' outcomes using a difference-in-differences design. Following this design, we estimated impacts as the regression-adjusted difference in outcomes during the intervention period for Medicare FFS beneficiaries attributed to the 14 treatment panels and those attributed to 42 matched comparison panels minus any difference in outcomes for these two groups during a one-year baseline period. We selected the comparison panels from panels in Maryland that did not participate in the HCIA intervention but that performed similarly to the treatment panels on quality and cost measures in CareFirst's PCMH program for commercial members. This choice of comparison group reflected our intention to estimate the marginal impact of extending the program to Medicare beneficiaries. We constructed all study outcomes, and beneficiary-level covariates for our regression models, using Medicare claims and enrollment data (Appendix 1).

We prespecified primary tests for estimating the impact of CareFirst's program. These tests outlined the evidence we would need to conclude that the program was effective, and the awardee and staff at the Center for Medicare & Medicaid Innovation (CMMI) reviewed them. Each test specified a population, outcome, time period, expected direction of effect, and threshold that we count as substantively important. The purpose of these primary tests was to

focus the impact evaluation on hypotheses that would provide the most robust evidence about program effectiveness. We used the results from the primary and secondary tests (robustness checks) to draw conclusions about program impacts in the evaluation domains.

For this report, we made one substantive change to our primary tests from those specified in the third annual report. In that report, we anticipated being able to calculate all outcomes through the end of the extension period, and therefore defined the primary tests in all four evaluation domains to include the two quarters of the no-cost extension. However, as previously mentioned, we were unable to estimate impacts for the quality-of-care outcomes and processes domains during the final six months of the intervention due to challenges posed by the switch from ICD9 to ICD10 codes. Therefore, we have updated our primary tests in these domains (Section IV.A.3) to estimate impacts only through the end of the original award period.

To implement the difference-in-differences design and conduct the primary tests presented in this report, we used a multivariate linear model. Appendix 2 provides details on the regression methods. The unit of analysis was the beneficiary-quarter, with one outcome per person for each quarter that the person met sample eligibility criteria (described in Section IV.A.2). The model's explanatory variables included beneficiary-level covariates (such as age and diagnoses) defined before the intervention began; whether the beneficiary was assigned to a treatment or a comparison panel; an indicator for each panel (which accounted for differences between panels in their patients' outcomes at baseline); indicators for each post-intervention quarter; and an interaction of a beneficiary's treatment status with each post-intervention quarter. The estimated relationship between the interaction term and the outcome in a given quarter is the impact estimate for that quarter. It measures the average difference between outcomes for beneficiaries assigned to the treatment and comparison panels during that period, subtracting out any differences between these groups during the four baseline quarters. The model used robust standard errors to account for clustering of outcomes across quarters for the same beneficiary and a dummy variable for each practice (fixed effects) to account for clustering of outcomes for beneficiaries assigned to the same practice. We averaged the impact estimates across relevant quarters to generate our impact estimates for the primary test periods. We used one-sided tests to assess the statistical significance of these impact estimates, using a threshold for significance of $p = 0.10$. We adjusted the p-values for multiple comparisons we made within a domain (but not across domains).

2. Definitions of the treatment and comparison groups

Treatment group. The treatment group consists of Medicare FFS beneficiaries served by the 14 treatment panels in four baseline quarters before the intervention began (August 1, 2012, to July 31, 2013) and 10 intervention quarters (August 1, 2013, to January 31, 2016). We included January 2016, even though the program officially ended in December 2015, because this enabled us to include one more intervention quarter from the third year of program operations—when CareFirst expected impacts to be largest.

We constructed the treatment group in three steps.

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

Comparison group. The comparison group consists of Medicare FFS beneficiaries whom we assigned to 42 matched comparison panels in each of the baseline and intervention quarters. The comparison panels were similar to the treatment panels during the baseline period on factors that can influence patients' outcomes, especially those factors that CareFirst used when determining panels to recruit for the intervention. We demonstrated the extent to which the treatment and comparison panels were well matched in the third annual report. This section describes how we constructed the matched comparison group.

We identified the 42 comparison panels in four steps:

1. At our request, CareFirst provided a list of all 149 panels (of 438) in the commercial program that met the following criteria that all 14 treatment panels also met: (1) located in Maryland, (2) joined the commercial PCMH program when it began in 2011, and (3) served at least 1,000 CareFirst members in 2012.
2. We developed matching variables, defined at the start of the intervention (August 1, 2013), for all treatment and potential comparison panels. These variables included characteristics of the panel overall (for example, the number of PCPs in the panel and the panel's quality and financial performance in the commercial PCMH program) and characteristics of the

Medicare FFS beneficiaries assigned to the panels (for example, mean Hierarchical Condition Category score and utilization in the baseline period). We did not include measures of quality-of-care processes in the matching because we had not yet calculated these measures when we completed matching (in spring 2015). When assigning Medicare beneficiaries to the comparison panels, we used the same attribution and panel assignment logic that we used for the treatment panels, as described previously.

3. We narrowed the pool of 149 to 101 potential comparison panels that, like the treatment panels, (1) had an average of at least 500 assigned Medicare FFS beneficiaries during the four baseline quarters, (2) had at least five PCPs at the start of the intervention, and (3) were located in urban or suburban areas.

4. We used propensity-score methods to select 42 comparison panels from the pool of 101 that were similar to the 14 treatment panels on the matching variables, which we presented in detail in the third annual report (Peterson et al. 2017). The propensity score is the predicted probability, based on all of a panel's matching variables, that a given panel was selected for treatment (Stuart 2010). It collapses all of the matching variables into a single number for each panel that can be used to assess how similar panels are to one another. By matching each treatment panel to one or more comparison panels with similar propensity scores, we generated a comparison group that is similar, on average, to the comparison group on the matching variables. The approach, however, does not ensure that each comparison panel matches exactly to its treatment panel on all matching variables. We prioritized one matching variable—whether a panel is virtual—by requiring that a virtual treatment panel could match only to a virtual comparison panel, and a nonvirtual treatment panel could match only to a nonvirtual comparison panel. Such panels were likely to have fewer resources, and greater coordination challenges, than the nonvirtual panels, which were part or all of a single, larger practice.

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

After completing the matching, we assigned Medicare FFS beneficiaries to the comparison practices in each intervention quarter using the same rules applied to the treatment group.

3. Revised primary tests

Table IV.1 shows the primary tests for CareFirst, by domain. As described previously, we have revised the time period covered by the primary tests for our impact analyses in the quality-of-care processes and outcomes domains because of the change in coding on Medicare claims (Section IV.A.1). All other elements of the primary tests—outcomes, populations, and substantive thresholds—remain unchanged from the third annual report.

Table IV.1. Specification of the primary tests for CareFirst BlueCross BlueShield (revised)

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

Table IV.1 *(continued)*

Note: We revised the primary tests because, as described in the text, we had two fewer quarters of data than initially expected for outcomes in the quality-of-care processes and outcome domains.

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

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

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

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

Our rationale for selecting these primary tests is as follows:

- **Outcomes.** CareFirst's central goal was to reduce hospitalizations, emergency department (ED) visits, and Medicare Part A and B spending, so our primary tests address these three outcomes. In addition, the primary tests address two quality-of-care outcomes the intervention is expected to affect: ambulatory care-sensitive condition (ACSC) admissions and 30-day unplanned hospital readmissions. Finally, we include three quality-of-care process measures that, based on CareFirst's program design, we think the program could improve: (1) a composite measure for whether a beneficiary with diabetes received all four recommended processes of care during the year (HbA1c test, lipid profile, dilated eye exam, and nephropathy screening); (2) receipt of a complete lipid profile for people with ischemic vascular disease (IVD); and (3) receipt of a follow-up ambulatory care visit with a primary care or specialist provider within 14 days of hospital discharge. Although CareFirst did not set explicit targets for these particular quality-of-care process measures, the outcome incentive awards incentivized improvements in processes of care for chronic illnesses and the care coordination intervention could be expected to improve 14-day follow-up rates.
- **Time period.** CareFirst expected participating panels to show substantial impacts by their second program year. For this reason, for the service use and spending domains, our primary tests cover the 1.5 years from August 2014 through January 2016 (or the 5th through 10th intervention quarters [I5 through I10]), a period that began one year after the program started in August 2013. The final impact analysis includes one month beyond the intervention end date (December 31, 2015) so that we could include outcomes for the quarter that runs from November 2015 to January 2016, most of which fell during the program's operational period. Because we were unable to include in our analysis any measures of ACSC admissions, 30-day unplanned readmissions, or ambulatory care within 14 days of hospital admissions for the final six months of program operations, the primary test period for these three outcomes is the second year of program operations (I5 through I8). Most of the measures are defined quarterly, so our impact estimates represent averages across relevant quarters. In contrast, because the quality-of-care process measures for IVD and diabetes are defined over a year, our primary tests assess impacts during the second full year of program operations (a period corresponding to I5 through I8).
- **Population.** For all but the three quality-of-care process measures, the population includes all Medicare FFS beneficiaries (excluding those dually eligible for Medicaid) assigned to the 14 treatment panels. This corresponds to CareFirst's definition of its target population. For the diabetes and IVD quality-of-care process measures, we limit the sample to beneficiaries ages 18 to 75 with diabetes or ages 18 and older with IVD, respectively, and who were observable in Medicare FFS claims for all 12 months of the measurement year. For the 14-day follow-up measure, we limit the sample in each quarter to those who had at least one qualifying hospitalization during the quarter for which we could observe whether the beneficiary had a 14-day follow-up visit.
- **Direction (sign) of the impact estimate.** For the quality-of-care process measures, we expect the impact estimate to be positive, signaling an increase in the percentage of patients

receiving recommended care. For all other outcomes, we expect the impact estimates to be negative, indicating a reduction in service use or overall expenditures.

- **Substantive thresholds.** Some impact estimates could be large enough to be policy relevant to CMMI and other stakeholders even if they are not statistically significant; for this reason, we prespecified thresholds for substantive importance. We express this threshold as a percentage change from the counterfactual—that is, the outcomes that beneficiaries in the treatment group would have had if they had not received the HCIA-funded intervention. For all but the quality-of-care process measures, the 4 to 5 percent thresholds we chose (depending on the outcome) are 75 percent of CareFirst's expected effects during the primary test periods. (We use 75 percent recognizing that CareFirst could still be considered successful if it approached, but did not fully achieve, its anticipated effects.) The 15 percent threshold for the quality-of-care process measures is taken from relevant literature (Peikes et al. 2011; Rosenthal et al. 2016) because CareFirst did not specify by how much it expected to improve these outcomes.

4. Secondary tests (robustness checks)

We conducted secondary quantitative tests to corroborate the findings from the primary tests. This is important because some of the differences observed between the treatment and comparison groups in the primary test results could reflect limitations of the non-experimental impact evaluation design or random fluctuations in the data. We have greater confidence in the primary results if they are generally consistent with the expected broader pattern of results from the secondary tests.

We conducted two sets of secondary tests for CareFirst.

1. We estimated the program's impacts on all-cause admissions and total Medicare spending during two intervention periods, in addition to those specified in the primary tests: (1) the first 6 months after the panels joined the intervention (I1 and I2), and (2) months 7 to 12 after the panels joined the intervention (I3 and I4). Because we and CareFirst expected program impacts to increase over time, with little or no impacts in the first few months of the program, the following pattern would be highly consistent with an effective intervention: little to no measured effects in the first two quarters, growing effects in quarters 3 and 4, and the largest impacts in quarters 5 through 10. In contrast, very large differences in outcomes (favorable or unfavorable) in the first 6 intervention months could suggest a limitation in the comparison group, rather than true intervention impacts.
2. We reestimated impacts on all-cause admissions and total Medicare spending, but limiting to beneficiaries assigned to the treatment and comparison groups by the start of the period, either baseline or intervention. This restriction prevents addition to the intervention sample over time. Differences in sample addition between the treatment and comparison groups could bias the impact results if the sample members added over time differ from earlier sample members (for example, they are younger and healthier). This could create differences in mean outcomes between the treatment and comparison groups that are unrelated to the intervention. We have explored this possibility because, as we describe in Section IV.B.1, the rate of net sample growth during the intervention period was slightly

higher for the comparison group (growth of 26.5 percent from I1to I10) than for the treatment group (growth of 23.6 percent). It appears differences in sample addition drive the differences in net sample growth because the rate of sample loss was the same during the intervention period (results not shown). That is, the percentage of beneficiaries assigned to treatment and comparison groups in the first quarter but lost to follow-up (due to death, moving into managed care, dual enrollment in Medicaid, or moving out of state) by the end of the intervention period is exactly the same (11.0 percent) in the treatment and comparison groups.

B. Beneficiaries' outcomes and intervention impacts

In this section, we first present sample sizes and mean outcomes by quarter for the treatment and comparison groups. We present sample sizes and mean outcomes across all study quarters—not simply the two new quarters—to show how those final quarters of data fit into the larger trends from the earlier periods These mean outcomes provide context for understanding the difference-in-differences estimates that follow; however, the differences in mean outcomes are not regression-adjusted and not impact estimates. Next, we present the results of the primary tests for service use and spending, followed by the results of the secondary tests (robustness checks). These assess whether the primary test results are plausible given the secondary test results and the implementation evidence. We then draw conclusions about program impacts in each domain.

As described earlier, we present quantitative results only for the service use and spending outcomes because these are the results we have updated through the end of the extension period. We do not present results for the quality-of-care processes or quality-of-care outcomes domains because we were unable to extend our evaluation of these outcomes beyond the original three-year award period. However, we summarize our impact conclusions across all four study domains—including the two quality-of-care domains—in Section IV.B.6 so that we present all of the final impact conclusions in this report.

1. Sample sizes

In the first baseline quarter (B1), the treatment group includes 29,405 beneficiaries assigned to the 14 participating panels and the comparison group includes 59,655 beneficiaries assigned to the 42 comparison panels (Table IV.2). The sample sizes increase modestly during the four baseline quarters (by 11 percent from B1 to B4). This net increase indicates that sample addition (due to beneficiaries being newly attributed to the treatment or comparison practices) exceeds sample attrition (due to beneficiaries dying, switching from FFS Medicare to managed care, moving out of state, or dually enrolling in Medicaid). The sample sizes drop modestly from the last baseline quarter to the first intervention quarter, reflecting that the sample definition (Section IV.A.2) retains sample members in successive baseline and intervention quarters, even if they are no longer attributed to the treatment or comparison panel, but not between the baseline and intervention periods. The sample increases modestly during the intervention period, again reflecting greater sample addition than attrition over time. The net sample increase during the intervention period is smaller for the treatment group (23.6 percent from I1 to I10) than the comparison group (26.5 percent over the same period).

Table IV.2. Unadjusted mean outcomes for service use and spending, by treatment status and quarter

	Number of Medicare FFS beneficiaries (panels)			All-cause inpatient admissions (#/1,000 beneficiaries/quarter)			Outpatient ED visit rate (#/1,000 beneficiaries/quarter)			Medicare Part A and B spending ($/beneficiary/month)		
Q	T	C (no wgt)	C (wgt)	T	C	Diff (%)	T	C	Diff (%)	T	C	Diff (%)
Baseline period (August 1, 2012–July 31, 2013)												
B1	29,405 (14)	59,655 (42)	29,456	78.7	78.0	0.7 (0.9%)	82.7	82.5	0.1 (0.2%)	$997	$960	$37 (3.8%)
B2	30,607 (14)	62,539 (42)	30,877	79.6	80.0	-0.4 (-0.5%)	75.5	80.1	-4.6 (-5.8%)	$956	$973	$-16 (-1.7%)
B3	32,126 (14)	64,101 (42)	31,707	81.9	81.4	0.5 (0.6%)	74.7	73.7	1.0 (1.3%)	$1,003	$985	$18 (1.8%)
B4	32,840 (14)	65,864 (42)	32,939	77.5	74.8	2.7 (3.6%)	83.9	86.9	-3.0 (-3.5%)	$1,001	$986	$15 (1.5%)
Intervention period (August 1, 2013–January 31, 2016)												
I1	31,492 (14)	61,216 (42)	31,134	77.2	70.9	6.2 (8.8%)	80.4	81.2	-0.9 (-1.0%)	$1,015	$948	$67 (7.1%)
I2	32,843 (14)	64,307 (42)	32,652	74.6	74.4	0.2 (0.2%)	77.1	74.3	2.7 (3.6%)	$952	$949	$4 (0.4%)
I3	33,514 (14)	66,298 (42)	33,568	75.6	70.7	4.9 (7.0%)	78.4	77.5	0.9 (1.2%)	$997	$922	$75 (8.1%)
I4	34,582 (14)	68,662 (42)	34,614	74.4	71.7	2.7 (3.7%)	88.8	89.3	-0.5 (-0.5%)	$985	$978	$7 (0.7%)
I5	35,526 (14)	71,712 (42)	35,823	72.9	70.7	2.2 (3.2%)	83.6	87.2	-3.6 (-4.2%)	$1,003	$1,024	$-21 (-2.0%)
I6	36,504 (14)	73,688 (42)	36,944	78.6	72.2	6.5 (9.0%)	83.1	84.6	-1.5 (-1.8%)	$989	$938	$51 (5.4%)
I7	36,936 (14)	74,351 (42)	37,408	70.7	68.1	2.6 (3.8%)	79.1	85.9	-6.8 (-7.9%)	$1,039	$979	$59 (6.0%)
I8	37,576 (14)	75,946 (42)	38,229	69.3	67.5	1.9 (2.8%)	86.6	94.5	-7.9 (-8.4%)	$1,033	$1,011	$22 (2.2%)
I9	38,227 (14)	77,164 (42)	38,774	69.5	67.2	2.3 (3.4%)	86.8	88.1	-1.2 (-1.4%)	$1,033	$1,011	$23 (2.2%)
I10	38,926 (14)	78,448 (42)	39,385	68.1	68.0	0.2 (0.2%)	78.6	86.8	-8.3 (-9.5%)	$940	$957	$-17 (-1.8%)

Table IV.2 *(continued)*

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

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

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

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

2. Mean outcomes for the treatment and comparison groups

Service use. All-cause inpatient admissions declined substantially—by 16 to 17 percent—for both the treatment and comparison groups from B3 to I10. In all but one of the study quarters, the rates were modestly higher (0.2 to 9.0 percent higher) for the treatment group than the comparison group, without any consistent trend of increasing or decreasing differences.

The outpatient ED visit rates fluctuated over time and were generally similar between the treatment and comparison groups. The treatment group had moderately lower rates (by 7.9 to 9.5 percent) than the comparison group in three of the last four intervention quarters.

Spending. Mean spending was close to $1,000 per beneficiary per month for all baseline and intervention quarters for both treatment and comparison groups. The treatment group means ranged from 2.0 percent lower to 8.1 percent higher than the comparison group means across quarters, but these differences neither increased nor decreased over time.

3. Primary tests results

Service use. The treatment group's admission rate was 2.0 percent higher, and the outpatient ED visit rate was 3.0 percent lower, than the estimated counterfactuals. (The estimated counterfactual—that is, the estimated outcome the treatment group members would have had in the absence of the intervention—is the treatment group mean minus the difference-in-differences estimate.) Neither of these differences was statistically significant or larger than the prespecified substantive threshold. After combining results across the two outcomes in this domain, the outcomes for the treatment group were almost identical to (0.5 percent lower than) the estimated counterfactual. Power to detect effects that were the size of the substantive thresholds was marginal for the admissions and outpatient ED visit measures individually (62.8 and 73.4 percent, respectively) but good (82.2 percent) for the two outcomes combined.

Spending. The treatment group averaged $1,006 per beneficiary per month in Part A and B spending during the I5 through I10, a value 0.1 percent (or $1) lower than the estimated counterfactual. This difference was much smaller than the substantive threshold of 4 percent. Statistical power to detect an effect the size of the substantive threshold was marginal (65.0 percent).

Table IV.3. Results of primary tests for CareFirst for the service use and spending domains

Primary test definition					Statistical power to detect an effect that is[a]		Results			
Domain (# of tests in domain)	Outcome (units)	Time period for impacts	Population	Substantive threshold (expected direction of effect)[b]	Size of the substantive threshold	Twice the substantive threshold[c]	Treatment group mean	Regression-adjusted difference between the treatment and estimated counterfactual[b] (standard error)	Percentage difference[d]	*p*-value[e]
Service use (2)	All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	Average over intervention quarters 5 through 10	Medicare FFS beneficiaries assigned to treatment panels	5.0% (-)	62.8	97.4	71.5	1.4 (2.2)	2.0%	0.62
	Outpatient ED visits (#/1,000 beneficiaries/quarter)	Average over intervention quarters 5 through 10	Medicare FFS beneficiaries assigned to treatment panels	5.0% (-)	73.4	99.4	83.0	-2.5 (2.2)	-3.0%	0.22
	Combined (%)	Average over intervention quarters 5 through 10	Medicare FFS beneficiaries assigned to treatment panels	5.0% (-)	82.2	99.9	n.a.	n.a.	-0.5%	0.42
Spending (1)	Medicare Part A and B spending ($/beneficiary/month)	Average over intervention quarters 5 through 10	Medicare FFS beneficiaries assigned to treatment panels	4.0% (-)	65.0	98.0	$1,006	-$1 (24.2)	-0.1%	0.49

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

Note: The results for each outcome are based on a difference-in-differences regression model. Estimates are calculated for Medicare beneficiaries who are observable in the relevant time period: that is, beneficiaries who are enrolled in Medicare FFS (Part A and B), are alive, and have Medicare as their primary payer.

[a] The power calculation is based on actual standard errors from the analysis. For example, in the last row, a 4.0 percent effect on Medicare Part A and B spending (from the counterfactual of $1,006 - $1 = $1,005) would be a change of $40. Given the standard error of $24 from the regression model, we would be able to detect a statistically significant result 65.0 percent of the time if the impact was truly -$40, assuming a one-sided statistical test at the $p = 0.10$ significance level.

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

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

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

[e] *p*-values test the null hypothesis that the regression-adjusted difference-in-differences estimate is greater than or equal to zero (a one-sided test). Because it is a one-sided test, as the difference-in-differences estimate approaches infinity in an unfavorable direction, the *p*-value approaches 1, whereas it would approach 0 in a two-sided test. We adjusted the *p*-values for the multiple (two) comparisons made within the service use domain.

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

n.a. = not applicable.

4. Secondary test results

Estimates during the first intervention year (August 1, 2013, to July 31, 2014). As shown in Table IV.4, the differences in admissions and spending between the treatment group and its estimated counterfactual were small (2.5 percent or less) and not statistically significant during the two secondary test periods: the first six months of the intervention (I1 and I2) and the next six months (I3 and I4). These results support the credibility of the comparison group because we do not see large differences (favorable or unfavorable) during the first year of panel participation, a period during which we and CareFirst did not expect to see large program effects. This increased confidence in the comparison group, in turn, gives us greater confidence in the primary test results and, eventually, the overall conclusions of the impact evaluation.

Estimates limiting the sample to prevent sample addition. The secondary test results limited to those beneficiaries attributed at the start of the baseline or intervention period are consistent with the primary test results. They show no statistically significant or substantively large differences between the treatment group and its estimated counterfactual for inpatient admissions or Medicare spending.

5. Consistency of impact estimates with implementation findings

The impact estimates in the primary tests are plausible given the implementation findings. The primary test results showed no statistically significant or substantively important favorable effects during the 1.5-year primary test period (August 2014 through January 2016). The implementation evidence shows the intervention was active during these 18 months. For example, as described in Section III.B, the treatment panels provided care coordination services to 1,000 and 1,400 high-risk Medicare beneficiaries each month during this period. However, even with a well-implemented intervention, it is possible that the program was not able to change beneficiaries' or providers' behaviors in ways that would affect patients' outcomes.

6. Conclusions about program impacts, by domain

Based on the evidence presented in this report for the service use and spending domains—and based on the evidence we presented in the third annual report (Peterson et al. 2017) for the quality-of-care processes and outcomes domains—we draw the following final conclusions about the program's impacts on patients' outcomes. Table IV.5 summarizes these conclusions and their supporting evidence.

Table IV.4. Results of secondary tests for CareFirst

Secondary test definition				Results			
Domain	Outcome (units)	Time period for impacts	Population	Treatment group mean	Regression-adjusted difference between treatment and the estimated counterfactual (standard error)	Percentage difference[a]	*p*-value[b]
Estimates during the first intervention year (August 1, 2013–July 31, 2014)							
Service use	All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	Intervention quarters 1 and 2	All Medicare FFS beneficiaries assigned to treatment panels	75.8	1.0 (2.7)	1.4%	0.65
	All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	Intervention quarters 3 and 4	All Medicare FFS beneficiaries assigned to treatment panels	74.9	1.8 (2.7)	2.5%	0.75
Spending	Medicare Part A and B spending ($/beneficiary/month)	Intervention quarters 1 and 2	All Medicare FFS beneficiaries assigned to treatment panels	$984	$9 ($27)	1.0%	0.63
	Medicare Part A and B spending ($/beneficiary/month)	Intervention quarters 3 and 4	All Medicare FFS beneficiaries assigned to treatment panels	$991	$17 ($29)	1.7%	0.72
Estimates limiting the sample to prevent sample addition after the first baseline or intervention quarter							
Service use	All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	Intervention quarters 5–10	Medicare FFS beneficiaries assigned to treatment panels in the first baseline or first intervention quarter	74.6	0.9 (2.4)	1.3%	0.65
Spending	Medicare Part A and B spending ($/beneficiary/month)	Intervention quarters 5–10	Medicare FFS beneficiaries assigned to treatment panels in the first baseline or first intervention quarter	$1,032	-$2 (26.6)	-0.2%	0.47

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

Table IV.4 *(continued)*

Note: The results for each outcome are based on a difference-in-differences regression model. Estimates are calculated for Medicare beneficiaries who are observable in the relevant time period: that is, beneficiaries who are enrolled in Medicare FFS (Part A and B), are alive, and have Medicare as their primary payer.

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

[b] *p*-values test the null hypothesis that the regression-adjusted difference-in-differences estimate is greater than or equal to zero (a one-sided test). Because it is a one-sided test, as the difference-in-differences estimate approaches infinity in an unfavorable direction, the *p*-value approaches 1, whereas it would approach 0 in a two-sided test.The *p*-values from the secondary test results were not adjusted for multiple comparisons within or across domains.

FFS = fee-for-service.

- **The intervention did not have a substantively large impact on quality-of-care processes or service use.** For all outcomes in these domains, the primary test results were neither substantively large nor statistically significant. The statistical power to detect effects in these domains was good (more than 75 percent). Specifically, in the quality-of-care processes domain, power was good for each of the measures in the domain (lipid testing for people with IVD, receipt of four recommended care processes for those with diabetes, and ambulatory care follow-up visit within 14 days of a hospital stay). In the service use domain, power was good for the combined impact estimate across two outcomes in the domain (all-cause inpatient admissions and outpatient ED visits). The secondary test results support these primary test results by (1) showing no impacts in the first program year (when none were expected) and (2) demonstrating that differential sample addition over time between the treatment and comparison groups did not drive results. These conclusions are also consistent with implementation findings because, although the program was implemented reasonably well, it is plausible the program did not have its intended effects.
- **The program had a substantively large *unfavorable* effect on quality-of-care outcomes.** The primary tests presented in the third annual report showed a substantively large unfavorable estimate for quality-of-care outcomes, driven by a large unfavorable estimate for 30-day unplanned readmissions, in particular. However, the standard errors were large for both this estimate and the estimated combined effect in the domain. Therefore, we have low confidence in the conclusion of substantively unfavorable impacts. It is possible that the large observed point estimates were due to chance rather than true unfavorable impacts. However, there is a potentially plausible explanation for how the program could have worsened quality-of-care outcomes. Though we have no direct evidence to suggest this happened, it is possible that (1) PCPs diverted attention from lower- to higher-risk patients (for example, to focus time on coordinating care for high-risk patients); or (2) participating in the intervention prevented treatment panels from undertaking other quality initiatives to reduce readmission rates for their Medicare population that comparison panels might have.
- **The program had an indeterminate effect on Medicare spending.** The primary test results were neither substantively large nor statistically significant. However, the statistical power was marginal (65.0 percent) to detect effects the size of the substantive threshold. As a result, null findings from the primary test in this domain could be due to (1) the program truly not having a substantively large effect or (2) the program having a substantively large effect, but our tests failing to detect it. The fact that we observed no declines in service use (which the awardee anticipated would drive reductions in spending)—and the fact that primary tests for service use were well powered—suggests the more likely explanation is lack of effects on spending.

Table IV.5. Conclusions about the impacts of CareFirst's HCIA program on patients' outcomes, by domain

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

Sources: Tables IV.3 and IV.4 for the service use and spending domains, and the third annual report (Peterson et al. 2017) for the quality-of-care processes and quality-of-care outcomes domains.

HCIA = Health Care Innovation Award.

V. DISCUSSION AND CONCLUSION

The conclusions about program impacts in this report are consistent with those in the third annual report. This report updates our analysis from the third annual report by adding CareFirst's final six months of program operations to our impact evaluation of the service use and spending domains. Because CareFirst expected program impacts to grow over the first two years of program operations, these six months—which began the third year of operations—are when CareFirst expected to have fully realized its program impacts. However, even including these six months, we continued to find that the program did not have a substantively large impact on service use (ED visits or inpatient admissions). Similarly, we continued to find no measurable impact on Medicare Part A and B spending.

For the quality-of-care domains (processes and outcomes), we did not add any new data to this report. Therefore, our conclusions—and the data that support them—remain unchanged from the third annual report. Specifically, we conclude that the program had neither a substantively large nor statistically significant impact on quality-of-care processes. Further, it might have had a substantively large *unfavorable* impact on quality-of-care outcomes—driven by large increases in 30-day unplanned hospital readmissions. Our conclusions for the quality-of-care domains might have changed had we been able to include the final six months of program operations, as

initially planned. However, the estimates for those final six months would have had to be very large if they were to change the impact estimates over the full primary test period.

Our findings contrast with CareFirst's assessment of its program. CareFirst estimates that its intervention reduced overall Medicare spending by $65 million after subtracting the cost of the intervention. The calculations program staff used to reach this conclusion assumed that panel costs would have grown by 2.5 percent per year absent the intervention (CareFirst 2016). We believe this assumed cost growth was too high, as shown by the essentially zero cost growth we observed in comparison panels.

We estimated the marginal effect of extending the commercial PCMH program to Medicare beneficiaries because that is the most policy-relevant analysis for CMS. However, if there were any spillover effects from the commercial program (for example, if comparison panels changed their referral patterns and succeeded in reducing specialty costs among Medicare beneficiaries and among commercial members), we would expect these spillover effects to be small. The program's main focus was individualized care coordination services—which presumably could not have spillover effects.

However, we did find favorable trends in the (unadjusted) treatment group outcomes over time. Specifically, inpatient admissions declined by 17 percent from the baseline period to the end of the intervention. Further, spending did not increase over the 2.5-year period, remaining close to $1,000 per beneficiary per month in all quarters. However, these favorable trends were matched by similar trends in the comparison group, so forces outside of the program likely drove them. In Maryland, as in the country as a whole, Medicare hospitalization rates have declined in the past decade, and cost growth has been modest in the past five years (Krumholz et al. 2015; Boards of Trustees 2016; Maryland Health Care Commission 2016). These falling admission rates and low cost growth might result from a combination of improved health among patients, hospitals' responses to new incentives to reduce readmissions, a shift in location of hospital services from inpatient to outpatient settings, and—in Maryland in particular—the initiation of global hospital budgets in 2013 that created strong incentives to reduce avoidable hospital admissions (Krumholz et al. 2015; Patel et al. 2015).

The lack of observed program impacts are not due to failure to implement CareFirst's program. After a 13-month delay in program operations, CareFirst largely implemented its three program components (care coordination, financial incentives, and technical assistance) as planned.

Rather, the lack of measured effects might be due to one of three factors. First, although the program was generally implemented as planned, a few key implementation barriers might have limited the effectiveness of care coordination services. The illness burden scores might not have consistently flagged beneficiaries who were clinically unstable and thus could benefit most from care plans. Rather, beneficiaries' high burden scores might have reflected a recent acute event not tied to a chronic illness. Further, as several respondents noted in interviews, it was difficult to adjust the care coordination process—originally designed for commercial patients—to the Medicare population due to their generally greater clinical complexity.

Second, there might have been limitations in the core design of the intervention itself. Our evaluation was not designed to identify specific limitations in program design that could account for lack of effects. However, a quantitative analysis of CareFirst's data suggests that program impacts for those receiving care coordination would have to be very large—perhaps unrealistically large—to drive the targeted reduction in overall spending. Specifically, based on the percentage of the panels' Medicare beneficiaries who enrolled in care coordination services (9 percent) before or during the primary test period and their total spending relative to the average beneficiary's spending in the treatment group (about 2:1), we estimate the program would have had to reduce spending for those receiving care coordination services by 22 percent to achieve the intended full-panel reduction in spending of 4 percent. Such large reductions could be difficult to achieve, particularly given the challenges noted previously in systematically identifying care coordination participants and in adapting care coordination strategies from the commercial population.

Finally, the method CareFirst used to calculate panel performance might have muted an important feedback loop to panels. By using a benchmark (2.5 percent) that was well above actual spending growth, CareFirst calculated and paid large financial incentives to participating panels. If CareFirst had used a benchmark closer to actual spending growth in Maryland, fewer panels would have earned incentive awards and payments would have been smaller. This might have sent more appropriate signals to panels that they had to continue to adapt their interventions for Medicare beneficiaries to meet program aims.

In conclusion, we found no evidence of favorable impacts on service use, spending, or quality of care that would support expanding the current version of the CareFirst program to Medicare beneficiaries more broadly. Further revisions to the model could make the program more successful for Medicare beneficiaries.

REFERENCES

Boards of Trustees, Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds. "2016 Annual Report of the Boards of Trustees of the Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds." Washington, DC: Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds, 2016. Available at https://www.cms.gov/research-statistics-data-and-systems/statistics-trends-and-reports/reportstrustfunds/downloads/tr2016.pdf. Accessed April 6, 2017.

CareFirst Blue Cross BlueShield. "Program Description and Guidelines for CareFirst Patient-Centered Medical Home Program and Total Care and Cost Improvement Program." Baltimore, MD: CareFirst of Maryland, Inc., 2016. Available at https://provider.carefirst.com/carefirst-resources/provider/pdf/pcmh-program-description-guidelines.pdf. Accessed March 27, 2017.

Krumholz, H.M., S.V. Nuti, N.S. Downing, S.L. Normand, and Y. Wang. "Mortality, Hospitalizations, and Expenditures for the Medicare Population Aged 65 Years or Older, 1999–2013." *JAMA,* vol. 314, no. 4, 2015, pp. 355–365. doi: 10.1001/jama.2015.8035

Maryland Health Care Commission. "Annual Report on Selected Maryland General and Special Hospital Services: Fiscal Year 2016." Baltimore, MD: Maryland Health Care Commission, 2016. Available at http://mhcc.maryland.gov/mhcc/pages/hcfs/hcfs_hospital/documents/acute_care/chcf_acute_care_license_rpt_2016_20151130.pdf. Accessed April 6, 2017.

Patel, A., R. Rajkumar, J.M. Colmers, D. Kinzer, P.H. Conway, and J.M. Sharfstein. "Maryland's Global Hospital Budgets: Preliminary Results from an All-Payer Model." *New England Journal of Medicine,* vol. 373, no. 20, 2015, pp. 1899–1901. doi: 10.1056/NEJMp1508037

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

Peterson, Greg, Kristin Geonnotti, Lauren Hula, Boyd Gilman, Laura Blue, Keith Kranker, Kate Stewart, Sheila Hoag, and Lorenzo Moreno. "CareFirst BlueCross BlueShield," in Greg Peterson, Laura Blue, Lorenzo Moreno, Keith Kranker, Boyd Gilman, Kate Stewart, Sandi Nelson, Kristin Geonnotti, Sheila Hoag, Andrew McGuirk, Ken Peckham. "Evaluation of Health Care Innovation Awards (HCIA): Primary Care Redesign Programs. Third Annual Report to CMS." Princeton, NJ: Mathematica Policy Research, March 2017. Available at https://downloads.cms.gov/files/cmmi/hcia-primarycareredesign-thirdannualrpt.pdf. Accessed April 6, 2017.

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

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

CHAPTER 2

FINGER LAKES HEALTH SYSTEMS AGENCY

Randall Blair, Rachel Shapiro, Rebecca Coughlin, Greg Peterson, Boyd Gilman, Laura Blue, Keith Kranker, Kate Stewart, and Sheila Hoag

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

FINGER LAKES HEALTH SYSTEMS AGENCY

CHAPTER SUMMARY

Introduction. Finger Lakes Health Systems Agency (FLHSA) used its $26.6 million Health Care Innovation Award (HCIA) to improve primary care delivery in 68 practices in the greater Rochester, New York, area. The investment in transforming participating practices into patient-centered medical homes (PCMHs) targeted all patients served by these practices, and its intensive care management services targeted high-risk Medicare and Medicaid beneficiaries—estimated to be about 1 percent of practices' patients. FLHSA aimed to reduce the total cost of care by 3 percent by improving intermediate health outcomes and quality of care for all patients—particularly high-risk Medicare and Medicaid beneficiaries—thus reducing potentially preventable hospital admissions, hospital readmissions, and avoidable emergency department (ED) visits. FLHSA received the award in 2012 and implemented its HCIA intervention with three cohorts of participating practices over 30 months from January 2013 to June 2015. The Centers for Medicare & Medicaid Services allowed FLHSA to extend the intervention by 12 months—from July 2015 to June 2016—and spend unused funds. The program design remained largely similar during the extension period, although care managers' salaries were not covered during the extension period (as they were during the award period).

Objectives. The third annual report (Blair et al. 2017) presented implementation and impact findings from the original award period for the first two cohorts of practices that participated in the FLHSA intervention. This report updates impact and implementation findings to include the 12-month no-cost extension period. Updating impact estimates to include these last 12 months of the intervention is important because FLHSA expected impacts to grow over time, such that impacts for Cohort 1 and 2 practices would be larger during the extension period than in preceding years. Including these additional 12 months also enables us to include the third cohort of practices in the impact evaluation for some outcomes, as these practices were expected to start registering impacts during the extension period, corresponding to their second full year of the intervention. Due to data availability constraints related to the conversion of *International Classification of Diseases,* 9th edition (ICD-9) to 10th edition (ICD-10) diagnosis codes in Medicare claims in late 2015, this report updates our findings from the previous report only for the service use and spending outcomes. The outcomes in our other two evaluation domains, quality-of-care processes and quality-of-care outcomes, cover largely the same time periods and cohorts as presented in the third annual report (Blair et al. 2017), although we have updated the analysis with refreshed claims data for that period.

Methods. To estimate impacts, we compared outcomes for Medicare fee-for-service (FFS) beneficiaries served by 55 of the 68 participating practices with outcomes for Medicare FFS beneficiaries served by 158 matched comparison practices that did not participate in the HCIA program, adjusting for observed differences in outcomes for the two groups during a one-year baseline period. Because FLHSA targeted high-risk Medicare and Medicaid beneficiaries with intensive care management, for most outcomes we estimated the intervention's impact on high-risk Medicare FFS beneficiaries served by the practices, in addition to the intervention's impact on all Medicare FFS beneficiaries.

Implementation. FLHSA implemented the practice transformation and care management components of the intervention largely as planned during the original award period. All enrolled practices successfully hired care managers within six months of enrollment; they worked with clinical advisors to provide high-risk patients with intensive care management services—including connecting patients with services and coaching them on self-management. As of July 2015, FLHSA care managers had provided about 2 percent of their practices' patients with intensive and nonintensive care management services. FLHSA practice improvement advisors also helped practices transform into PCMHs; this included making greater use of electronic health records and implementing weekly care team huddles and monthly care team meetings for communication and planning purposes. Although the practices implemented the intervention as planned, care managers reported that they had insufficient time to manage patient caseloads, associated reporting requirements, other FLHSA commitments, and meetings and reporting tasks outside of the FLHSA intervention. FLHSA continued to deliver its intervention largely as intended during the HCIA no-cost extension period from July 2015 to June 2016. However, 10 of the original 68 practices did not continue participating in the FLHSA intervention during the extension period, partly because the intervention discontinued care managers' salaries.

Impacts on patients' outcomes. The impact estimates indicate that the intervention improved Medicare FFS beneficiaries' outcomes in the quality-of-care processes domain. A 4 percent improvement in inpatient admissions followed by an ambulatory care visit with a primary care or specialist provider within 14 days drove this impact. The favorable impacts were modest in size: they were smaller than the prespecified threshold of 15 percentage points for substantively large effects. It is possible that differences in the 14-day follow-up measure emerged between the treatment and comparison groups *before* the intervention start date; thus, these estimated impacts might reflect, in part, practice improvement efforts outside the scope of the HCIA intervention. The program did not measurably improve outcomes in the quality-of-care outcomes, service use, or spending domains. Statistical power to detect effects was good for the service use domain, so it is most likely that the program had no substantively large effects on outcomes in the domain (outpatient ED visits and inpatient admissions). However, the evaluation was not well powered for outcomes in the quality-of-care outcomes domain (inpatient admissions for ambulatory care-sensitive conditions and 30-day unplanned hospital readmissions) or for spending, so the lack of measured effects might be either because the program truly did not have effects or it did but our tests failed to detect them. It is unclear whether the estimates for Medicare FFS beneficiaries would generalize to Medicare Advantage and Medicaid beneficiaries who were in FLHSA's target population but, due to data availability, not in the evaluation's treatment group.

Conclusion. The impact conclusions fully align with those in our third annual report, even after adding 12 months of data and Cohort 3 practices to the service use and spending domains. As noted in that report, the lack of effects on service use appears not to be due to a failure to engage primary care providers or generally implement the program as planned. Rather, the lack of effects might be a result of (1) unforeseen implementation barriers, including limited staff time for care management and transformation activities; (2) overly ambitious goals given the relatively small portion of patients receiving intensive services; and (3) a nontrivial portion of practices that left the intervention before the no-cost extension period.

Summary of intervention and impact results for FLHSA

Intervention description			
Awardee description		Community health planning and convening organization in Rochester, New York	
Award amount ($ millions)		$26.6 million	
Award extended beyond June 2015?		Yes (12 months)	
Location		6 counties in greater Rochester area[a] (urban, suburban, and rural)	
Target population		All patients served by 68 primary care practices, which enrolled in the intervention in three cohorts	
Interventions		Identified care gaps among the full patient population at participating practices and developed care plans for high-risk patients • 5 HCIA-funded practice improvement advisors helped practices improve team communication, use EHRs to identify care gaps, and streamline workflows • FLHSA paid each PCP $20,000 to participate in the intervention • FLHSA hired 70 care managers to (1) coach high-need patients on self-management, (2) coordinate care with providers, and (3) connect patients with social services	
Metrics of intervention delivered		• Weekly care team huddles and monthly care team meetings at all practices • Care managers hired at all practices • Care manager services provided to 17,484 patients as of June 30, 2015	
Impact evaluation methods			
Core design		Difference-in-differences model with matched comparison group	
Treatment group	Definition	Medicare FFS beneficiaries attributed to 55 practices FLHSA enrolled by July 1, 2013	
	# of beneficiaries during primary test period[b]	9,286 to 20,706 per quarter	
Comparison group definition		Medicare FFS beneficiaries attributed to 158 matched comparison practices	
Impact results: Quality-of-care processes domain			
Ambulatory care visit within 14 days of discharge (% of beneficiaries/quarter)		Comparison mean[c]	67.7%
		Impact estimate (% difference)	+3.0 pp (+4.4%)*
Received recommended lipid test, for patients with IVD (% of beneficiaries/year)		Comparison mean[c]	76.5
		Impact estimate (% difference)	-0.8 pp (-1.0%)
Received an HbA1c test, for patients with diabetes (% of beneficiaries/year)[e]		Comparison mean[c]	88.7
		Impact estimate (% difference)	+0.8 pp (+1.0%)
Received a complete lipid profile, for patients with diabetes (% of beneficiaries/year)[d]		Comparison mean[c]	80.2%
		Impact estimate (% difference)	+2.3 pp (+2.8%)
Combined impact estimate[e]		(+1.8%)**	
Impact conclusion		**Statistically significant favorable effect**	
Impact results: Quality-of-care outcomes domain			
30-day unplanned hospital readmissions (#/1,000 beneficiaries/quarter)		Comparison mean[c]	14.2
		Impact estimate (% difference)	+0.1 (+0.7%)
Inpatient admissions for ACSC conditions (#/1,000 beneficiaries/quarter)		Comparison mean[c]	15.4
		Impact estimate (% difference)	-0.1 (-0.4%)
Combined impact estimate[e]		(+2.0%)[f]	
Impact conclusion		Indeterminate effect	
Impact results: Service use domain			
All-cause inpatient admissions (#/1,000 beneficiaries /quarter)		Comparison mean[c]	80
		Impact estimate (% difference)	+4.2 (+5.2%)
Outpatient ED visits (#/1,000 beneficiaries/quarter)		Comparison mean[c]	168.7
		Impact estimate (% difference)	+3.3 (+1.9%)
Combined impact estimate[e]		(+3.6%)[g]	
Impact conclusion		No substantively large effect	
Impact results: Spending domain			
Medicare Part A and B spending ($/beneficiary/month)		Comparison mean[d]	$828
		Impact estimate (% difference)	+$7 (+0.8%)
Combined impact conclusion[e]		(+0.7%)[h]	
Impact conclusion		Indeterminate effect	

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

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

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

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

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

[d] To assess care processes for people with diabetes, we focused on the two recommended processes that FLHSA targeted: HbA1c tests and lipid profiles.

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

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

[g] FLHSA's combined impact estimate for the service use domain comprises the estimates of two measures in this table (all-cause inpatient admissions and outpatient ED visits among all beneficiaries) and two measures not reported in this table (all-cause inpatient admissions and outpatient ED visits among only high-risk beneficiaries) but that are reported in the full chapter for FLHSA.

[h] FLHSA's combined impact estimate for the spending domain comprises the estimates of one measure in this table (Medicare Part A and B spending among all beneficiaries) and one measure not reported in this table (Medicare Part A and B spending among only high-risk beneficiaries) but that is reported in the full chapter for FLHSA.

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

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

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

ACSC = ambulatory care-sensitive condition; ED = emergency department; EHR = electronic health record; FFS = fee-for-service; FLHSA = Finger Lakes Health Systems Agency; HCIA = Health Care Innovation Award; IVD = ischemic vascular disease; PCP = primary care provider; pp = percentage point.

I. INTRODUCTION

Finger Lakes Health Systems Agency (FLHSA), a community health planning organization and convening agency in Rochester, New York, received a $26.6 million Health Care Innovation Award (HCIA) to implement an intervention to transform primary care processes and delivery in primary care practices in six counties in the greater Rochester area. FLHSA received the award in 2012 and implemented its HCIA intervention with three cohorts of participating practices over the course of 2.5 years from January 2013 to June 2015. In our third annual report (Blair et al. 2017), we estimated the impacts of the FLHSA intervention on quality-of-care processes and outcomes, service use, and Medicare spending during these 2.5 years of program operations. The analysis in that report drew upon data from the first two of FLHSA's three cohorts of participating practices.

In 2015, FLHSA received an extension—at no additional cost to the Centers for Medicare & Medicaid Services (CMS)—to continue to provide services for an additional 12 months, through June 30, 2016. This report updates our third annual report to include impact estimates for Medicare fee-for-service (FFS) beneficiaries through these additional 12 months of program operations of the no-cost extension period, and to include the third cohort of participating practices in impact estimates for some outcomes. Specifically, we update the impact estimates to include the final 12 months of program operations and the third practice cohort for outcomes in two of the four domains we have selected for our impact evaluation: service use and spending. As explained in Section IV.A.1, we do not add the final 12 months of outcomes or the third practice cohort to our impact analyses for most outcomes in the other two study domains—quality-of-care processes or quality-of-care outcomes—due to a change in how Medicare claims record the diagnoses needed to calculate these measures. Using the longest outcome period available for each domain, we draw final conclusions about the FLHSA intervention's impacts in all four study domains. As in the third annual report, we also discuss possible explanations for the observed effects. Other updates from the last report include a brief review of the intervention that FLHSA delivered during the extension period, including relevant quantitative measures of the intervention delivered, and an assessment of baseline equivalence between practices that participated in the FLHSA program (treatment practices) and similar nonparticipating practices (comparison practices) across all three practice cohorts (as opposed to the first two cohorts only).

In Section II of this report, we briefly describe the design of the FLHSA program. Then, in Section III we summarize the model's implementation during both the original award and no-cost extension periods. In Section IV we present impact evaluation methods, characteristics of Medicare FFS beneficiaries at the treatment and comparison practices, and estimated impacts on patients' outcomes over the full intervention period, including the extension period. In Section V we discuss the findings.

II. PROGRAM OVERVIEW

FLHSA recruited practices in three cohorts (or Cohorts 1, 2, and 3), selecting practices that (1) served a large number of Medicare and Medicaid beneficiaries relative to other practices in the region; (2) had used electronic health records (EHRs) for at least six months; (3) had two to seven full-time-equivalent physicians, nurse practitioners, or physician assistants; and (4) demonstrated a sufficient level of readiness to participate in the program, as determined by practices' demonstrated

leadership and level of interest in the intervention. Starting in January 2013 with Cohort 1, FLHSA's intervention had three components:

1. **Practice transformation.** FLHSA practice improvement advisors worked with practice champions—primary care physicians from each practice who served as the main points of contact with FLHSA program staff—and other practice staff to redesign primary care processes, culture, and workforce to transform 68 practices into patient-centered medical homes (PCMHs). Practice improvement advisors held weekly or biweekly meetings with practice staff to identify and work on quality improvement projects to help staff transform their practices. Critical practice transformation activities included weekly huddles (meetings during which the care team reviewed patients' information to prepare for upcoming appointments); monthly care team meetings of primary care providers, nurses, and support staff to discuss specific patients' care plans; and practice champions' participation in learning collaboratives to discuss practice transformation and improvement projects.
2. **Care management.** The care management component focused on providing intensive care management services to high-risk patients, estimated to be about 1 percent of practices' patient population. FLHSA clinical advisors helped participating practices to train, deploy, and support practice-based care managers in providing intensive care management and linking patients with community resources. Care managers, who were fully funded by the HCIA during the initial award period (but not the 12-month extension), screened practice populations to identify high-risk patients who qualified for intensive care management services, and reached out to patients to invite them to participate. Care managers started with 20 to 40 intensive patients each, eventually building up to a caseload of 40 to 60 patients, and contacted these patients at least monthly to check on their conditions, provide behavioral coaching, schedule needed care, and link patients with community resources.
3. **Communitywide outcomes-based payment model.** FLHSA leadership worked with two insurers to develop a communitywide outcomes-based payment model to ensure sustainability of program activities and personnel after the HCIA period. FLHSA leadership expected that the combined shared savings payments to practices would cover continuing practice transformation costs and the cost of employing a care manager after the intervention had ceased.

FLHSA expected practice transformation efforts to improve clinical care by increasing screenings and other preventive care measures for all patients, thereby proactively identifying and treating patients' medical issues. This in turn was expected to reduce inpatient admissions for ambulatory care-sensitive conditions (ACSCs) and all-cause inpatient admissions, which would reduce total spending at participating practices. In addition, FLHSA expected direct care management services and better access to medical care to enhance high-risk patients' clinical care and self-management, thus keeping these patients' chronic conditions under control and reducing outpatient emergency department (ED) visits, inpatient admissions for ACSCs, and all-cause inpatient admissions. These reductions would, in turn, further reduce total spending for high-risk patients. FLHSA's goals were to reduce the total cost of care by 3 percent by reducing potentially preventable hospital admissions by 25 percent, reducing 30-day readmissions by 25 percent, and reducing avoidable ED visits by 15 percent by the end of the award. FLHSA officials expected the

program to have no effects in practices' first year of participation, half of the maximum effect in the second year, and the full effects in the third year and beyond.

The program design remained largely similar during the 12-month no-cost extension period that began in July 2015, although FLHSA adapted its reimbursement for practices' care management and transformation efforts in two key ways. First, award funds no longer covered the salary and benefits of care managers during the extension period, whereas they were fully covered during the original award period. Second, FLHSA implemented a new reimbursement system for practice transformation, whereby practices were paid upon completing deliverables or reaching milestones. (This differed from reimbursement during the original award period, which did not depend on performance.) Although FLHSA changed its reimbursement approach during the extension period, core program activities and tasks remained essentially the same as during the original award period. For example, three of the four deliverables that FLHSA required of its participating practices during the extension period—attending the practice champion learning collaboratives, attending the care manager learning collaboratives, and reporting data via EHRs—were core practice transformation and care management activities during the original award period.

III. PROGRAM IMPLEMENTATION

In this section, we first summarize our assessment of program implementation during the original three-year award period, which we presented in greater detail in the third annual report (Blair et al. 2017). We then review implementation success during the 12-month extension period.

A. Implementation during the original three-year award period

FLHSA implemented the practice transformation and care management components of the intervention largely as planned from January 2013 to June 2015. Several measures capture the generally successful implementation:

- FLHSA met or exceeded its staffing goals for the HCIA-funded intervention. Each practice met FLHSA's goal to employ at least one care manager, resulting in a total of 70 embedded care managers across the 68 practices during the award period.
- As of July 2015, FLHSA care managers had provided services (both intensive and otherwise) to 17,484 distinct patients, exceeding the target cumulative enrollment of 13,564 patients (about half of whom were expected to receive intensive care management) for the entire award period. This was about 2 percent of the approximately 750,000 patients at participating practices.
- Practice improvement advisors met regularly with practice champions and other staff to identify and implement quality improvement projects, improve communication pathways among practice staff, use EHRs to improve care processes, and improve practice workflows. FLHSA staff also organized monthly learning collaboratives to support practice champions and facilitate learning across practices. By July 2015, two-thirds of Cohort 1 practice champions and nearly all Cohort 2 and 3 practice champions had participated in learning collaboratives.
- Practice staff successfully implemented monthly care team meetings and weekly care team huddles to coordinate care. When the program began, only 40 percent of Cohort 1 practices, 11

percent of Cohort 2 practices, and no Cohort 3 practices reported holding monthly care team meetings, but by July 2015, all Cohort 1 practices, 96 percent of Cohort 2 practices, and 75 percent of Cohort 3 practices held monthly care team meetings. Similarly, at the start of program implementation, 40 percent of Cohort 1 practices, 28 percent of Cohort 2 practices, and 88 percent of Cohort 3 practices reported they held weekly huddles, whereas, by July 2015, all practices held weekly huddles.

Although FLHSA implemented the intervention largely as planned in terms of staffing, given the greater use of EHRs and care team huddles and meetings, staff reported having limited time to devote to transformation activities. For practice champions, the HCIA-funded intervention was only part of their job, and they reported that weekly care team huddles, monthly care team meetings, and learning collaboratives often took more time than they had available. In addition, care managers reported that they often had insufficient time to manage large patient caseloads and associated reporting requirements, combined with additional FLHSA commitments (such as attending learning collaboratives) and other meetings and reporting tasks outside of the FLHSA intervention. Despite their time limitations, practice staff perceived the intervention as a relative advantage compared with the standard delivery of care because of its increased emphasis on the care team, the presence of a practice-based care manager, and improved communication with patients.

The third intervention component, working with two local insurers to implement a community outcomes-based payment model, was not implemented as planned. Over the original award period—and unanticipated by FLHSA when it applied for HCIA funding—two regional accountable care organizations (ACOs) formed, and most practices participating in the HCIA-funded intervention joined these ACOs. Although FLHSA did develop its planned outcomes-based payment model during the award period, the model was not widely used because ACOs and local insurers could not agree on outcomes-based payments, particularly with respect to compensation for care management.

All 68 practices in all three cohorts fulfilled their initial two-year commitment to the program. However, three Cohort 1 practices left the demonstration in early 2015. This left a total of 65 active practices by June 2015, the end of the original award period.

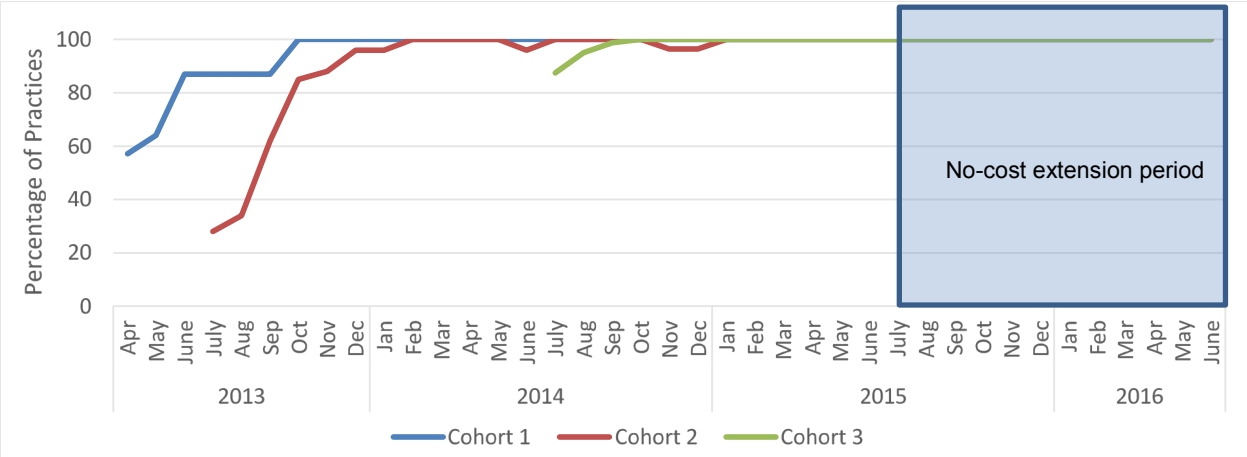
B. Implementation during the 12-month extension period

Participating practices could decide to continue or discontinue their transformation and care management efforts during the no-cost extension period. Most practices continued into the extension period despite the loss of care manager salaries by paying care managers through other sources. However, seven practices (one Cohort 1, five Cohort 2, and one Cohort 3 practice) ceased participation before the extension period began. The most common reasons for dropping out of the intervention, according to FLHSA reports, were practices' inability to pay a care manager and inability to devote staff time to practice transformation.

Among the 58 remaining practices, FLHSA continued to deliver its intervention largely as intended during the no-cost extension period from July 2015 to June 2016. The following evidence supports this conclusion:

- Care managers remained in 56 of the 58 remaining practices. Care managers continued to review population-based reports, participate in learning collaboratives, and provide high-risk patients with intensive care management services. (Unfortunately, there are no data on the number of active care management cases during the extension period because practices stopped reporting this metric to FLHSA at the end of the original award period.) In addition, all active care managers in Cohort 1 and 2 practices—and most active care managers in Cohort 3 practices—participated in learning collaboratives.
- All 58 remaining practices continued the practice transformation component of the intervention under the guidance of FLHSA practice improvement advisors. At the end of the quarter running from April to June 2016, all practices in all cohorts continued to hold at least one huddle per week (Figure III.1) and most practices continued to complete care team meetings each month (Figure III.2). At the end of the no-cost extension period, practice champions' participation in learning collaboratives was 50 percent for Cohort 1, 85 percent for Cohort 2, and 100 percent for Cohort 3 (not shown in figures).

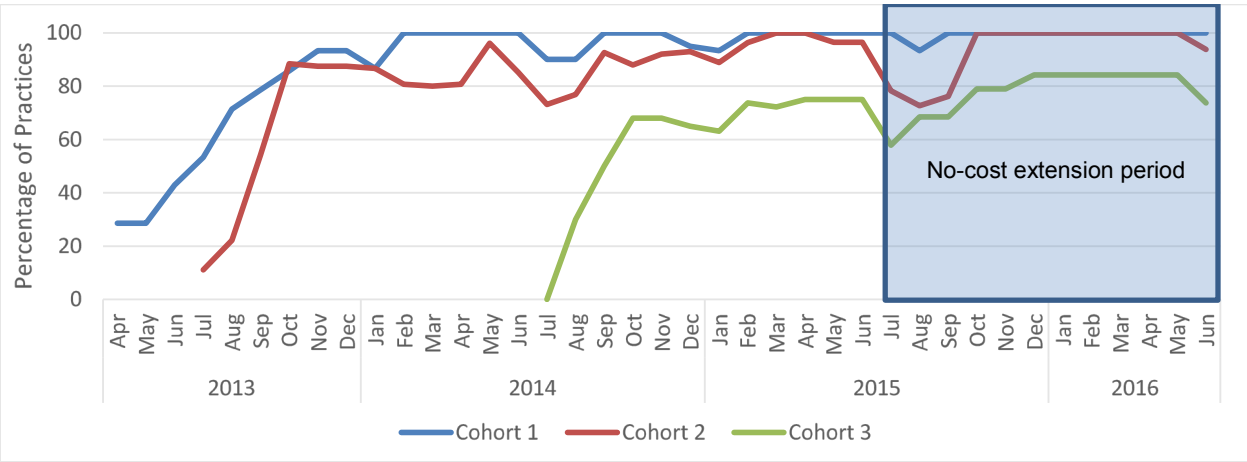
Figure III.1. Practices that held weekly care team huddles, by cohort



Source: FLHSA monitoring data.

FLHSA = Finger Lakes Health Systems Agency.

Figure III.2. Practices that held monthly care team meetings, by cohort



Source: FLHSA monitoring data.
FLHSA = Finger Lakes Health Systems Agency.

IV. PROGRAM IMPACTS ON PATIENTS' OUTCOMES

This section of the report draws final conclusions about the impacts of FLHSA's HCIA program on patients' outcomes in four domains: quality-of-care processes, quality-of-care outcomes, service use, and spending. Our impact conclusions for the service use and spending domains reflect the full 3.5 years of program operations, including the 12-month no-cost extension period. However, due to constraints in available data, described later, our conclusions for the other two domains are based on the first 2.0 to 2.5 years of program operations only. The impact conclusions presented in this chapter represent the combined effect of the intervention's practice transformation and care management components, but do not reflect FLHSA's implementation of an outcomes-based payment model. As discussed in Section III.A, FLHSA developed a new outcomes-based payment model with local insurers, but did not use the model as anticipated during the original award period or no-cost extension period.

We first summarize the methods for estimating impacts, which are the same as those in the third annual report (which describes them more fully). Next, we discuss the characteristics of Medicare FFS beneficiaries served by treatment practices at baseline—that is, when the program started operations—and we assess baseline equivalence between beneficiaries at treatment and comparison practices. Lastly, we describe the impact estimates, their plausibility given implementation findings, and our conclusions about program impacts in each domain.

A. Methods

1. Overview

We estimated the impact of FLHSA's HCIA program on patients' outcomes using a difference-in-differences design. This model estimated impacts as the regression-adjusted difference in outcomes for Medicare FFS patients served by 58 treatment practices and those served by 158

matched comparison practices, subtracting out the difference in outcomes between these groups during a year-long baseline period. We conducted an intent-to-treat analysis, and thus included in our treatment group the 10 treatment practices that dropped out of the FLHSA intervention before the no-cost extension period. Keeping all practices in our treatment group enables us to assess the effect of the FLHSA intervention under real-world circumstances, in which practice participation is not 100 percent during the entire award and extension periods.

We prespecified primary tests to determine the impact of the HCIA program. These tests outlined the evidence we would need to conclude that the program was effective, and the awardee and staff at the Center for Medicare & Medicaid Innovation (CMMI) reviewed these. Each test specified a population, outcome, time period, expected direction of effect, and threshold that we count as substantively important. The purpose of these primary tests was to focus the impact evaluation on hypotheses that would provide the most robust evidence about program effectiveness. We used the results from the primary and secondary tests (robustness checks) to draw conclusions about program impacts in each of the four evaluation domains.

To implement the difference-in-differences design and conduct the primary tests, we used Medicare claims for beneficiaries served by the treatment and comparison practices to develop nine outcome variables in four domains, as well as covariates that captured beneficiaries' characteristics before the intervention began. The domains, outcome variables, and covariates are identical to those defined in the third annual report. (Appendix 1 in the third annual report describes construction of the outcomes and covariates and Appendix 2 lists the covariates.) The outcomes in the quality-of-care outcomes, service use and spending domains are defined quarterly—for example, the number of inpatient admissions per 1,000 beneficiaries per quarter, or the mean per-beneficiary-per-month Medicare Part A and B spending in a quarter—whereas most outcomes in the quality-of-care processes domain are defined annually. We used an ordinary least squares regression model to estimate impacts. The model used robust standard errors to account for clustering of outcomes across quarters for the same beneficiary and a dummy variable for each practice (fixed effects) to account for clustering of outcomes for beneficiaries assigned to the same practice. We averaged the impact estimates across relevant quarters (or years) to generate our impact estimates for the primary test periods. We used one-sided tests to assess the statistical significance of these impact estimates, using a threshold for significance of $p = 0.1$.

We made one important change to the impact design we expected to implement when we wrote our third annual report. We were unable to include the final 12 months of program operations—and by extension, data for the third cohort of practices—in most impact estimates for two study domains, quality-of-care processes and quality-of-care outcomes, because those final 12 months fell primarily after providers switched from using International Classification of Diseases, 9th edition (ICD-9) to 10th edition (ICD-10) codes for describing a person's medical conditions on claims. (This switch took effect October 1, 2015.) All of the measures in the quality-of-care processes and outcomes domains rely on diagnosis codes recorded in claims. Thus the switch from ICD-9 to ICD-10 codes prevented us from calculating the outcome measures in the same way across the full study period, which is essential for estimating unbiased impact estimates. Therefore, for these two domains, we limited the impact estimates entirely to the ICD-9 period, which generally fell in the

first 2.5 years of program operations. As a result, findings for these two domains are similar to those in the third annual report, as they are based on largely the same data and cohorts.

2. Definition of treatment and comparison groups

Treatment group. The treatment group consisted of Medicare FFS patients served by 16 Cohort 1 practices, 21 Cohort 2 treatment practices, and 18 Cohort 3 practices (for a total of 55 treatment practices among the 68 participating practices in all three cohorts), which we observed in 4 quarters before the intervention began (or baseline quarters) and up to 14 quarters after the intervention began (or intervention quarters). The baseline period was January 1, 2012, to December 31, 2012, for Cohort 1 practices; July 1, 2012, to June 30, 2013, for Cohort 2 practices; and July 1, 2013, to June 30, 2014, for Cohort 3 practices. Cohort 1 practices had 14 intervention quarters totaling 42 months (from January 1, 2013, to June 30, 2016); Cohort 2 practices had 12 intervention quarters totaling 36 months (from July 1, 2013, to June 30, 2016); and Cohort 3 practices had 8 intervention quarters totaling 24 months (from July 1, 2014, to June 30, 2016).

We excluded 13 participating practices from the research sample—7 practices because they were federally qualified health centers (FQHCs) and had no suitable comparison. (We attempted to match these FQHCs to nonparticipating FQHCs in New York State using practice-level data from the Health Resources & Services Administration Data Warehouse. However, we found that participating FQHCs were much larger in size and their patient populations differed systematically from other FQHCs in the state.) We excluded another 2 practices because they served psychiatric or pediatric populations only, and 4 practices because they had no attributed Medicare patients in at least one quarter of the evaluation baseline period, and were thus incompatible with the statistical regression model used to measure impacts (described in Appendix 2).

We constructed the treatment group in three steps. First, in each baseline and intervention month, we attributed beneficiaries to the primary care practice whose providers (physicians, nurse practitioners, or physician assistants) provided the plurality of primary care services in the past 24 months. Second, in each baseline and intervention period, we assigned each patient to the first treatment practice to which he or she was attributed in that period, and continued to assign him or her to that practice for all quarters in the period. This assignment rule—which is distinct from the attribution method—ensures that patients did not exit the treatment group during the intervention period solely because the intervention succeeded in reducing their service use (including visits at treatment practices). Third, we applied additional restrictions to refine the analysis sample in each quarter. A patient assigned to a treatment practice in a quarter was included in the analysis sample for that quarter if he or she (1) had observable outcomes for at least one day in the quarter and (2) lived in New York for at least one day of the quarter. (Outcomes were observable if beneficiaries were enrolled in Medicare FFS [Part A and B], were alive, and had Medicare as their primary payer.)

In addition to this full treatment sample, we defined a subset of patients who were at high risk of hospitalizations and other expensive medical care. This high-risk subgroup enabled us to conduct primary tests examining whether any observed effects concentrated among high-risk members. This would be expected, given that FLHSA targeted its care management services to high-risk Medicare and Medicaid beneficiaries. In each baseline quarter, we defined the evaluation's high-risk

subgroup to consist of Medicare FFS beneficiaries with a Hierarchical Condition Category (HCC) score in the top third of all treatment group members with observable outcomes at the start of the baseline period.

Comparison group. The comparison group consisted of Medicare FFS beneficiaries we assigned to 158 matched comparison practices in each of the baseline and intervention quarters. The comparison practices were similar to the treatment practices during the baseline period on factors that could influence patients' outcomes, especially those factors that FLHSA used when deciding which practices to recruit for the intervention. (See the third annual report for more detail on how we identified the comparison group.) Section IV.C shows the balance we achieved between the two groups on the matching variables.

We identified the 158 comparison practices in several steps. First, we limited the potential comparison practices to the approximately 2,000 non-FQHC primary care practices in New York State located outside of the greater New York City area, the 6 counties in which FLHSA operated, and 13 counties in New York that had relatively strong participation in federal primary care initiatives. Second, we constructed matching variables, defined before the start of the intervention for all treatment and potential comparison practices. These variables include characteristics of the practices, of all Medicare FFS beneficiaries assigned to the practices, and of high-risk beneficiaries assigned to the practices. (Section IV.C provides additional detail on matching data and results.) Third, we narrowed the pool of potential comparison practices by excluding those practices that participated in either the Multi-Payer Advanced Primary Care Practice demonstration or the Comprehensive Primary Care initiative or had an average of fewer than 25 assigned Medicare FFS beneficiaries during any of the four baseline quarters. Fourth, we used propensity-score methods to select 158 comparison practices that were similar to the 55 treatment practices on the matching variables. After completing the matching process, we assigned Medicare FFS beneficiaries to the comparison practices in each intervention quarter, applying the same rules we used for the treatment group. We also defined a high-risk subgroup of the comparison group using the same rules as for the treatment group.

3. Revised primary tests

Table IV.1 shows the primary tests for FLHSA, by domain. We have revised these tests since the third annual report, reflecting our understanding of recent changes in Medicare diagnosis codes. Only the primary test time periods have changed since the third annual report; the outcomes, populations, and substantive thresholds remain unchanged.

Our rationale for selecting these primary tests is as follows:

- **Outcomes.** FLHSA's central goal was to reduce ED visits, 30-day unplanned readmissions, ACSC admissions, and total medical spending. FLHSA did not explicitly state that it expected to reduce all-cause hospital admissions. However, through the expected reductions in ACSC admissions, FLHSA would also reduce all-cause admissions (although by a smaller percentage change). We thus assessed program effects on all five of these outcomes. We also included four quality-of-care process measures that, based on FLHSA's program design and core monitoring indicators, we think the program could improve: (1) a measure for whether a beneficiary with

diabetes received an HbA1c test, (2) a measure for whether a beneficiary with diabetes received a lipid profile, (3) receipt of a complete lipid profile among people with ischemic vascular disease (IVD), and (4) receipt of a follow-up ambulatory care visit with a primary care or specialist provider within 14 days of hospital discharge.

- **Time period.** In determining the time period for primary tests, we used the guiding principle of maximizing the number of quarters and cohorts for which valid data were available for each outcome, after excluding the first year of data (when no effects were expected for most outcomes). For all outcomes in the service use and spending domains, we analyzed program impacts from early 2014 to mid-2016 among patients in Cohort 1 (the last 2.5 years or 10 quarters of implementation), from mid-2014 to mid-2016 for Cohort 2 (the last 2.0 years or 8 quarters of implementation), and from mid-2015 to mid-2016 for Cohort 3 (the last year or 4 quarters of implementation). This corresponds to intervention quarters 5 through 14 (I5 through I14) for Cohort 1, I5 through I12 for Cohort 2, and I5 through I8 for Cohort 3. Due to data availability constraints associated with the ICD-10 conversion in October 2015, our analysis of quality-of-care process measures for IVD and diabetes is limited to quarters I5 through I8 for Cohorts 1 and 2 and our analysis of the 14-day follow-up measure following discharge and 30-day readmissions is limited to quarters I5 through I10 for Cohort 1 and I5 through I8 for Cohort 2. Also related to the ICD-10 conversion issue, our analysis of ACSC admissions is limited to quarters I5 through I11 for Cohort 1, I5 through I9 for Cohort 2, and I5 for Cohort 3.
- **Population.** FLHSA's practice transformation and care management components aimed to generate impacts among all patients, but the impacts of care management services were expected to be concentrated among high-risk patients. To capture potential effects on all Medicare beneficiaries as well as high-risk Medicare beneficiaries, we included both groups in our primary tests on quality-of-care outcomes, service use, and spending. For the diabetes and IVD quality-of-care process measures, we followed the definitions of existing quality measures and limited the population to beneficiaries ages 18 to 75 with diabetes or ages 18 and older with IVD, respectively, and who were observable in Medicare FFS claims for all 12 months of the measurement year. For the 14-day follow-up measure, we limited the sample in each quarter to those beneficiaries who had at least one index hospitalization during the quarter for which we could observe whether the person had all hospitalizations followed by a 14-day follow-up visit.
- **Direction (sign) of the impact estimate.** We expected the impact estimate to be positive for the quality-of-care process measures, signaling an increase in the percentage of people receiving recommended care. For all other outcomes, we expected the impact estimates to be negative, indicating a reduction in service use or overall expenditures.
- **Substantive thresholds.** Some impact estimates could be large enough to be substantively interesting to CMMI and other stakeholders even if they are not statistically significant; for this reason, we specified thresholds for what we call substantive importance. We expressed the threshold as a percentage change from the counterfactual—that is, the outcomes that beneficiaries in the treatment group would have had if they had not received the treatment.
  - For the full patient population, the 3 and 2 percent thresholds we chose for all-cause inpatient admissions and total spending, respectively, are 75 percent of FLHSA's expected

effects among all three cohorts during the primary test period (I5 through I14). (We use 75 percent recognizing that FLHSA could still be considered successful if it approached, but did not achieve, its fully anticipated effects.) The 5 percent threshold for the remaining outcomes among the full patient population is extrapolated from the literature (Peikes et al. 2011), which suggests that impacts of this size should be considered policy-relevant even though they are smaller than the impacts FLHSA anticipated. (By the third year of the intervention, the awardee expected a decrease of 25 percent in potentially preventable hospitalizations and 30-day hospital readmissions, and a decrease of 15 percent in ED visits among its full patient population.)

- For the high-risk patient population, the 5 and 3 percent thresholds we chose for all-cause inpatient admissions and total spending, respectively, are 75 percent of our estimate of FLHSA's expected effects among high-risk beneficiaries for all three cohorts during the primary test period (I5 through I14). This estimate is based on the percentage of high-risk beneficiaries in the population and their proportion of the full population's utilization and costs. The 15 percent threshold for the remaining outcomes is extrapolated from the literature (Peikes et al. 2011) for the same reason stated earlier. That is, the literature indicates that effects of this size should be considered policy-relevant even though they are smaller than our calculation of FLHSA's expected effects for high-risk beneficiaries).
- The 15 percent threshold for the quality-of-care process measures is also extrapolated from the literature (Peikes et al. 2011; Rosenthal et al. 2016) because FLHSA did not specify by how much it expected to improve these outcomes.

4. Secondary tests (robustness checks)

We also conducted secondary quantitative tests to help corroborate the findings from the primary tests. This is important because differences observed between the treatment and comparison groups in the primary test results could reflect the non-experimental design of our evaluation or random fluctuations in the data. We have greater confidence in the primary test results if they are generally consistent with the expected broader pattern of results.

We conducted three sets of secondary tests for FLHSA:

1. First, we estimated the program's impacts on the full Medicare FFS population and the high-risk Medicare FFS population during the first 12 months after the practices joined the intervention (quarters I1 through I4). We did this for three domains: quality-of-care processes, service use, and spending. For the quality-of-care processes domain, if the primary test results suggested favorable findings, we would expect to find positive impacts during the first year. This is because we expected impacts on quality-of-care processes to materialize in the first 12 months of the intervention, given that all treatment practices hired care managers within 3 months of the intervention's start date and they began providing care management services shortly thereafter. Because we and FLHSA expected program impacts on service use and spending to increase over time, with little or no impacts in the first few months of the program, we expected no measurable effects in the first 12 intervention months in the service use and spending domains. If we found large differences in outcomes (favorable or unfavorable) in these domains in the first 12 intervention months, this could suggest a limitation in the

comparison group, not true program impacts. We excluded quality-of-care outcomes from this set of secondary tests, as FLHSA did not state whether it anticipated effects in this domain during the first program year.

2. Second, we reestimated impacts on all outcomes among the full Medicare FFS population and the high-risk Medicare FFS population, limiting the sample to beneficiaries assigned to the treatment and comparison groups by the start of the period, either baseline or intervention. This restriction prevents addition to the intervention sample over time. It is possible that differences in sample addition between the treatment and comparison groups could bias the impact results to some degree if the sample members added over time differed from earlier sample members (for example, if they were younger and healthier). This could create differences in mean outcomes between the treatment and comparison groups that were unrelated to the HCIA-funded intervention. We have explored this possibility because, as we will describe in Section IV.D, the rate of net sample growth during the intervention period was slightly higher for the comparison group (growth of 17.1 percent from I1 to I8) than for the treatment group (growth of 12.4 percent over the same period) for the service use and spending domains.

3. Third, we reestimated impacts on all outcomes among the full and the high-risk Medicare FFS populations, limiting the sample to beneficiaries assigned to Cohort 1 and 2 treatment and comparison practices. Excluding Cohort 3 practices from impact estimates controls for highly differential sample addition in Cohort 3 that, as we describe in Section IV.D.1, occurred from I1 to I8: the rate of net sample growth during the intervention period was much higher for the Cohort 3 comparison group (growth of 35.3 percent from I1 to I8) than for the Cohort 3 treatment group (growth of 13.3 percent over the same period). This set of secondary tests assesses whether Cohort 3 sample members who joined practices during the intervention period might bias our impact results, while allowing for sample addition in the other two cohorts.

Table IV.1. Specification of the primary tests for FLHSA

Domain (number of tests in the domain)[a]	Outcome (units)	Time period for impacts (controlling for baseline differences)[b]	Population	Expected direction of effect (+ or -) and substantive threshold (impact as percentage of the counterfactual)[c]
Quality-of-care processes (4)	Received an HbA1c test (binary [yes or no]/beneficiary/year)	Average over intervention quarters 5–8 for Cohorts 1 and 2[d]	Medicare FFS beneficiaries assigned to treatment groups with diabetes and ages 18 to 75	15.0% (+)
	Received a lipid profile(binary [yes or no]/beneficiary/year)	Average over intervention quarters 5–8 for Cohorts 1 and 2[d]	Medicare FFS beneficiaries assigned to treatment groups with diabetes and ages 18 to 75	15.0% (+)
	Received complete lipid profile in the year (binary [yes or no]/beneficiary/year)	Average over intervention quarters 5–8 for Cohorts 1 and 2[d]	Medicare FFS beneficiaries assigned to treatment groups with IVD and ages 18 or older	15.0% (+)
	All inpatient admissions within a quarter were followed by an ambulatory care visit with a primary care or specialist provider within 14 days (binary [yes or no]/beneficiary/year)	Average over intervention quarters 5–10 for Cohort 1 and 5–8 for Cohort 2	Medicare FFS beneficiaries assigned to treatment groups with at least one hospital stay in the quarter	15.0% (+)
Quality-of-care outcomes (4)	Inpatient admissions for ambulatory care-sensitive conditions (#/beneficiary/quarter)	Average over intervention quarters 5–11 for Cohort 1, 5–9 for Cohort 2, and 5 for Cohort 3	Medicare FFS beneficiaries assigned to treatment groups	5.0% (-)
	30-day unplanned hospital readmissions (#/beneficiary/quarter)	Average over intervention quarters 5–10 for Cohort 1 and 5–8 for Cohort 2	Medicare FFS beneficiaries assigned to treatment groups	5.0% (-)
	Inpatient admissions for ambulatory care-sensitive conditions (#/beneficiary/quarter)	Average over intervention quarters 5–11 for Cohort 1, 5–9 for Cohort 2, and 5 for Cohort 3	High-risk Medicare FFS beneficiaries assigned to treatment groups	15.0% (-)
	30-day unplanned hospital readmissions (#/beneficiary/quarter)	Average over intervention quarters 5–10 for Cohort 1 and 5–8 for Cohort 2	High-risk Medicare FFS beneficiaries assigned to treatment groups	15.0% (-)
Service use (4)	All-cause inpatient admissions (#/beneficiary/quarter)	Average over intervention quarters 5–14 for Cohort 1, 5–12 for Cohort 2, and 5–8 for Cohort 3	Medicare FFS beneficiaries assigned to treatment groups	3.0% (-)
	Outpatient ED visit rate (#/beneficiary/quarter)	Average over intervention quarters 5–14 for Cohort 1, 5–12 for Cohort 2, and 5–8 for Cohort 3	Medicare FFS beneficiaries assigned to treatment groups	5.0% (-)
	All-cause inpatient admissions (#/beneficiary/quarter)	Average over intervention quarters 5–14 for Cohort 1, 5–12 for Cohort 2, and 5–8 for Cohort 3	High-risk Medicare FFS beneficiaries assigned to treatment groups	5.0% (-)
	Outpatient ED visit rate (#/beneficiary/quarter)	Average over intervention quarters 5–14 for Cohort 1, 5–12 for Cohort 2, and 5–8 for Cohort 3	High-risk Medicare FFS beneficiaries assigned to treatment groups	15.0% (-)

Table IV.1 *(continued)*

Domain (number of tests in the domain)[a]	Outcome (units)	Time period for impacts (controlling for baseline differences)[b]	Population	Expected direction of effect (+ or -) and substantive threshold (impact as percentage of the counterfactual)[c]
Spending (2)	Medicare Part A and B spending ($/beneficiary/month)	Average over intervention quarters 5–14 for Cohort 1, 5–12 for Cohort 2, and 5–8 for Cohort 3	Medicare FFS beneficiaries assigned to treatment groups	2.0% (-)
	Medicare Part A and B spending ($/beneficiary/month)	Average over intervention quarters 5–14 for Cohort 1, 5–12 for Cohort 2, and 5–8 for Cohort 3	High-risk Medicare FFS beneficiaries assigned to treatment groups	3.0% (-)

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

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

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

[d] For most measures, we take the average across quarterly impact estimates. For the diabetes and IVD quality-of-care process measures, we used an annual impact estimate for the second program year (corresponding to intervention quarters 5 through 8).

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

B. Characteristics of the treatment group at baseline

This section describes the characteristics of the treatment group at the start of the intervention (January 1, 2013, for Cohort 1; July 1, 2013, for Cohort 2; and July 1, 2014, for Cohort 3). This description updates the one we presented in the third annual report by adding Cohort 3. We also show this information in the second column of Table IV.2, which has the characteristics of all treatment practices, pooled across all three cohorts. (Table IV.2 serves a second purpose—to show the equivalence of the treatment and comparison practices in all three cohorts at the start of the intervention—which we describe in Section IV.C.)

Characteristics of the practices overall. Our analysis includes 55 treatment practices at the start of the intervention, none of which were FQHCs. Because we conduct an intent-to-treat analysis, these 55 include the 10 practices that dropped out of the FLHSA intervention before the no-cost extension period. Almost all treatment practices had providers receiving payment from CMS for meaningful use of EHRs (96 percent). This latter proportion is consistent with FLHSA's targeting, as one of the program's eligibility criteria was an EHR system that practice staff used actively for at least a year. Treatment practices had 5.5 total clinicians, on average. The large majority of practices' clinicians in the treatment group had a primary care specialty.

Characteristics of the practices' Medicare FFS beneficiaries. The demographic characteristics of all Medicare FFS beneficiaries assigned to the treatment group during the baseline period were, overall, comparable to nationwide Medicare FFS averages. Beneficiaries in the treatment group also had hospital and ACSC admission rates, 30-day readmission rates, and HCC scores that were comparable to national averages. However, the mean outpatient ED visit rate (148 per 1,000 beneficiaries per quarter) was higher than the national average of 105. In part, this might reflect the proportion of dually eligible beneficiaries in treatment practices, which, at 31 percent, is higher than the national average of 20 percent among Medicare FFS beneficiaries. People dually enrolled in Medicare and Medicaid tend to have higher ED visit rates than Medicare beneficiaries who are not dually enrolled (Medicare Payment Advisory Commission 2016).

Characteristics of the practices' high-risk Medicare FFS beneficiaries. The high-risk beneficiaries in the treatment group had substantially greater health care needs during the baseline period than the full treatment group (Table IV.2). Their mean HCC risk score was more than twice the mean for all treatment group members (2.3 versus 1.1), consistent with how the group was defined. Further, they had more than twice the all-cause inpatient admissions and Medicare spending than the full population of attributed beneficiaries.

Table IV.2. Characteristics of treatment and comparison practices before the intervention start date (January 1, 2013, for Cohort 1; July 1, 2013, for Cohort 2; and July 1, 2014, for Cohort 3)

Characteristic	Treatment practices (N = 55)	Unmatched comparison pool (N = 537)	Matched comparison group (N = 158)	Absolute difference[a]	Standardized difference[b]	Medicare FFS national average
Exact-match variable[c]						
Non-FQHC	100.0	100.0	100.0	0	0	n.a.
Propensity-matched variables[d]						
Characteristics of a practice's location(s)						
Located in an urban zip code (%)	83.6	80.9	87.3	-3.7	-0.09	NA
Zip code poverty rate (%)[e,f]	14.8	13.8	17.0	-2.1	-0.18	NA
Located in a health professionals shortage area (primary care)[f]	96.4	56.4	75.3	21.1	0.49	NA
Medicare Advantage penetration rate[f]	56.7	31.4	37.7	19.1	1.54	NA
Characteristics of all Medicare FFS patients attributed to practices during the baseline year[g]						
Number of FFS beneficiaries	351.7	404.5	362.8	-11.1	-0.05	n.a.
HCC risk score	1.12	1.17	1.13	-0.01	-0.07	1.0
All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	79.5	81.8	77.9	1.6	0.08	74[h]
Outpatient ED visit rate (#/1,000 beneficiaries/quarter)	147.7	124.6	149.6	-1.9	-0.03	105[i]
Medicare Part A and B spending ($/beneficiary/month)	730	773	741	-12	-0.07	860[j]
30-day unplanned hospital readmission (#/1,000 beneficiaries/quarter)	13.1	12.1	12.0	1.1	0.15	NA
Inpatient admissions for ACSCs (#/1,000 beneficiaries/quarter)	14.9	15.7	14.1	0.7	0.11	11.8[k]
Dually eligible beneficiaries (%)	30.8	19.3	31.4	-0.6	-0.03	19.9[l]
Disability as original reason for Medicare entitlement (%)	43.6	28.9	41.0	2.6	0.15	16.7[m]
Age (years)	67.1	71.3	67.3	-0.3	-0.05	71[n]
Female (%)	59.2	58.2	58.3	1.0	0.13	54.7[m]
Race: white (%)	81.3	88.5	83.9	-2.6	-0.14	81.8[m]
Receipt of recommended lipid profile, among those with diabetes ages 18 to 75 (%)	85.7	86.3	84.8	0.8	0.08	NA
Receipt of recommended hemoglobin A1c test, among those with diabetes ages 18 to 75 (%)	90.5	89.1	90.5	-0.0	-0.01	NA
Receipt of recommended lipid profile, among those with IVD ages 18 or older (%)	79.5	81.7	79.4	0.1	0.01	NA
Receipt of an ambulatory care visit within 14 days of any hospital discharges in the quarter, among those with at least one discharge in the quarter (%)	67.8	62.4	66.8	1.1	0.12	NA

Table IV.2 *(continued)*

Characteristic	Treatment practices (N = 55)	Unmatched comparison pool (N = 537)	Matched comparison group (N = 158)	Absolute difference[a]	Standardized difference[b]	Medicare FFS national average
Characteristics of high-risk Medicare FFS patients attributed to practices during the baseline year[g]						
Number of high-risk FFS beneficiaries	85.0	101.2	87.4	-2.4	-0.04	n.a.
HCC risk score	2.27	2.34	2.28	-0.01	-0.06	n.a.
All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	179.2	179.4	179.4	-0.3	-0.00	74[h]
Outpatient ED visit rate (#/1,000 patients/quarter)	257.7	204.0	242.9	14.8	0.12	105[i]
Medicare Part A and B spending ($/beneficiary/month)	1519	1570	1530	-11	-0.03	860[j]
30-day unplanned hospital readmission (#/1,000 beneficiaries/quarter)	37.0	32.8	35.3	1.7	0.07	NA
Inpatient admissions for ACSCs (#/beneficiary/quarter)	40.1	41.6	41.4	-1.3	-0.06	11.8[k]
Characteristics of the practices						
Meaningful use of EHR (%)[o]	96.4	75.1	95.5	0.8	0.03	n.a.
Patient-centered medical home[p]	20.0	9.2	17.1	2.9	0.08	NA
Owned by hospital or health system (%)	67.3	31.5	65.0	2.3	0.05	n.a.
Number of clinicians at practice	5.5	3.4	5.4	0.1	0.04	n.a.
Practices' clinicians with a primary care specialty (%)	94.2	92.7	92.1	2.2	0.13	n.a.

Sources: Analysis of the Medicare Enrollment Database and claims data accessed through the Virtual Research Data Center at CMS. Zip code household income data merged from the American Community Survey ZIP Code Characteristics. Characteristics of the practices come from SK&A, a health care data vendor, and the National Committee for Quality Assurance.

Notes: The characteristics for the treatment and their matched comparison practices are defined at the time the treatment practice joined the intervention (January 1, 2013, for Cohort 1 practices; July 1, 2013, for Cohort 2 practices; and July 1, 2014, for Cohort 3 practices).

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

Absolute differences might not be exact due to rounding.

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

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

[c] Exact match means that we required that non-FQHCs match only to non-FQHCs.

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

[e] Average poverty rate associated with each practice's zip code, merged from the American Community Survey.

[f] The propensity-score model did not include these variables due to concerns that they would generate potential imbalances among the critical matching variables; crucial matching variables include all variables on patient and practice characteristics in this table.

[g] The baseline year is January 1, 2012, to December 31, 2012, for Cohort 1 practices; July 1, 2012, to June 30, 2013, for Cohort 2 practices; and July 1, 2013, to June 30, 2014, for Cohort 3 practices.

[h] Health Indicators Warehouse (2014b).

[i] Gerhardt et al. (2014).

[j] Boards of Trustees (2013).

[k] This rate is for beneficiaries ages 65 and older (Truven Health Analytics 2015).

[l] Medicare Payment and Advisory Commission (2016).

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

[n] Health Indicators Warehouse (2014b).

Table IV.2 *(continued)*

[o] Meaningful use of EHRs is calculated as the percentage of practices with at least one provider (NPI) working in the practice who received financial incentives for meaningful use of certified EHRs through Medicare or Medicaid during the baseline period. Data on meaningful use of EHRs were merged from CMS data.

[p] NCQA Patient-Centered Medical Home (PCMH) Recognition. Data on practices with NCQA recognition were merged from the NCQA database.

ACSC = ambulatory care-sensitive condition; CMS = Centers for Medicare & Medicaid Services; ED = emergency department; EHR = electronic health record; FFS = fee-for-service; FQHC = federally qualified health center; HCC = Hierarchical Condition Category; IVD = ischemic vascular disease; NCQA = National Committee for Quality Assurance; NPI = National Provider Identifier.

NA = not available.

n.a. = not applicable.

C. Equivalence of treatment and comparison groups at baseline

Demonstrating similarity between the treatment and comparison groups at the start of the intervention is important for the evaluation design. This similarity increases the credibility of a key assumption underlying difference-in-differences models—that the change over time in outcomes for the comparison group is the same change that would have happened for the treatment group, had the treatment group not received the intervention. In the third annual report, we assessed equivalence across only the first two cohorts. As we now include Cohort 3 practices in the primary tests, we must assess baseline treatment–comparison equivalence across all three cohorts.

Table IV.2 shows that the 55 treatment practices and the 158 selected comparison practices were similar at the start of the intervention on most matching variables. By construction, there were no differences between the two groups on the exact matching variable of whether the practice was an FQHC. The treatment and matched comparison group beneficiaries differed somewhat on the variables we matched through propensity scores (the second and third panels of Table IV.2 for all beneficiaries and high-risk beneficiaries, respectively), but the standardized differences across the propensity-score matching variables were all within our target of 0.25 standardized differences, and most were within 0.15 standardized differences (the 0.25 target is an industry standard; for example, see Institute of Education Sciences [2014]). This includes patients' demographic characteristics, Medicare FFS beneficiaries' and high-risk Medicare FFS beneficiaries' utilization and costs, and four quality-of-care process measures. Similarly, all differences between treatment and comparison group practices' characteristics (the fourth panel of Table IV.2) are within our target of 0.25 standardized differences. This includes practices' EHR use, medical home designation, ownership, and number of clinicians.

However, there are two important treatment–comparison differences in the characteristics of practices' locations (the first panel of Table IV.2). Namely, the Medicare Advantage penetration rate in counties with treatment practices was higher than that of comparison practices. In addition, a higher proportion of treatment practices than comparison practices were located in a health professionals shortage area (96 versus 75 percent of comparison practices). The difference-in-differences impact estimation model accounts for these baseline treatment–comparison differences. However, given differential Medicare Advantage penetration rates between treatment and comparison groups, it is important to conduct a sensitivity test that prohibits sample addition during the intervention period (Section IV.A.4). This sensitivity test assesses the extent of any possible bias that could result from differential sample addition to

treatment and comparison groups over time (although it cannot correct for bias that might result from differential sample attrition to Medicare managed care).

D. Beneficiaries' outcomes and intervention impacts

In this section, we first present sample sizes and mean outcomes, by quarter, for the treatment and comparison groups. These mean outcomes provide context for understanding the difference-in-differences estimates that follow. (We show means for all quarters, not just those included in primary tests, to assess trends over the entire baseline and intervention periods.) Notably, these differences in mean outcomes are not regression-adjusted and not impact estimates by themselves. Next, we present the results of the primary tests, by domain. Then we present the results of the secondary tests (robustness checks) and assess whether the primary test results are plausible given the secondary test results and the implementation evidence. We end with conclusions about program impacts in each domain.

1. Sample sizes

In this analysis, sample sizes vary substantially by outcome and quarter (Tables IV.3 through IV.7). This reflects two phenomena: (1) distinct implementation periods for each of the three cohorts, such that Cohort 1 practices have two more quarters of data than Cohort 2 practices, and six more quarters of data than Cohort 3 practices; and (2) the ICD-10 conversion issue discussed in Section IV.A.1, which affects the number of quarters of available data for several quality-of-care processes and quality-of-care outcomes measures. Due to the ICD-10 conversion issue, the analysis of nearly all outcomes in the quality-of-care processes and quality-of-care outcomes domains excludes Cohort 3 practices and additional quarters. (The sole exception is ACSC admissions, for which we include one additional quarter of follow-up data for Cohort 1 and 2 practices beyond the data shown in the third annual report, as well as one quarter of data for Cohort 3 practices.) As a result, sample sizes in these two domains are largely similar to those presented in the third annual report. In contrast, sample sizes in the service use and spending domains are generally larger than those reported in the third annual report, reflecting the addition of Cohort 3 practices.

Unlike in the third annual report, we observe substantial and differential sample addition between the treatment and comparison groups, primarily driven by the addition of Cohort 3. Specifically, for both the treatment and comparison groups, total sample sizes increased from I1 to I8 for outcomes in the service use and spending domains, reflecting greater sample addition than attrition over time (Table IV.6). However, the net sample increase from I1 to I8 was lower in the treatment group (12.4 percent) than the comparison group (17.1 percent), largely because of differences in Cohort 3 (a 13.3 percent increase in the treatment group but a 35.3 percent increase in the comparison group; results not shown). (Section IV.A.4 discusses this differential growth in the size of treatment and comparison groups in the context of sensitivity tests.)

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

		Number of Medicare FFS beneficiaries (practices)			Mean outcomes		
Period	Quarter(s)	T	C (not weighted)	C (weighted)	T	C	Difference (%)
Among those with diabetes and ages 18 to 75, received A1c screening (binary [yes or no]/beneficiary/year)							
Baseline	B1–B4[a]	2,208 (37)	5,901 (108)	2,421	90.5	90.0	0.5 (0.5%)
Intervention	I1–I4[a]	2,106 (37)	5,526 (108)	2,310	91.1	89.3	1.8 (2.0%)
	I5–I8[a]	1,926 (37)	5,141 (108)	2,137	89.5	88.8	0.7 (0.8%)
Among those with diabetes and ages 18 to 75, received lipid panel (binary [yes or no]/beneficiary/year)							
Baseline	B1–B4[a]	2,208 (37)	5,901 (108)	2,421	84.1	85.9	-1.8 (-2.1%)
Intervention	I1–I4[a]	2,106 (37)	5,526 (108)	2,310	84.4	85.3	-0.9 (-1.1%)
	I5–I8[a]	1,926 (37)	5,141 (108)	2,137	82.5	82.3	0.1 (0.2%)
Among those with IVD and ages 18 or older, received lipid panel (binary [yes or no]/beneficiary/year)							
Baseline	B1–B4[a]	3,198 (37)	9,778 (108)	3,760	78.8	78.5	0.3 (0.4%)
Intervention	I1–I4[a]	2,953 (37)	9,111 (108)	3,571	78.2	78.7	-0.5 (-0.6%)
	I5–I8[a]	2,685 (37)	8,582 (108)	3,314	75.7	76.5	-0.8 (-1.1%)
Among those with at least one inpatient admission in the quarter, all inpatient admissions in the quarter were followed by an ambulatory care visit with a primary care or specialist provider within 14 days of discharge (binary [yes or no]/beneficiary/year)							
Baseline	B1	925 (37)	2,139 (108)	809	66.3	66.8	-0.5 (-0.8%)
	B2	934 (37)	2,316 (108)	892	67.1	67.9	-0.7 (-1.1%)
	B3	886 (37)	2,292 (108)	851	69.3	66.9	2.4 (3.6%)
	B4	1,001 (37)	2,432 (108)	897	71.1	65.7	5.5 (8.3%)

Table IV.3 *(continued)*

Period	Quarter(s)	Number of Medicare FFS beneficiaries (practices) T	C (not weighted)	C (weighted)	Mean outcomes T	C	Difference (%)
Intervention	I1	858 (37)	2,091 (108)	810	72.0	69.2	2.8 (4.1%)
	I2	920 (37)	2,136 (108)	873	74.2	66.3	7.9 (11.9%)
	I3	907 (37)	2,098 (108)	854	71.7	63.3	8.4 (13.3%)
	I4	919 (37)	2,223 (108)	862	72.5	63.9	8.6 (13.4%)
	I5	948 (37)	2,230 (108)	884	72.0	68.5	3.5 (5.1%)
	I6	897 (37)	2,269 (108)	862	70.3	64.0	6.3 (9.8%)
	I7	984 (37)	2,318 (108)	893	70.1	67.2	2.9 (4.4%)
	I8	959 (36)	2,309 (108)	864	71.4	65.3	6.1 (9.4%)
	I9	654 (16)	1,172 (51)	531	67.3	68.1	-0.8 (-1.2%)
	I10	631 (16)	1,240 (51)	577	72.7	64.6	8.2 (12.7%)

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

Notes: The baseline quarters are measured relative to the start of the baseline period (January 1, 2012, for Cohort 1; July 1, 2012, for Cohort 2; and July 1, 2013, for Cohort 3). For example, the first baseline quarter (B1) for Cohort 1 runs from January 1, 2012, to March 31, 2012. The intervention quarters are measured relative to the start of the intervention period on January 1, 2013, for Cohort 1; July 1, 2013, for Cohort 2; and July 1, 2014, for Cohort 3. For example, the first intervention quarter for Cohort 1 (I1) runs from January 1, 2013, to March 31, 2013. In each period (baseline or intervention), the treatment group each quarter includes all beneficiaries assigned to a treatment practice by the start of the quarter and who met other sample criteria—that is, they were enrolled in FFS Medicare and were living in New York. In each period, the comparison group includes all beneficiaries assigned to a comparison practice by the start of the quarter and who met the other sample criteria.

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

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

B = baseline; C = comparison; FFS = fee-for-service; I = intervention; IVD = ischemic vascular disease; T = treatment.

Table IV.4. Unadjusted mean outcomes (quality-of-care outcomes) measured for Medicare FFS beneficiaries, by treatment status and quarter

Inpatient admissions for ACSCs (#/1,000 beneficiaries/quarter)						
	Number of Medicare FFS beneficiaries (practices)					
Q	T	C (no wgt)	C (wgt)	T	C	Diff (%)
Baseline period (January 1, 2012, to December 31, 2012, for Cohort 1 practices; July 1, 2012, to June 30, 2013, for Cohort 2 practices; and July 1, 2013, to June 30, 2014, for Cohort 3 practices)						
B1	18,608 (55)	50,877 (158)	18,471	15.5	11.8	3.7 (31.0%)
B2	19,249 (55)	52,858 (158)	19,111	15.3	15.3	0.1 (0.4%)
B3	19,301 (55)	53,526 (158)	19,393	14.5	13.7	0.8 (6.2%)
B4	19,770 (55)	54,934 (158)	19,954	16.6	13.6	3.0 (22.1%)
Intervention period (January 1, 2013, to June 30, 2016, for Cohort 1 practices; July 1, 2013, to June 30, 2016, for Cohort 2 practices; and July 1, 2014, to June 30, 2016, for Cohort 3 practices)						
1	18,422 (55)	50,489 (158)	18,598	16.3	13.5	2.8 (21.0%)
I2	19,078 (55)	52,520 (158)	19,447	14.4	15.3	-0.8 (-5.5%)
I3	19,362 (55)	53,249 (158)	19,845	14.3	13.8	0.5 (3.3%)
I4	19,789 (55)	54,668 (158)	20,479	15.7	13.7	2.0 (14.6%)
I5	20,014 (55)	55,526 (158)	20,775	16.4	11.8	4.7 (39.6%)
I6	14,830 (37)	38,759 (108)	15,376	15.6	13.5	2.1 (15.4%)
I7	15,022 (37)	39,058 (108)	15,497	15.1	13.4	1.7 (12.9%)
I8	15,341 (37)	39,829 (108)	15,777	14.1	11.5	2.5 (21.9%)
I9	15,648 (37)	40,182 (108)	15,929	15.0	12.4	2.6 (21.4%)
I10	9,234 (16)	19,893 (51)	9,376	16.7	14.8	1.9 (12.8%)
I11	9,391 (16)	20,003 (51)	9,422	14.9	13.8	1.1 (7.7%)

30-day unplanned hospital readmissions (#/1,000 beneficiaries/quarter)						
	Number of Medicare FFS beneficiaries (practices)					
Q	T	C (no wgt)	C (wgt)	T	C	Diff (%)
Baseline period (January 1, 2012, to December 31, 2012, for Cohort 1 practices and July 1, 2012, to June 30, 2013, for Cohort 2 practices)						
B1	14,096 (37)	35,602 (108)	13,927	12.0	11.2	0.8 (7.2%)
B2	14,502 (37)	36,965 (108)	14,397	13.6	11.1	2.5 (22.3%)
B3	14,547 (37)	37,532 (108)	14,664	14.2	9.7	4.4 (45.6%)
B4	14,860 (37)	38,345 (108)	15,018	12.9	12.2	0.7 (5.5%)
Intervention period (January 1, 2013, to June 30, 2016, for Cohort 1 practices and July 1, 2013, to June 30, 2016, for Cohort 2 practices)						
I1	13,689 (37)	35,275 (108)	14,155	10.7	10.5	0.3 (2.7%)
I2	14,131 (37)	36,503 (108)	14,642	12.0	11.6	0.3 (2.9%)
I3	14,278 (37)	36,867 (108)	14,781	12.9	10.3	2.5 (24.6%)
I4	14,556 (37)	37,691 (108)	15,115	13.0	11.6	1.4 (12.2%)
I5	14,617 (37)	38,050 (108)	15,149	13.5	11.9	1.6 (13.1%)
I6	14,830 (37)	38,759 (108)	15,376	13.1	9.6	3.5 (36.0%)
I7	15,022 (37)	39,058 (108)	15,497	14.6	11.4	3.1 (27.5%)
I8	15,341 (37)	39,829 (108)	15,777	12.1	10.3	1.8 (17.3%)
I9	9,044 (16)	19,644 (51)	9,222	17.7	12.9	4.8 (36.9%)
I10	9,234 (16)	19,893 (51)	9,376	15.2	15.0	0.2 (1.1%)
I11	-[a]	-[a]	-[a]	-[a]	-[a]	-[a]

Table IV.4 *(continued)*

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

Notes: The baseline quarters are measured relative to the start of the baseline period on January 1, 2012, for Cohort 1; July 1, 2012, for Cohort 2; and July 1, 2013, for Cohort 3. For example, the first baseline quarter (B1) for Cohort 1 runs from January 1, 2012, to March 31, 2012. The intervention quarters are measured relative to the start of the intervention period on January 1, 2013, for Cohort 1; July 1, 2013, for Cohort 2; and July 1, 2014, for Cohort 3. For example, the first intervention quarter for Cohort 1 (I1) runs from January 1, 2013, to March 31, 2013. In each period (baseline or intervention), the treatment group each quarter includes all beneficiaries assigned to a treatment practice by the start of the quarter and who met other sample criteria—that is, they were enrolled in FFS Medicare and were living in New York. In each period, the comparison group includes all beneficiaries assigned to a comparison practice by the start of the quarter and who met the other sample criteria.

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

[a] No data are available for 30-day unplanned hospital readmissions for I11, the switch from ICD-9 to ICD-10 codes prevented us from calculating this outcome measure for I11 the same way as in previous quarters.

ACSC = ambulatory care-sensitive condition; B = baseline; C = comparison; Diff = difference; FFS = fee-for-service; I = intervention; ICD-9 = *International Classification of Diseases,* 9th edition; no wgt = unweighted; Q = quarter; T = treatment; wgt = weighted.

Table IV.5. Unadjusted mean outcomes (quality-of-care outcomes) measured for high-risk Medicare FFS beneficiaries, by treatment status and quarter

	Inpatient admissions for ACSCs (#/1,000 beneficiaries/quarter)					
	Medicare FFS beneficiaries (practices)					
Q	T	C (no wgt)	C (wgt)	T	C	Diff (%)
Baseline period (January 1, 2012, to December 31, 2012, for Cohort 1 practices; July 1, 2012, to June 30, 2013, for Cohort 2 practices; and July 1, 2013, to June 30, 2014, for Cohort 3 practices)						
B1	4,785 (55)	13,367 (158)	4,712	39.5	35.7	3.8 (10.6%)
B2	4,760 (55)	13,365 (158)	4,716	45.0	43.0	2.0 (4.6%)
B3	4,574 (55)	13,101 (158)	4,612	35.0	36.9	-2.0 (-5.3%)
B4	4,487 (55)	12,942 (158)	4,566	46.8	41.2	5.6 (13.5%)
Intervention period (January 1, 2013, to June 30, 2016, for Cohort 1 practices; July 1, 2012, to June 30, 2016, for Cohort 2 practices; and July 1, 2014, to June 30, 2016, for Cohort 3 practices)						
I1	4,782 (55)	13,268 (157)	4,754	45.0	37.8	7.2 (19.1%)
I2	4,725 (55)	13,249 (157)	4,764	36.6	44.0	-7.4 (-16.8%)
I3	4,579 (55)	12,986 (157)	4,707	35.6	38.1	-2.5 (-6.6%)
I4	4,469 (55)	12,784 (157)	4,646	42.3	37.6	4.7 (12.5%)
I5	4,363 (55)	12,534 (157)	4,546	43.3	32.8	10.5 (32.1%)
I6	3,138 (37)	8,427 (108)	3,265	39.5	38.1	1.4 (3.6%)
I7	3,035 (37)	8,245 (108)	3,197	42.2	40.1	2.1 (5.2%)
I8	2,947 (37)	8,109 (108)	3,154	36.0	32.3	3.7 (11.4%)
I9	2,851 (37)	7,892 (108)	3,070	42.8	34.4	8.4 (24.5%)
I10	1,575 (16)	3,666 (51)	1,743	44.4	48.5	-4.1 (-8.4%)
I11	1,536 (16)	3,549 (51)	1,686	39.1	45.9	-6.9 (-15.0%)

	30-day unplanned hospital readmissions (#/1,000 beneficiaries/quarter)					
	Medicare FFS beneficiaries (practices)					
Q	T	C (no wgt)	C (wgt)	T	C	Diff (%)
Intervention period (January 1, 2013, to June 30, 2016, for Cohort 1 practices and July 1, 2012, to June 30, 2016, for Cohort 2 practices)						
B1	3,585 (37)	9,230 (108)	3,520	36.3	34.4	1.9 (5.5%)
B2	3,548 (37)	9,217 (108)	3,523	36.1	32.8	3.3 (9.9%)
B3	3,410 (37)	9,036 (108)	3,443	37.2	27.1	10.1 (37.3%)
B4	3,338 (37)	8,899 (108)	3,395	31.2	37.2	-6.1 (-16.3%)
Intervention period (January 1, 2013, to June 30, 2016, for Cohort 1 practices and July 1, 2012, to June 30, 2016, for Cohort 2 practices)						
I1	3,551 (37)	9,251 (108)	3,630	30.1	27.1	3.1 (11.4%)
I2	3,503 (37)	9,185 (108)	3,591	31.4	31.2	0.2 (0.6%)
I3	3,402 (37)	8,973 (108)	3,514	34.4	28.5	5.9 (20.6%)
I4	3,328 (37)	8,800 (108)	3,456	34.0	28.8	5.1 (17.8%)
I5	3,223 (37)	8,594 (108)	3,350	39.1	30.4	8.7 (28.6%)
I6	3,138 (37)	8,427 (108)	3,265	32.8	22.9	9.9 (43.1%)
I7	3,035 (37)	8,245 (108)	3,197	38.6	34.4	4.1 (12.0%)
I8	2,947 (37)	8,109 (108)	3,154	32.9	25.4	7.5 (29.6%)
I9	1,626 (16)	3,770 (51)	1,784	56.6	35.3	21.3 (60.4%)
I10	1,575 (16)	3,666 (51)	1,743	39.4	40.1	-0.7 (-1.9%)
I11	-[a]	-[a]	-[a]	-[a]	-[a]	-[a]

Table IV.5 *(continued)*

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

Notes: The baseline quarters are measured relative to the start of the baseline period on January 1, 2012, for Cohort 1; July 1, 2012, for Cohort 2; and July 1, 2013, for Cohort 3. For example, the first baseline quarter (B1) for Cohort 1 runs from January 1, 2012, to March 31, 2012. The intervention quarters are measured relative to the start of the intervention period on January 1, 2013, for Cohort 1; July 1, 2013, for Cohort 2; and July 1, 2014, for Cohort 3. For example, the first intervention quarter (I1) for Cohort 1 runs from January 1, 2013, to March 31, 2013. In each period (baseline or intervention), the treatment group each quarter includes all beneficiaries assigned to a treatment practice by the start of the quarter and who met other sample criteria—that is, they were enrolled in FFS Medicare and were living in New York. In each period, the comparison group includes all beneficiaries assigned to a comparison practice by the start of the quarter and who met the other sample criteria.

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

[a] No data are available for 30-day unplanned hospital readmissions for I11, the switch from ICD-9 to ICD-10 codes prevented us from calculating this outcome measure for I11 the same way as in previous quarters.

ACSC = ambulatory care-sensitive condition; B = baseline; C = comparison; Diff = difference; FFS = fee-for-service; I = intervention; ICD-9 = *International Classification of Diseases,* 9th edition; no wgt = unweighted; Q = quarter; T = treatment; wgt = weighted.

Table IV.6. Unadjusted mean outcomes (service use and spending) measured for all Medicare FFS beneficiaries, by treatment status and quarter

	Number of Medicare FFS beneficiaries (practices)			All-cause inpatient admissions (#/1,000 beneficiaries/quarter)			Outpatient ED visit rate (#/1,000 beneficiaries/quarter)			Medicare Part A and B spending ($/beneficiary/month)		
Q	T	C (no wgt)	C (wgt)	T	C	Diff (%)	T	C	Diff (%)	T	C	Diff (%)
Baseline period (January 1, 2012, to December 31, 2012, for Cohort 1 practices; July 1, 2012, to June 30, 2013, for Cohort 2 practices; and July 1, 2013, to June 30, 2014, for Cohort 3 practices)												
B1	18,608 (55)	50,877 (158)	18,471	79.8	75.7	4.1 (5.4%)	153.2	141.7	11.5 (8.1%)	$701	$691	$9 (1.4%)
B2	19,249 (55)	52,858 (158)	19,111	83.7	78.5	5.1 (6.5%)	161.1	143.2	18.0 (12.6%)	$761	$746	$15 (2.0%)
B3	19,301 (55)	53,526 (158)	19,393	79.1	75.6	3.5 (4.6%)	152.5	143.9	8.6 (6.0%)	$746	$742	$4 (0.5%)
B4	19,770 (55)	54,934 (158)	19,954	87.2	78.1	9.1 (11.7%)	151.3	140.7	10.6 (7.5%)	$799	$768	$31 (4.1%)
Intervention period (January 1, 2013, to June 30, 2016, for Cohort 1 practices; July 1, 2013, to June 30, 2016, for Cohort 2 practices; and July 1, 2014, to June 30, 2016, for Cohort 3 practices)												
I1	18,422 (55)	50,489 (158)	18,598	79.0	75.3	3.8 (5.0%)	154.4	141.4	13.0 (9.2%)	$751	$733	$18 (2.5%)
I2	19,078 (55)	52,520 (158)	19,447	82.2	76.1	6.1 (8.0%)	162.9	150.1	12.8 (8.5%)	$790	$752	$38 (5.1%)
I3	19,362 (55)	53,249 (158)	19,845	83.3	76.1	7.1 (9.3%)	159.4	148.6	10.8 (7.3%)	$767	$752	$16 (2.1%)
I4	19,789 (55)	54,668 (158)	20,479	83.4	76.6	6.8 (8.9%)	168.5	148.6	19.9 (13.4%)	$821	$787	$34 (4.3%)
I5	20,014 (55)	55,526 (158)	20,775	84.6	74.7	9.9 (13.3%)	165.4	147.5	17.9 (12.2%)	$828	$765	$63 (8.2%)
I6	20,296 (55)	56,644 (158)	21,224	80.7	73.5	7.1 (9.7%)	166.7	154.3	12.4 (8.0%)	$815	$791	$24 (3.0%)
I7	20,356 (55)	56,809 (158)	21,404	87.3	78.4	9.0 (11.5%)	169.8	160.2	9.7 (6.0%)	$834	$814	$20 (2.5%)
I8	20,706 (55)	57,783 (158)	21,792	82.6	73.5	9.1 (12.3%)	169.8	156.4	13.3 (8.5%)	$839	$803	$37 (4.6%)
I9	15,648 (37)	40,182 (108)	15,928	85.2	72.1	13.1 (18.1%)	162.2	147.9	14.3 (9.7%)	$805	$786	$18 (2.3%)
I10	15,925 (37)	40,777 (108)	16,201	79.9	75.8	4.1 (5.5%)	174.8	156.4	18.5 (11.8%)	$825	$867	$-43 (-5.0%)
I11	15,940 (37)	40,707 (108)	16,203	79.9	74.7	5.1 (6.8%)	176.4	151.2	25.2 (16.7%)	$821	$825	$-4 (-0.5%)
I12	16,086 (37)	40,996 (108)	16,327	81.3	73.3	8.0 (10.9%)	163.9	145.0	19.0 (13.1%)	$841	$831	$10 (1.2%)
I13	9,286 (16)	19,872 (51)	9,413	88.6	81.8	6.8 (8.3%)	185.8	149.7	36.2 (24.2%)	$837	$874	$-37 (-4.2%)
I14	9,343 (16)	19,869 (51)	9,401	91.6	74.4	17.2 (23.1%)	185.3	157.0	28.4 (18.1%)	$905	$860	$45 (5.3%)

Table IV.6 *(continued)*

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

Notes: The baseline quarters are measured relative to the start of the baseline period on January 1, 2012, for Cohort 1; July 1, 2012, for Cohort 2; and July 1, 2013, for Cohort 3. For example, the first baseline quarter (B1) for Cohort 1 runs from January 1, 2012, to March 31, 2012. The intervention quarters are measured relative to the start of the intervention period on January 1, 2013, for Cohort 1; July 1, 2013, for Cohort 2; and July 1, 2014, for Cohort 3. For example, the first intervention quarter for Cohort 1 (I1) runs from January 1, 2013, to March 31, 2013. In each period (baseline or intervention), the treatment group each quarter includes all beneficiaries assigned to a treatment practice by the start of the quarter and who met other sample criteria—that is, they were enrolled in FFS Medicare and were living in New York. In each period, the comparison group includes all beneficiaries assigned to a comparison practice by the start of the quarter and who met the other sample criteria. See text for details.

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

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

Table IV.7. Unadjusted mean outcomes (service use and spending) measured for high-risk Medicare FFS beneficiaries, by treatment status and quarter

	Number of Medicare FFS beneficiaries (practices)			All-cause inpatient admissions (#/1,000 beneficiaries/quarter)			Outpatient ED visit rate (#/1,000 beneficiaries/quarter)			Medicare Part A and B spending ($/beneficiary/month)		
Q	T	C (no wgt)	C (wgt)	T	C	Diff (%)	T	C	Diff (%)	T	C	Diff (%)
Baseline period (January 1, 2012, to December 31, 2012, for Cohort 1 practices and July 1, 2012, to June 30, 2013, for Cohort 2 practices)												
B1	4,785 (55)	13,367 (158)	4,712	183.5	180.8	2.7 (1.5%)	247.3	224.0	23.3 (10.4%)	$1,511	$1,514	$-4 (-0.3%)
B2	4,760 (55)	13,365 (158)	4,716	192.9	184.0	8.9 (4.8%)	270.2	230.2	39.9 (17.3%)	$1,602	$1,539	$63 (4.1%)
B3	4,574 (55)	13,101 (158)	4,612	173.2	170.4	2.8 (1.6%)	245.5	236.5	9.0 (3.8%)	$1,560	$1,510	$49 (3.2%)
B4	4,487 (55)	12,942 (158)	4,566	180.1	169.8	10.3 (6.1%)	243.2	226.2	16.9 (7.5%)	$1,534	$1,508	$26 (1.7%)
Intervention period (January 1, 2013, to June 30, 2016, for Cohort 1 practices; July 1, 2012, to June 30, 2016, for Cohort 2 practices; and July 1, 2014, to June 30, 2016, for Cohort 3 practices)												
I1	4,782 (55)	13,268 (157)	4,754	184.4	168.7	15.7 (9.3%)	249.1	236.1	13.0 (5.5%)	$1,617	$1,628	$-11 (-0.7%)
I2	4,725 (55)	13,249 (157)	4,764	184.6	174.9	9.6 (5.5%)	277.1	252.1	25.0 (9.9%)	$1,619	$1,558	$61 (3.9%)
I3	4,579 (55)	12,986 (157)	4,707	178.9	169.3	9.6 (5.6%)	264.4	247.6	16.7 (6.8%)	$1,537	$1,511	$26 (1.7%)
I4	4,469 (55)	12,784 (157)	4,646	189.1	167.7	21.4 (12.7%)	288.0	249.4	38.6 (15.5%)	$1,650	$1,591	$59 (3.7%)
I5	4,363 (55)	12,534 (157)	4,546	188.9	171.0	17.8 (10.4%)	264.9	246.2	18.8 (7.6%)	$1,715	$1,598	$117 (7.3%)
I6	4,261 (55)	12,303 (157)	4,472	175.3	157.5	17.8 (11.3%)	267.1	253.5	13.5 (5.3%)	$1,604	$1,536	$68 (4.4%)
I7	4,109 (55)	11,967 (157)	4,362	191.3	181.1	10.2 (5.6%)	271.2	279.4	-8.2 (-2.9%)	$1,646	$1,649	$-3 (-0.2%)
I8	3,989 (55)	11,748 (157)	4,300	178.7	163.9	14.8 (9.0%)	295.1	258.5	36.6 (14.1%)	$1,691	$1,668	$22 (1.3%)
I9	2,851 (37)	7,892 (108)	3,070	197.1	151.0	46.1 (30.6%)	259.6	249.8	9.7 (3.9%)	$1,573	$1,540	$33 (2.2%)
I10	2,768 (37)	7,727 (108)	2,999	179.9	165.8	14.1 (8.5%)	290.3	260.7	29.6 (11.4%)	$1,630	$1,710	$-80 (-4.7%)
I11	2,684 (37)	7,454 (108)	2,896	181.1	167.3	13.8 (8.2%)	280.6	258.5	22.2 (8.6%)	$1,699	$1,642	$57 (3.5%)
I12	2,582 (37)	7,252 (108)	2,827	175.4	156.9	18.6 (11.8%)	290.3	248.3	42.0 (16.9%)	$1,620	$1,563	$57 (3.7%)
I13	1,413 (16)	3,301 (51)	1,592	204.5	184.6	20.0 (10.8%)	304.7	260.0	44.7 (17.2%)	$1,759	$1,794	$-36 (-2.0%)
I14	1,366 (16)	3,181 (51)	1,520	217.4	165.9	51.5 (31.1%)	352.1	264.6	87.5 (33.1%)	$1,950	$1,831	$119 (6.5%)

Table IV.7 *(continued)*

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

Notes: The baseline quarters are measured relative to the start of the baseline period on January 1, 2012, for Cohort 1; July 1, 2012, for Cohort 2; and July 1, 2013 for Cohort 3. For example, the first baseline quarter (B1) for Cohort 1 runs from January 1, 2012, to March 31, 2012. The intervention quarters are measured relative to the start of the intervention period on January 1, 2013, for Cohort 1; July 1, 2013, for Cohort 2; and July 1, 2014, for Cohort 3. For example, the first intervention quarter (I1) for Cohort 1 runs from January 1, 2013, to March 31, 2013. In each period (baseline or intervention), the treatment group each quarter includes all beneficiaries assigned to a treatment practice by the start of the quarter and who met other sample criteria—that is, they were enrolled in FFS Medicare and were living in New York. In each period, the comparison group includes all beneficiaries assigned to a comparison practice by the start of the quarter and who met the other sample criteria. See text for details.

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

2. Mean outcomes for the treatment and comparison groups

Quality-of-care processes. Means in this domain resemble those presented in the third annual report. For the HbA1c and lipid test measures, practices' performance declined from baseline to the second year of the intervention, with no obvious differences between treatment and comparison groups (Table IV.3). With respect to the 14-day follow-up measure, treatment and comparison practices were well matched at the start of the baseline period. However, treatment practices performed better on the follow-up measure in the last baseline quarter relative to the comparison group, and performed better than comparison practices throughout the intervention period, with the exception of only one quarter (I9).

Quality-of-care outcomes. Means in this domain are largely similar to those presented in the third annual report, with no distinguishable trends among, or differences between, treatment and comparison groups during the intervention period for the full sample or high-risk beneficiaries (Tables IV.4 and IV.5). There is one exception: 30-day readmissions among the high-risk group increased dramatically in I9 among treatment group practices (but not among comparison practices), before returning in I10 to a level on par with levels in the baseline and early implementation periods.

Service use. All-cause inpatient admissions among the full sample fluctuated from I1 to I12, with treatment beneficiaries having moderately higher rates than comparison beneficiaries (Table IV.6). Similarly, among high-risk beneficiaries, the treatment group had moderately higher rates than the comparison group. In I14, admissions increased again among Cohort 1 treatment practices, but decreased among Cohort 1 comparison practices. Among all three cohorts, outpatient ED visit rates among the full Medicare FFS sample generally increased during the intervention period, with no distinguishable differences between the treatment and comparison groups (Tables IV.6 and IV.7).

Spending. Mean Medicare Part A and B spending among all Medicare FFS beneficiaries generally increased for the treatment and comparison groups from I1 to I8 (all three cohorts), from I9 to I12 (Cohorts 1 and 2), and from I13 to I14 (Cohort 1) (Table IV.6). However, there were no distinguishable differences between treatment and comparison group spending, either for all or high-risk beneficiaries (Tables IV.6 and IV.7).

3. Primary tests results

Overview. The impact estimates in this report extend our results from the third annual report to include the final months of FLHSA's HCIA intervention covered by the no-cost extension. For the quality-of-care processes domain, we found statistically significant favorable differences between the treatment and comparison groups during the primary test period. For the quality-of-care outcomes, service use, and spending domains, we found no statistically significant or substantively important effects in either a favorable or an unfavorable direction. However, statistical power was limited for outcomes in the quality-of-care outcomes and spending domains. Table IV.8 summarizes all primary test results.

Table IV.8. Results of primary tests for FLHSA

Primary test definition					Statistical power[a] to detect an effect that is		Results			
Domain (# of tests in domain)	Outcome (units)	Time period for impacts	Population	Substantive threshold (expected direction of effect)[b]	Size of the substantive threshold	Twice the size of the substantive threshold[c]	Treatment group mean	Regression-adjusted difference between the treatment and estimated counterfactual (standard error)[b]	Percentage difference[d]	*p*-value[e]
Quality of care processes (4)	Received an HbA1c test (binary [yes or no]/beneficiary/year)	Intervention quarters 5–8[f] for Cohorts 1 and 2	Medicare FFS beneficiaries ages 18 to 75 with diabetes assigned to treatment practices	15.0% (+)	>99.9%	>99.9%	89.5%	0.8 (1.2)	1.0%	0.46
	Received a lipid profile(binary [yes or no]/beneficiary/year)	Intervention quarters 5–8[f] for Cohorts 1 and 2	Medicare FFS beneficiaries ages 18 to 75 with diabetes assigned to treatment practices	15.0% (+)	>99.9%	>99.9%	82.5%	2.3 (1.5)	2.8%	0.21
	Received complete lipid profile in the year (binary [yes or no]/beneficiary/year)	Intervention quarters 5–8[f] for Cohorts 1 and 2	Medicare FFS beneficiaries ages 18 or older with ischemic vascular disease assigned to treatment practices	15.0% (+)	>99.9%	>99.9%	75.7%	-0.8 (1.3)	-1.0%	0.53
	All inpatient admissions within a quarter were followed by an ambulatory care visit with a primary care or specialist provider within 14 days (binary [yes or no]/beneficiary/year)	Intervention quarters 5–10 for Cohort 1 and 5–8 for Cohort 2	Medicare FFS beneficiaries with at least one hospital stay in the quarter assigned to treatment practices	15.0% (+)	>99.9%	>99.9%	70.7%	3.0 (1.6)	4.4%	0.10
	Combined (%)	Varies by test	Varies by test	15.0% (+)	>99.9%	>99.9%	n.a.	n.a.	1.8%	0.05
Quality-of-care outcomes (4)	Inpatient admissions for ambulatory care-sensitive conditions (#/1,000 beneficiaries/ quarter)	Intervention quarters 5–11 for Cohort 1, 5–9 for Cohort 2, and 5 for cohort 3	All observable Medicare FFS beneficiaries assigned to treatment practices	5.0% (-)	28.3%	55.3%	15.4	-0.1 (1.1)	-0.4%	0.50

Table IV.8 *(continued)*

Primary test definition					Statistical power[a] to detect an effect that is		Results			
Domain (# of tests in domain)	Outcome (units)	Time period for impacts	Population	Substantive threshold (expected direction of effect)[b]	Size of the substantive threshold	Twice the size of the substantive threshold[c]	Treatment group mean	Regression-adjusted difference between the treatment and estimated counterfactual (standard error)[b]	Percentage difference[d]	*p*-value[e]
	30-day unplanned hospital readmissions (#/1,000/quarter)	Intervention quarters 5-10 for Cohort 1 and 5–8 for Cohort 2	All observable Medicare FFS beneficiaries assigned to treatment practices	5.0% (-)	22.3%	40.4%	14.3	0.1 (1.4)	0.7%	0.50
	Inpatient admissions for ambulatory care-sensitive conditions (#/1,000 beneficiaries/ quarter)	Intervention quarters 5–11 for Cohort 1, 5–9 for Cohort 2, and 5 for Cohort 3	All observable high-risk Medicare FFS beneficiaries assigned to treatment practices	15.0% (-)	62.4%	97.2%	41.0	-2.6 (4.1)	-5.9%	0.45
	30-day unplanned hospital readmissions (#/1,000/quarter)	Intervention quarters 5–10 for Cohort 1 and 5–8 for Cohort 2	All observable high-risk Medicare FFS beneficiaries assigned to treatment practices	15.0% (-)	41.1%	79.7%	39.9	4.7 (5.0)	13.4%	0.65
	Combined (%)	Varies by test	Varies by test	10.0% (-)	46.9%	87.0%	n.a.	n.a.	2.0%	0.59
Service use (4)	All-cause inpatient admissions (#/1,000 beneficiaries/ quarter)	Average over intervention quarters 5–14 for Cohort 1, 5–12 for Cohort 2, and 5–8 for Cohort 3	All observable Medicare FFS beneficiaries assigned to treatment practices	3.0% (-)	35.9%	71.3%	84.2	4.2 (2.6)	5.2%	0.84
	Outpatient ED visits (#/1,000 beneficiaries /quarter)	Average over intervention quarters 5–14 for Cohort 1, 5–12 for Cohort 2, and 5–8 for Cohort 3	All observable Medicare FFS beneficiaries assigned to treatment practices	5.0% (-)	70.6%	99.1%	172.0	3.3 (4.6)	1.9%	0.56
	All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	Average over intervention quarters 5–14 for Cohort 1, 5–12 for Cohort 2, and 5–8 for Cohort 3	All observable high-risk Medicare FFS beneficiaries assigned to treatment practices	5.0% (-)	38.5%	75.6%	189.0	15.1 (8.8)	8.7%	0.87

Table IV.8 *(continued)*

Primary test definition					Statistical power[a] to detect an effect that is		Results			
Domain (# of tests in domain)	Outcome (units)	Time period for impacts	Population	Substantive threshold (expected direction of effect)[b]	Size of the substantive threshold	Twice the size of the substantive threshold[c]	Treatment group mean	Regression-adjusted difference between the treatment and estimated counterfactual (standard error)[b]	Percentage difference[d]	*p*-value[e]
	Outpatient ED visits (#/1,000 beneficiaries/ quarter)	Average over intervention quarters 5–14 for Cohort 1, 5–12 for Cohort 2, and 5–8 for Cohort 3)	All observable high-risk Medicare FFS beneficiaries assigned to treatment practices	15.0% (-)	94.2%	>99.9%	287.6	-3.6 (15.3)	-1.2%	0.50
	Combined (%)	Average over intervention quarters 5–14 for Cohort 1, 5–12 for Cohort 2, and 5–8 for Cohort 3	Varies by test	7.0% (-)	84.1%	>99.9%	n.a.	n.a.	3.6%	0.88
Spending (2)	Medicare Part A and B spending ($/beneficiary/month)	Average over intervention quarters 5–14 for Cohort 1, 5–12 for Cohort 2, and 5–8 for Cohort 3	All observable Medicare FFS beneficiaries assigned to treatment practices	2.0% (-)	32.6	64.8	$835	$7 (19.9)	0.8%	0.55
	Medicare Part A and B spending ($/beneficiary/month)	Average over intervention quarters 5–14 for Cohort 1, 5–12 for Cohort 2, and 5–8 for Cohort 3	All observable high-risk Medicare FFS beneficiaries assigned to treatment practices	3.0% (-)	30.5	60.4	$1,689	$8 (65.3)	0.5%	0.51
	Combined (%)	Average over intervention quarters 5–14 for Cohort 1), 5–12 for Cohort 2, and 5–8 for Cohort 3)	Varies by test	2.5% (-)	33.8	67.2	n.a.	n.a.	0.7%	0.59

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

Note: The results for each outcome are based on a difference-in-differences regression model, as described in Appendix 2. Estimates are calculated for Medicare beneficiaries who are observable in the relevant time period: that is, beneficiaries who are enrolled in Medicare FFS (Part A and B), are alive, and have Medicare as their primary payer. Additional sample restrictions apply to the quality-of-care process measures; see text for details.

Table IV.8 *(continued)*

[a] The power calculation is based on actual standard errors from the analysis. For example, in the second-to-last row, a 3.0 percent effect on Medicare Part A and B spending (from the counterfactual of $1,689 + $8 = $1,697) would be a change of $68. Given the standard error of $65 from the regression model, we would be able to detect a statistically significant result 30.5 percent of the time if the impact was truly -$68, assuming a one-sided statistical test at the $p = 0.10$ significance level.

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

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

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

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

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

ED = emergency department; FFS = fee-for-service; FLHSA = Finger Lakes Health Systems Agency; HCIA = Health Care Innovation Award.

n.a. = not applicable.

Quality-of-care processes. The likelihood of receiving an HbA1c test or a lipid profile for diabetes was 1.0 and 2.8 percent higher, respectively, for the treatment group (a favorable estimate) than the estimated counterfactual. (Our estimate of the counterfactual—the outcome the treatment group members would have had in the absence of the HCIA-funded intervention—is the treatment group mean minus the difference-in-differences estimate.) We do not consider these favorable point estimates to be substantively large because both were smaller than the substantive threshold for these outcomes of 15 percent. In addition, these favorable results were not statistically significant.

The likelihood of receiving a lipid profile for IVD was 1.0 percent lower for the treatment group (an unfavorable estimate) than the estimated counterfactual. We cannot conclude whether these unfavorable results are statistically significant because our one-sided statistical tests assess only improvements in outcomes.

The likelihood of receiving an ambulatory care visit within 14 days of hospital discharge was 4.4 percent higher in the treatment group than its estimated counterfactual, a favorable and statistically significant difference.

The combined estimate across the four measures in the quality-of-care processes domain was 1.8 percent, a favorable and statistically significant point estimate. Although the estimates for ambulatory care visits—and for all quality-of-care process measures combined—were statistically significant, they were smaller than the substantive threshold of 15 percent for these outcomes.

Quality-of-care outcomes. The rate of ACSC admissions for the treatment group during the primary test period was 0.4 percent lower than our estimate of the counterfactual for the full Medicare FFS population, and 5.9 percent lower than our estimate of the counterfactual for high-risk Medicare FFS beneficiaries—indicating a decrease in ACSC admissions, or a favorable outcome. In contrast, the rate of unplanned readmissions was 0.7 percent higher for the full Medicare FFS population and 13.4 percent higher than our estimate of the counterfactual for high-risk Medicare FFS beneficiaries (indicating an increase in readmissions or an unfavorable outcome). However, no differences were substantively large for ACSC admissions or 30-day readmissions. (As shown in Table IV.1, the threshold was 5 percent for Medicare FFS beneficiaries and 15 percent for high-risk beneficiaries.) After combining results across the two outcomes (and among both populations) in this domain, the combined effect was 2.0 percent, smaller than the substantive threshold of 10.0 percent and in an unfavorable direction.

The statistical power to detect effects the size of the substantive threshold was poor to marginal for ACSC admissions (28.3 percent for the Medicare FFS population and 62.4 percent for high-risk Medicare FFS beneficiaries) and poor for 30-day unplanned readmissions (22.3 percent for the Medicare FFS population and 41.1 percent for high-risk Medicare FFS beneficiaries). Power was also poor (46.9 percent) for the combined effect in the domain.

Service use. The treatment group's admission rate was 5.2 percent higher for the full Medicare FFS population and 8.7 percent higher than our estimate of the counterfactual for high-risk Medicare FFS beneficiaries; the differences were unfavorable. The treatment group's

outpatient ED rate was 1.9 percent higher than our estimate of the counterfactual among the full Medicare FFS population and 1.2 percent lower among high-risk Medicare FFS beneficiaries; these differences were neither statistically significant nor substantively large. After combining results across the two outcomes in this domain, the outcomes for the treatment group were 3.6 percent higher than the estimated counterfactual. Power to detect effects that were the size of the substantive thresholds was poor for the admissions measure (35.9 and 38.5 percent for all patients and high-risk beneficiaries, respectively); marginal for the outpatient ED visit measure for all patients (70.6 percent); and good for the outpatient ED visit measure for high-risk beneficiaries (94.2 percent) and the combined outcome measure (84.1 percent).

Spending. For the full Medicare FFS population, the treatment group averaged $835 per beneficiary per month in Part A and B spending during the 5th through 10th intervention quarters, 0.8 percent (or $7) higher than the estimated counterfactual. Among high-risk Medicare FFS beneficiaries, spending was also similar between the treatment and comparison groups. Among both groups, treatment–comparison differences were smaller than the substantive thresholds of 2 and 3 percent for all and high-risk Medicare FFS beneficiaries, respectively. Statistical power to detect an effect the size of the substantive threshold was poor for the individual spending and combined outcomes (ranging from 30.5 to 33.8 percent).

4. Secondary test results

Estimates during the first intervention year (January 2013 to December 2014 for Cohort 1, July 2013 to June 2014 for Cohort 2, and July 2014 to June 2015 for Cohort 3). As shown in Table IV.9, most differences in service use and spending outcomes for the treatment group and its estimated counterfactual were small and not statistically significant during the first 12 months of the intervention (I1 through I4). These results support the credibility of the comparison group because we do not see large differences (favorable or unfavorable) in these outcomes during the first year of practice participation, a period during which we and FLHSA did not expect to see program effects in service use or spending. This increased confidence in the comparison group, in turn, gives us greater confidence in the primary test results and, eventually, the conclusions of the impact evaluation.

However, there were favorable (and statistically significant) differences in quality-of-care process measures for the treatment group and its estimated counterfactual during the first 12 months of the intervention—particularly with respect to inpatient admissions followed by an ambulatory care visit (increase of 8.0 percent) and beneficiaries with diabetes who received an HbA1c test during the year (increase of 2.0 percent) (Table IV.9). These differences in the first year are reasonable because we would anticipate that any program impacts on quality-of-care processes (as opposed to the other outcomes) could appear soon after the practices joined the intervention.

Estimates limiting the sample to prevent sample addition. We conducted additional secondary tests that limited the sample to those beneficiaries attributed at the start of the baseline or intervention periods. These tests used the same outcomes and time periods as the primary tests. The results of these secondary tests were consistent with the primary test results; they showed a statistically significant favorable difference between the treatment and comparison groups in hospitalizations followed by an ambulatory care visit within 14 days (and a statistically significant and favorable difference in the receipt of lipid profiles for beneficiaries with diabetes), but no differences in other outcomes in the quality-of-care processes domain, or any outcomes in the service use, quality-of-care outcomes, and spending domains for either the full Medicare FFS population or high-risk Medicare FFS beneficiaries (Table IV.9). These results support conclusions from the primary tests, and, because they are so similar to the primary test results, suggest that differential sample addition between treatment and comparison practice had no substantive effects on these results.

Estimates limiting the sample to the first two cohorts. We conducted additional secondary tests that limited the sample to Cohort 1 and 2 practices and their matched comparison practices to ensure that pronounced differential sample addition among treatment and comparison groups in Cohort 3 did not bias impact estimates. These tests used the same outcomes and time periods as the primary tests. The results of these secondary tests were fully consistent with the primary test results; they showed a statistically significant favorable difference between the treatment and comparison groups in hospitalizations followed by an ambulatory care visit within 14 days, but no differences in other outcomes in the quality-of-care processes domain or other domains (Table IV.9).

Table IV.9. Results of secondary tests for FLHSA

Secondary test definition				Results			
Domain (# of tests in domain)	Outcome (units)	Time period for impacts	Population	Treatment group mean	Regression-adjusted difference between the treatment and estimated counterfactual (standard error)	Percentage difference[a]	*p*-value[b]
Estimates during the first intervention year (January 1, 2013, to December 31, 2013, for Cohort 1; July 1, 2013, to June 30, 2014, for Cohort 2; and July 1, 2014, to June 30, 2015, for Cohort 3)							
Quality-of-care processes (4)	Received an HbA1c test (binary [yes or no]/beneficiary/year)	Intervention quarters 1–4 for Cohorts 1 and 2[c]	Medicare FFS beneficiaries ages 18 to 75 with diabetes assigned to treatment practices	91.1	1.8 (1.1)	2.0%	0.06
	Received a lipid profile(binary [yes or no]/beneficiary/year)	Intervention quarters 1–4 for Cohorts 1 and 2[c]	Medicare FFS beneficiaries ages 18 to 75 with diabetes assigned to treatment practices	84.4	1.2 (1.4)	1.4%	0.20
	Received complete lipid profile in the year (binary [yes or no]/beneficiary/year)	Intervention quarters 1–4 for Cohorts 1 and 2[c]	Medicare FFS beneficiaries ages 18 or older with ischemic vascular disease assigned to treatment practices	78.2	-0.2 (1.2)	-0.3%	0.57
	All inpatient admissions within a quarter were followed by an ambulatory care visit with a primary care or specialist provider within 14 days (binary [yes or no]/beneficiary/year)	Intervention quarters 1–4 for Cohorts 1 and 2[c]	Medicare FFS beneficiaries with at least one hospital stay in the quarter assigned to treatment practices	72.6	5.4 (1.6)	8.0%	<0.01
Service use (4)	All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	Intervention quarters 1–4 for Cohorts 1, 2, and 3	All observable Medicare FFS beneficiaries assigned to treatment practices	82.0	0.6 (2.7)	0.7%	0.58
	Outpatient ED visits (#/1,000 beneficiaries/quarter)	Intervention quarters 1–4 for Cohorts 1, 2, and 3	All observable Medicare FFS beneficiaries assigned to treatment practices	161.3	-1.1 (4.5)	-0.7%	0.40
	All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	Intervention quarters 1–4 for Cohorts 1, 2, and 3	All observable high-risk Medicare FFS beneficiaries assigned to treatment practices	184.2	9.1 (8.6)	5.2%	0.86
	Outpatient ED visits (#/1,000 beneficiaries/quarter)	Intervention quarters 1–4 for Cohorts 1, 2, and 3	All observable high-risk Medicare FFS beneficiaries assigned to treatment practices	269.6	-6.7 (13.0)	-2.4%	0.30
Spending (2)	Medicare Part A and B spending ($/beneficiary/month)	Intervention quarters 1–4 for Cohorts 1, 2, and 3	All observable Medicare FFS beneficiaries assigned to treatment practices	$782	$14 (20.6)	1.9%	0.76
	Medicare Part A and B spending ($/beneficiary/month)	Intervention quarters 1–4 for Cohorts 1, 2, and 3	All observable high-risk Medicare FFS beneficiaries assigned to treatment practices	$1,606	$33 (63.7)	2.1%	0.70

Table IV.9 *(continued)*

Secondary test definition				Results			
Domain (# of tests in domain)	Outcome (units)	Time period for impacts	Population	Treatment group mean	Regression-adjusted difference between the treatment and estimated counterfactual (standard error)	Percentage difference[a]	*p*-value[b]
Estimates limiting the sample to prevent sample addition after the first baseline or intervention quarter							
Quality of care processes (4)	Received an HbA1c test (binary [yes or no]/beneficiary/year)	Intervention quarters 5–8[c] for Cohorts 1 and 2	Medicare FFS beneficiaries ages 18 to 75 with diabetes assigned to treatment practices	90.1	0.8 (1.3)	0.9%	0.25
	Received a lipid profile(binary [yes or no]/beneficiary/year)	Intervention quarters 5–8[c] for Cohorts 1 and 2	Medicare FFS beneficiaries ages 18 to 75 with diabetes assigned to treatment practices	83.1	2.1 (1.6)	2.5%	0.10
	Received complete lipid profile in the year (binary [yes or no]/beneficiary/year)	Intervention quarters 5–10 for Cohort 1 and 5–8 for Cohort 2	Medicare FFS beneficiaries ages 18 or older with ischemic vascular disease assigned to treatment practices	76.2	-0.6 (1.3)	-0.7%	0.66
	All inpatient admissions within a quarter were followed by an ambulatory care visit with a primary care or specialist provider within 14 days (binary [yes or no]/beneficiary/year)	Intervention quarters 5–10 for Cohort 1 and 5–8 for Cohort 2	Medicare FFS beneficiaries with at least one hospital stay in the quarter assigned to treatment practices	72.2	3.9 (1.7)	5.7%	0.01
Quality-of-care outcomes (4)	Inpatient admissions for ambulatory care-sensitive conditions (#/1,000 beneficiaries/quarter)	Intervention quarters 5–11 for Cohort 1, 5–9 for Cohort 2, and 5 for Cohort 3	All observable Medicare FFS beneficiaries assigned to treatment practices	16.0	-0.8 (1.2)	-4.5%	0.27
	30-day unplanned hospital readmissions (#/1,000 beneficiaries/quarter)	Intervention quarters 5–10 for Cohort 1 and 5–8 for Cohort 2	All observable Medicare FFS beneficiaries assigned to treatment practices	14.8	0.3 (1.5)	2.3%	0.59
	Inpatient admissions for ambulatory care-sensitive conditions (#/1,000 beneficiaries/quarter)	Intervention quarters 5–11 for Cohort 1, 5–9 for Cohort 2, and 5 for Cohort 3	All observable high-risk Medicare FFS beneficiaries assigned to treatment practices	41.3	-3.3 (4.3)	-7.3%	0.23
	30-day unplanned hospital readmissions (#/1,000 beneficiaries/quarter)	Intervention quarters 5–10 for Cohort 1 and 5–8 for Cohort 2	All observable high-risk Medicare FFS beneficiaries assigned to treatment practices	40.3	4.7 (5.3)	13.3%	0.81

Table IV.9 *(continued)*

Secondary test definition				Results			
Domain (# of tests in domain)	Outcome (units)	Time period for impacts	Population	Treatment group mean	Regression-adjusted difference between the treatment and estimated counterfactual (standard error)	Percentage difference[a]	*p*-value[b]
Service use (4)	All-cause inpatient admissions (#/1,000 beneficiaries/ quarter)	Intervention quarters 5–14 for Cohort 1, 5–12 for Cohort 2, and 5–8 for Cohort 3	All observable Medicare FFS beneficiaries assigned to treatment practices	86.6	4.8 (2.9)	5.9%	0.95
	Outpatient ED visits (#/1,000 beneficiaries/quarter)	Intervention quarters 5–14 for Cohort 1, 5–12 for Cohort 2, and 5–8 for Cohort 3	All observable Medicare FFS beneficiaries assigned to treatment practices	172.3	3.1 (5.3)	1.8%	0.72
	All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	Intervention quarters 5–14 for Cohort 1, 5–12 for Cohort 2, and 5–8 for Cohort 3	All observable high-risk Medicare FFS beneficiaries assigned to treatment practices	187.8	15.5 (9.4)	9.0%	0.95
	Outpatient ED visits (#/1,000 beneficiaries/ quarter)	Intervention quarters 5–14 for Cohort 1, 5–12 for Cohort 2, and 5–8 for Cohort 3	All observable high-risk Medicare FFS beneficiaries assigned to treatment practices	281.5	1.5 (16.5)	0.6%	0.54
Spending (2)	Medicare Part A and B spending ($/beneficiary/month)	Intervention quarters 5–14 for Cohort 1, 5–12 for Cohort 2, and 5–8 for Cohort 3	All observable Medicare FFS beneficiaries assigned to treatment practices	$840	-$1 (22.4)	-0.2%	0.47
	Medicare Part A and B spending ($/beneficiary/month)	Intervention quarters 5–14 for Cohort 1, 5–12 for Cohort 2, and 5–8 for Cohort 3	All observable high-risk Medicare FFS beneficiaries assigned to treatment practices	$1,663	-$1 (70.5)	-0.1%	0.49
Estimates limiting the sample to Cohort 1 and 2 practices (excluding Cohort 3)							
Quality-of-care processes (4)	Received an HbA1c test (binary [yes or no]/beneficiary/year)	Intervention quarters 5–8[c] for Cohorts 1 and 2	Medicare FFS beneficiaries ages 18 to 75 with diabetes assigned to treatment practices	89.5	0.8 (1.2)	1.0%	0.46
	Received a lipid profile(binary [yes or no]/beneficiary/year)	Intervention quarters 5–8[c] for Cohorts 1 and 2	Medicare FFS beneficiaries ages 18 to 75 with diabetes assigned to treatment practices	82.5	2.3 (1.5)	2.8%	0.21
	Received complete lipid profile in the year (binary [yes or no]/beneficiary/year)	Intervention quarters 5–8[c] for Cohorts 1 and 2	Medicare FFS beneficiaries ages 18 or older with ischemic vascular disease assigned to treatment practices	75.7	-0.8 (1.2)	-1.0%	0.53

Table IV.9 *(continued)*

Secondary test definition				Results			
Domain (# of tests in domain)	Outcome (units)	Time period for impacts	Population	Treatment group mean	Regression-adjusted difference between the treatment and estimated counterfactual (standard error)	Percentage difference[a]	*p*-value[b]
	All inpatient admissions within a quarter were followed by an ambulatory care visit with a primary care or specialist provider within 14 days (binary [yes or no]/beneficiary/year)	Intervention quarters 5–10 for Cohort 1 and 5–8 for Cohort 2	Medicare FFS beneficiaries with at least one hospital stay in the quarter assigned to treatment practices	70.7	3.0 (1.6)	4.4%	0.10
Quality-of-care outcomes (4)	Inpatient admissions for ambulatory care-sensitive conditions (#/1,000 beneficiaries/ quarter)	Intervention quarters 5–11 for Cohort 1 and 5–9 for Cohort 2	All observable Medicare FFS beneficiaries assigned to treatment practices	15.5	-0.3 (1.2)	-1.8%	0.50
	30-day unplanned hospital readmissions (#/1,000 beneficiaries/quarter)	Intervention quarters 5–10 for Cohort 1 and 5–8 for Cohort 2	All observable Medicare FFS beneficiaries assigned to treatment practices	14.3	0.1 (1.3)	0.7%	0.50
	Inpatient admissions for ambulatory care-sensitive conditions (#/1,000 beneficiaries/ quarter)	Intervention quarters 5–11 for Cohort 1 and 5–9 for Cohort 2	All observable high-risk Medicare FFS beneficiaries assigned to treatment practices	41.1	-3.3 (4.4)	-7.5%	0.41
	30-day unplanned hospital readmissions (#/1,000 beneficiaries/quarter)	Intervention quarters 5–10 for Cohort 1 and 5–8 for Cohort 2	All observable high-risk Medicare FFS beneficiaries assigned to treatment practices	39.9	4.7 (4.8)	13.4%	0.66
Service use (4)	All-cause inpatient admissions (#/1,000 beneficiaries/ quarter)	Intervention quarters 5–14 for Cohort 1 and 5–12 for Cohort 2	All observable Medicare FFS beneficiaries assigned to treatment practices	83.6	3.0 (2.9)	3.7%	0.66
	Outpatient ED visits (#/1,000 beneficiaries/quarter)	Intervention quarters 5–14 for Cohort 1 and 5–12 for Cohort 2	All observable Medicare FFS beneficiaries assigned to treatment practices	170.4	3.9 (5.0)	2.4%	0.58
	All-cause inpatient admissions (#/1,000 beneficiaries/quarter)	Intervention quarters 5–14 for Cohort 1 and 5–12 for Cohort 2	All observable high-risk Medicare FFS beneficiaries assigned to treatment practices	186.7	14.8 (9.7)	8.6%	0.82
	Outpatient ED visits (#/1,000 beneficiaries/ quarter)	Intervention quarters 5–14 for Cohort 1 and 5–12 for Cohort 2	All observable high-risk Medicare FFS beneficiaries assigned to treatment practices	280.4	-8.2 (16.6)	-2.8%	0.48

Table IV.9 *(continued)*

Secondary test definition				Results			
Domain (# of tests in domain)	Outcome (units)	Time period for impacts	Population	Treatment group mean	Regression-adjusted difference between the treatment and estimated counterfactual (standard error)	Percentage difference[a]	*p*-value[b]
Spending (2)	Medicare Part A and B spending ($/beneficiary/month)	Intervention quarters 5–14 for Cohort 1, 5–12 for Cohort 2, and 5–8 for Cohort 3	All observable Medicare FFS beneficiaries assigned to treatment practices	$829	$5 (22.0)	0.6%	0.52
	Medicare Part A and B spending ($/beneficiary/month)	Intervention quarters 5–14 for Cohort 1, 5–12 for Cohort 2, and 5–8 for Cohort 3	All observable high-risk Medicare FFS beneficiaries attributed to treatment practices	$1,671	$5 (73.8)	0.3%	0.50

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

Notes: The results for each outcome are based on a difference-in-differences regression model, as described in Appendix 2. Estimates are calculated for Medicare beneficiaries who are observable in the relevant time period: that is, beneficiaries who are enrolled in Medicare FFS (Part A and B), are alive, and have Medicare as their primary payer. We defined high-risk beneficiaries as those with a Hierarchical Condition Category score in the top one-third among all treatment group members at the beginning of the baseline period (for outcomes in the baseline period) or intervention period (for outcomes in the intervention period).

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

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

[c] For the three quality-of-care process measures for diabetes and ischemic vascular disease, we calculated outcomes over a year-long period (rather than quarters).

ED = emergency department; FFS = fee-for-service; FLHSA = Finger Lakes Health Systems Agency.

5. Consistency of impact estimates with implementation findings

The impact estimates in the primary tests are plausible given the implementation findings. Notably, statistically significant (albeit modest) impacts on ambulatory care visits with a primary care or specialist provider within 14 days of hospitalization likely reflect care managers' efforts to follow up with patients after a hospitalization, coordinate patients' care among medical and community providers, and connect patients with community-based service organizations and transportation services for their medical appointments. However, as shown in Table IV.3, favorable differences between treatment and the comparison practices on the 14-day follow-up measure first emerged in the last baseline quarter and the first intervention quarter, when impacts might not be expected because practices had not yet hired all of the nurse care managers who would help schedule ambulatory care follow-up visits. Although the favorable differences in the last baseline quarter and the first intervention quarter might have been due to chance, the favorable differences might also signal that practices were engaged in other practice improvement efforts separate from the intervention. Therefore, it is possible that our finding of program improvements in quality-of-care processes could reflect, in part, practice improvement efforts outside the scope of the HCIA intervention

The primary test results showed no statistically significant favorable effects during the primary test period for quality-of-care outcomes, service use, and spending. (However, the evaluation was not well powered to detect effects on quality-of-care outcomes or spending.) These primary test results are plausible given the implementation evidence. The program was active during the original award period and the no-cost extension period. Even with a well-implemented intervention, however, it is possible that the program was unable to change beneficiaries' or providers' behaviors in ways that would affect outcomes in these domains during the primary test period covered in this report. In the case of FLHSA, it is possible that the program's large investments in care management and practice transformation helped generate modest favorable impacts in quality-of-care processes, but those investments did not translate into desired reductions in hospitalizations and ED visits during the primary test period.

6. Conclusions about program impacts, by domain

Based on all evidence currently available, we draw the following conclusions about program impacts. Table IV.10 summarizes these conclusions and their support.

- **The program had a statistically significant favorable effect on quality-of-care processes.** For the ambulatory care visit with a primary care or specialist provider within 14 days and the combined outcome in this domain, we found statistically significant favorable impacts. The point estimates from the primary tests suggest that the favorable impacts were modest in size (given that the estimates were smaller than the prespecified substantive thresholds). The secondary test results support these primary test results by (1) showing impacts in the first program year (when the intervention would presumably begin to register an effect on these quality-of-care processes) and (2) demonstrating that differential sample addition over time between the treatment and comparison groups did not drive results. However, because treatment practices' performance on the 14-day follow-up measure was better than the comparison group in the last baseline quarter (and the first intervention

quarter when impacts on 14-day follow-up might not yet be expected), it is possible that the favorable impacts on quality cannot be fully attributed to the HCIA-funded intervention. As such, these findings should be interpreted with some caution.

- **The program had no substantively large effect on service use.** The primary test results were not statistically significant for any outcome or population in this domain, and the combined test in the domain was neither substantively large nor statistically significant. We believe this lack of observed effects means the program truly did not have substantively large effects because the statistical power was good to detect effects the size of the substantive threshold for outpatient ED visits among high-risk beneficiaries and the combined outcome (more than 75 percent). The conclusion of no substantively large effects is also consistent with implementation findings because, although implemented reasonably well, it is plausible the program did not have intended effects in the service use domain.
- **The program had an indeterminate effect on quality-of-care outcomes and Medicare spending.** The primary test results were not statistically significant for any outcome or population in these domains, and the combined tests in each domain were neither substantively large nor statistically significant. However, the statistical power was poor to detect effects the size of the substantive threshold. As a result, null findings from the primary tests in these domains could be due to (1) the program truly not having a substantively large effect or (2) the program having a substantively large effect but our tests failing to detect it. The fact that we observed no declines in service use (which FLHSA anticipated would lead to reduced spending)—and that some primary tests for service use were well powered—suggests that for spending, at least, lack of effects is the more likely explanation.

Table IV.10. Conclusions about the impacts of FLHSA's HCIA program on patients' outcomes, by domain

		Evidence supporting conclusion		
Domain	Conclusion	Primary test result(s) that supported conclusion	Primary test result(s) plausible given secondary tests?	Primary test result(s) plausible given implementation evidence?
Quality-of-care processes	Statistically significant favorable effect	• Estimate for an ambulatory care visit with a primary care or specialist provider within 14 days was favorable and statistically significant (after adjusting for four tests in domain) • Estimate for the combined outcome in the quality-of-care processes domain was favorable and statistically significant	Yes	Yes
Quality-of-care outcomes	Indeterminate effect	• No substantively large or statistically significant effects; poor to marginal power to detect effects	Yes	Yes
Service use	No substantively large effect	• No statistically significant effects and the combined test for all outcomes in the domain was neither statistically significant nor substantively large; well-powered to detect a substantively large effect on ED visits for high-risk beneficiaries and the combined outcome in the domain	Yes	Yes
Spending	Indeterminate effect	• No substantively large or statistically significant effects; poor power to detect effects	Yes	Yes

Sources: Tables IV.8 and IV.9.

ED = emergency department; FLHSA = Finger Lakes Health Systems Agency; HCIA = Health Care Innovation Award.

V. DISCUSSION AND CONCLUSION

In this report, we update impact and implementation findings from the third annual report (Blair et al. 2017) to include the 12-month no-cost extension period. We found that FLHSA implemented the practice transformation and care management components of the intervention largely as planned during the original award period. In all practices, care managers provided high-risk patients with intensive care management services, and practice champions and care managers made efforts to transform their practices into PCMHs. However, 10 of the original 68 practices did not continue participating in the FLHSA intervention during the extension period, partly because the intervention discontinued care managers' salaries. In terms of impact results, we found that the program (1) had a favorable effect on quality-of-care processes, largely driven by an increase in ambulatory care visits with a primary care or specialist provider within 14 days; and (2) had no effect on service use (hospitalizations or outpatient ED visits). However, the

program had an indeterminate effect on quality-of-care outcomes (readmissions and ACSC admissions) and Medicare spending, due in part to limited statistical power to reliably detect effects.

Overall, these findings are fully consistent with those reported earlier in the third annual report (Blair et al. 2016). For the quality-of-care processes and outcomes domains, this consistency is expected because data constraints limited our ability to analyze additional quarters of data for most outcomes. For the service use domain, this consistency is important because it illustrates that, even after allowing additional time for impacts to accumulate, the program still did not measurably reduce service use.

Explaining indeterminate effects on spending

Our power to detect effects on Medicare Part A and B spending was poor. Specifically, if the program had succeeded in reducing costs by the value that we prespecified as substantively important (that is, 2 percent among the full population or 3 percent among a high-risk subgroup), our evaluation would have had only a 30 to 35 percent chance of detecting that impact (and thus a 65 to 70 percent chance of missing the impact). This low probability of detecting an effect, if there truly was one, reflects a simple fact: the variation in Medicare spending across practices and over time is large relative to the effect sizes that we would consider meaningful. In other words, although we might consider a 2 percent cost decrease among a general Medicare population to be substantively important, a true impact of that magnitude could easily be obscured in the data by random fluctuations in beneficiaries' costs. This differs from our impact estimates in the service use domain, in which we specified substantively important effects as those exceeding 5 to 15 percent (depending on the outcome and the population). The larger anticipated effect sizes greatly improved our chances of detecting an effect of that size amid statistical noise.

Explaining no substantively large effects on service use

The lack of effects on service use is not due to failure to implement the program or to engage providers as planned. In fact, available evidence indicates that FLHSA delivered a robust intervention during the 2.5-year award period and the 12-month extension period that followed. Rather, there are three likely explanations for the lack of observed impacts. The lack of effects might be a result of (1) unforeseen implementation barriers, including limited staff time to devote to transformation activities and care management; (2) overly ambitious goals given the relatively small portion of patients receiving intensive services; and (3) a nontrivial portion of practices that did not continue the intervention into the no-cost extension period. We discuss these three issues in more depth next.

First, although the program was generally implemented as planned, a few key implementation barriers might have limited the effectiveness of care management services in reducing utilization and costs. As noted in Section III.A, practice champions and care managers reported that they had limited time to devote to practice transformation and intensive care management activities, respectively. For practice champions, the HCIA-funded intervention was only part of their job, and they reported that weekly care team huddles, monthly care team

meetings, and learning collaboratives often took more time than they had available. Although care managers generally worked exclusively on HICA-funded activities, they often reported having insufficient time to manage large patient caseloads (ranging from 40 to 60 patients) and associated reporting requirements, combined with additional FLHSA commitments (such as attending learning collaboratives) and other meetings and reporting tasks. Care managers' high caseloads and reporting responsibilities could have negatively affected the quality or quantity of their interactions with patients, thus reducing the potential impact of care management services on patients' activation, self-management, access to care, and health outcomes.

Second, the intervention might have set overly ambitious goals at the outset. A premise of the intervention was that providing high-risk patients with intensive care management services would substantially reduce readmissions, potentially preventable hospitalizations, and avoidable ED visits, on the order of 15 to 25 percent. FLHSA gave direct care management services to some 17,500 people among 750,000 total patients, or 2.3 percent of all patients at participating practices. If this relatively small share of the patient population was expected to drive substantial reductions in readmissions and potentially preventable hospitalizations, the reductions would have to be substantially, potentially unrealistically large.

Third, the loss of 10 of the intervention's 68 participating practices (14.7 percent of the study sample) could contribute to the lack of observed impacts. Following the intent-to-treat principle, we kept all practices—including the 10 that dropped out of the intervention—in our treatment group throughout the intervention period. Impacts might have been larger had those 10 practices remained in the intervention. However, as noted earlier, keeping all practices in our treatment group enables us to assess the effect of the FLHSA intervention under real-world circumstances, in which practice participation is not 100 percent during the entire award and no-cost extension periods.

REFERENCES

Blair, Randall, Rachel Shapiro, Rebecca Coughlin, Greg Peterson, Boyd Gilman, Laura Blue, Keith Kranker, Kate Stewart, Sheila Hoag, and Lorenzo Moreno. "Findings for Finger Lakes Health Systems Agency," pp. 187–268. In Greg Peterson, Laura Blue, Lorenzo Moreno, Keith Kranker, Boyd Gilman, Kate Stewart, Sandi Nelson, Kristin Geonnotti, Sheila Hoag, Andrew McGuirk, and Ken Peckham. "Evaluation of Health Care Innovation Awards (HCIA): Primary Care Redesign Programs. Third Annual Report." Princeton, NJ: Mathematica Policy Research, March 2017. Available at https://downloads.cms.gov/files/cmmi/hcia-primarycareredesign-thirdannualrpt.pdf. Accessed April 6, 2017.

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

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

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

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

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

Health Indicators Warehouse. "Average Age of Medicare Beneficiaries (mean)." Hyattsville, MD: National Center for Health Statistics, HIW, 2014a. Accessed November 19, 2014.

Health Indicators Warehouse. "Hospital Inpatient Medicare Admissions (per 1,000 beneficiaries)." Hyattsville, MD: National Center for Health Statistics, HIW, 2014b. Accessed August 13, 2014.

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

Medicare Payment Advisory Commission. "Data Book: Beneficiaries Dually Eligible for Medicare and Medicaid." Washington, DC: Medicaid and CHIP Payment and Access Commission, 2016. Available at https://www.macpac.gov/wp-content/uploads/2015/01/Dually-Eligible-Beneficiares-DataBook.pdf. Accessed September 15, 2016.

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

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

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

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

CHAPTER 3

RESEARCH INSTITUTE AT NATIONWIDE CHILDREN'S HOSPITAL/ AKRON CHILDREN'S HOSPITAL

Eric Lammers, Joe Zickafoose, Keith Kranker, Greg Peterson, Kate Stewart, Laura Blue, Brenda Natzke, Boyd Gilman, and Sheila Hoag

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

RESEARCH INSTITUTE AT NATIONWIDE CHILDREN'S HOSPITAL/AKRON CHILDREN'S HOSPITAL

CHAPTER SUMMARY

Introduction. The Research Institute at Nationwide Children's Hospital (NCH) received a three-year, $13.2 million Health Care Innovation Award (HCIA) in partnership with Akron Children's Hospital (ACH) and Partners for Kids (PFK) to improve care and health and lower costs for children living in and near Columbus and Akron, Ohio, and enrolled in Medicaid, especially Medicaid managed care (MMC).

Objectives. Although NCH used the HCIA to implement a range of interventions for Medicaid-enrolled children, this report focuses on one key component implemented at NCH's partner ACH: providing peer support to caregivers of children with acute behavioral health care episodes and care management after discharge. In this chapter, we (1) describe the design and implementation of this intervention; (2) assess impacts of the intervention on quality-of-care processes and outcomes and participants' service use; and (3) draw lessons from the evaluation for other hospitals, health systems, or payers considering implementing a similar intervention.

Methods. To assess program implementation, we reviewed the awardee's program documents, conducted site visits and interviews with ACH program leadership and staff, and surveyed staff. To estimate program impacts on Medicaid beneficiaries, we used a difference-in-differences design with a matched comparison group. Our impact estimates measured the regression-adjusted differences in post-discharge outcomes between the beneficiaries discharged from the ACH behavioral health inpatient unit during an intervention enrollment period (N = 540) and matched comparison beneficiaries discharged from nine comparison hospitals (N = 1,952), minus the differences in post-discharge outcomes between Medicaid beneficiaries discharged from ACH and matched comparison beneficiaries before the intervention began.

Program design and implementation. ACH undertook two intervention components to improve care for children with behavioral health care needs: (1) peer-to-peer support (called *parent partners*) for caregivers of children with acute behavioral health care needs and (2) post-hospitalization behavioral health care management. The parent partner component created a new, paid position aimed to improve the family-centeredness of care. ACH hired parent partners whose own children had significant behavioral health care needs and who, as a result, had extensive experience with the behavioral health care system for children. An outside consultant trained parent partners in approaches to family-centered care, effective interactions with behavioral health clinicians, and motivational interviewing. Parent partners met in person with children's caregivers within 24 hours of a child's hospitalization, continued to interact with caregivers at least daily during the hospitalization, and participated daily in behavioral health team rounds. ACH aimed to staff the intervention with five parent partners and have a parent partner available seven days a week. However, some staff turnover occurred. For most of the intervention period, there were from 3.0 to 4.5 parent partner full-time equivalents (FTEs), and in the final five months it dropped to 1.4 FTEs.

ACH also created a new role for a post-discharge behavioral health care manager. The behavioral health care manager was a licensed independent social worker, functioning as a case manager, who conducted follow-up calls with caregivers one to two days after discharge, a few days before and after the first scheduled outpatient follow-up visit, and as needed thereafter. The care manager sought to identify and address barriers to behavioral health follow-up care.

Impacts on patients' outcomes. The evidence indicates that the behavioral health intervention improved patients' outcomes in one of the three domains that we were able to assess: service use. Specifically, the impact estimates showed a 14.6 percent decrease (a favorable effect) in behavioral health emergency department (ED) visits. This result exceeded our prespecified threshold of substantive importance, although it was not statistically significant ($p = 0.20$). We found no evidence of program impacts in the other two evaluation domains: quality-of-care processes or quality-of-care outcomes. Because we had good power to detect a substantively large effect on the one quality-of-care process measure—outpatient behavioral health visits within 30 days of discharge—the null finding in this domain most likely means the program truly did not have its intended impact on this outcome. In the quality-of-care outcomes domain, however, our study had limited power to detect substantive effects on the one measure in the domain: behavioral health readmissions within 60 days of discharge. Thus, we were unable to determine whether the intervention truly had no impact on this measure, or whether it did but our tests failed to detect the impact.

Conclusions. Our findings suggest the intervention (parent peer support and behavioral health care management) likely decreased behavioral health ED visits. Although peer support interventions are increasing in prominence for adults, the NCH/ACH program included, to our knowledge, the first peer support intervention targeting caregivers of children with acute behavioral health care needs. Lessons for organizations considering implementing an intervention like ACH's include the following:

- Targeting patients discharged from an inpatient behavioral health unit can be effective in identifying a steady flow of patients with similar needs who are at high risk of behavioral health ED visits and readmissions, creating substantial opportunities to reduce these events.
- Making peer partners a hired position, and providing robust training in motivational interviewing and how to communicate best with caregivers and behavioral health providers, can aid substantially in implementing the intervention as intended.
- Programs might face challenges in hiring and retaining experienced peer partners over time, in addition to challenges in persuading hospital providers that peer partners add value. These challenges can be addressed, in part, by vetting candidates for peer partner positions well, allowing flexibility for peer partners to attend to their own children's behavioral health needs, building rapport between peer partners and hospital staff through frequent face-to-face communication (for example, participating in hospital rounds), and having senior leadership within the hospital champion the program and explain it to hospital providers.

Overall, the findings of favorable impacts of this innovative program on outpatient behavioral health ED visits are promising. Further testing would be helpful, particularly given the statistical uncertainty of the estimates.

Summary of intervention and impact results for Nationwide Children's Hospital

Intervention description			
Awardee description		Two children's hospital systems (NCH and ACH) and a pediatric Medicaid accountable care organization affiliated with NCH	
Award amount ($ millions)		$13.2	
Award extended beyond June 2015?		Yes (6 months)	
Location		Columbus and Akron, Ohio (urban, suburban, and rural areas)	
Target population for component in the impact evaluation		All child Medicaid beneficiaries discharged from ACH's behavioral health inpatient unit	
Intervention component included in the impact evaluation[a]		• Parent partner peer support to caregivers of children with acute behavioral health care episodes • Care management for children after acute behavioral health care episodes	
Metrics of intervention delivered		• Enrolled 1,078 child Medicaid beneficiaries during the award period (639 during the period of the impacts analysis, August 2013 through May 2015) • Staffed with 3.0 to 4.5 parent partner FTEs during 18 of 23 months of the intervention period (goal was 5.0 FTEs); dropped to 1.4 FTEs during the final 5 months	
Impact evaluation methods			
Core design		Difference-in-differences model with matched comparison group	
Treatment group	Definition	Child Medicaid beneficiaries with a hospital stay in ACH's 12-bed behavioral health inpatient unit	
	# of beneficiaries	540	
Comparison group definition		Matched child Medicaid beneficiaries discharged from behavioral health inpatient units at nine comparison hospitals in Ohio	
Impact results: Quality-of-care processes domain			
Any outpatient behavioral health follow-up visit within 30 days of discharge (percentage)		Comparison mean[b]	87.2%
		Impact estimate (% difference)	+2.7 pp (+3.3%)
Impact conclusion[c]		No substantively large effect	
Impact results: Quality-of-care outcomes domain			
Any readmission with a behavioral health primary diagnosis within 60 days of discharge (percentage)		Comparison mean[b]	10.6%
		Impact estimate (% difference)	+0.5 pp (+4.6%)
Impact conclusion[c]		Indeterminate effect	
Impact results: Service use domain			
Number of outpatient behavioral health ED visits within 90 days of discharge (#/1,000 beneficiaries)		Comparison mean[b]	135.2
		Impact estimate (% difference)	-27.8 visits (-14.6%)
Impact conclusion[c]		**Substantively important (but not statistically significant) favorable effect**	

Note: See the NCH/ACH chapter for details on the intervention, impact methods, and impact results.

[a] The program at NCH and ACH had three primary components that were not included in the impact evaluation: (1) improving care for children with complex health needs using multidisciplinary teams and standardized care processes, (2) promoting delivery of an evidence-based therapy (progesterone) to women at risk for repeat premature delivery by communicating with and providing resources to obstetrics and neonatal care providers, and (3) enrolling children eligible for Medicaid based on disability into an existing Medicaid ACO (PFK) and creating a new Medicaid ACO in northeastern Ohio led by ACH.

[b] The comparison mean is the estimate of the outcome the treatment group beneficiaries would have had if they had not received the intervention. It is equal to the mean for the treatment group in the post-intervention cohort minus the impact estimate.

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

ACH = Akron Children's Hospital; ACO = accountable care organization; ED = emergency department; FTE = full-time equivalent; NCH = Nationwide Children's Hospital; PFK = Partners for Kids; pp = percentage point.

I. INTRODUCTION

This report presents findings from the evaluation of the Research Institute at Nationwide Children's Hospital (NCH) Health Care Innovation Award (HCIA). NCH, a pediatric hospital in Columbus Ohio, used the award funding to implement a range of interventions to improve outcomes for Medicaid-enrolled children in Ohio. In the second annual report (Zickafoose et al. 2016), we described the design and implementation of these interventions in detail. This report adds to those earlier findings by estimating the impacts of the award on children's outcomes. In consultation with the Center for Medicare & Medicaid Innovation (CMMI), we chose to focus this report on one intervention—peer support and care management for caregivers of children hospitalized due to a behavioral health condition—as it was implemented at one NCH partner site: Akron Children's Hospital (ACH). We focus on this intervention because it was a novel approach to improving care for children with acute behavioral health needs, was implemented largely as planned, and because a rigorous analysis of the program's impacts was feasible.

Section II summarizes NCH's HCIA-funded program overall and the evaluation's design for estimating impacts of the ACH behavioral health intervention. Section III describes the design and implementation of the ACH behavioral health intervention. Section IV describes our methods, results, and conclusions of estimating program impacts on patients' outcomes in three domains: (1) quality-of-care processes, (2) quality-of-care outcomes, and (3) service use. Finally, Section V discusses aspects of the ACH behavioral health intervention that were novel and provides lessons for hospitals, health systems or payers considering implementing or supporting a similar intervention.

II. OVERVIEW OF THE AWARD AND THE IMPACT EVALUATION

A. NCH's HCIA-funded program

NCH received a three-year, $13.2 million HCIA in partnership with ACH and another organization called Partners for Kids (PFK). The goal of their HCIA-funded program was to improve care and health and lower costs for children enrolled in Medicaid, especially Medicaid managed care (MMC). Table II.1 summarizes key features of the program. Both NCH and ACH are freestanding pediatric tertiary care hospitals located in Columbus and Akron, Ohio, respectively. In addition to inpatient care, both offer outpatient primary, specialty, and emergency care services throughout their regions. PFK is a physician–hospital organization formed by a partnership between NCH and independent providers throughout a 34-county region in central and southeastern Ohio. PFK began operations in 1994 and has evolved into a Medicaid accountable care organization (ACO) that covered care for about 300,000 MMC children at the time of the HCIA award in July 2012. Although NCH was the HCIA awardee and administered the funding, both NCH and ACH used award funds to deliver program services. PFK served in an advisory role.

The awardee's activities can be summarized in four components, each with specific but interrelated goals.

1. Behavioral health

NCH aimed to improve care for children with behavioral health care needs served by NCH and ACH. To do this, NCH and ACH provided (1) peer support (through a position known as *parent partners*) for caregivers of children with acute behavioral health care needs—as determined either by admission to an inpatient or crisis-intervention unit or by a provider's referral—and (2) behavioral health care management. Parent partners were people whose own children had serious behavioral health care needs; under the intervention, the parent partners would meet with caregivers of hospitalized children to help families navigate the behavioral health care system and improve the family-centeredness of care. Following the child's discharge, behavioral health care managers conducted follow-up calls with caregivers (typically parents) to identify and address barriers to follow-up care. The goals for this program component included (1) increasing the rate of outpatient behavioral health follow-up within 30 days of discharge to at least 85 percent; (2) reducing behavioral health-related hospital readmissions within 60 days by 10 percent; and (3) reducing post-discharge impairment, as measured by the Columbia Impairment Scale, by at least 15 percent between discharge and 30 days after discharge.

2. Complex care

NCH sought to improve care management and hospital care for children with complex chronic conditions served by NCH and ACH, specifically those children with neurological conditions and feeding tubes. The HCIA-funded program included outpatient care management and redesign of inpatient care processes. The goals for this program component included (1) reducing hospital inpatient days for children with feeding tubes by 10 percent and (2) increasing the proportion of tube-fed children with healthy weights by 10 percent.

3. Preventing premature births

The awardee aimed to reduce the rates of preterm births and related neonatal hospital care in the county surrounding ACH (Summit County). ACH promoted the use of an evidence-based therapy (progesterone) to prevent repeat premature delivery by at-risk pregnant women. Obstetrics providers prescribed this therapy, but program staff also worked with insurers, pharmacies, and home health agencies to overcome barriers to delivering the therapy. The goals for this program component included (1) increasing progesterone use in pregnant mothers in Summit County with previous preterm births by 10 percent, (2) reducing the preterm birth rate in Summit County by 20 percent, and (3) reducing neonatal intensive care unit days at ACH by 10 percent.

4. ACO expansion

NCH sought to enroll children eligible for Medicaid based on disability into an existing Medicaid ACO (that is, PFK) and to create a new Medicaid ACO in northeastern Ohio led by ACH. This component sought to build infrastructure and connect children to the other components and other existing health care services; it did not include direct services to participants. The goals for this component included (1) reducing per-member, per-month MMC costs for children eligible for Medicaid due to disability by 2 percent; and (2) reducing per-

member, per-month MMC costs for children eligible for Medicaid for reasons other than disability by 1 percent.

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

Program description	
Award amount	$13,160,092
Award start date	July 2012
Implementation date	November 2012 (first program component) January 2013 (behavioral health component)
Award end date	Original: June 2015 End date for award funding of behavioral health component and other direct program services: June 2015 Administrative end date with no-cost extension: December 2015
Awardee description	NCH partnered with ACH and PFK. NCH and ACH are freestanding pediatric tertiary care hospitals located in Columbus and Akron, Ohio, respectively. In addition to inpatient care, both offer outpatient primary, specialty, and emergency care services throughout their regions. PFK is a physician–hospital organization formed by a partnership between NCH and independent providers throughout a 34-county region in central and southeastern Ohio and served as a Medicaid ACO that covered care for about 300,000 children in MMC at the time the HCIA began.
Program overview	NCH and its partners aimed to improve care for children with behavioral health care needs, improve care for children with complex chronic conditions, reduce the rates of preterm births and related neonatal hospital care in Summit County, Ohio, and implement activities to expand the existing PFK Medicaid ACO for children and create a new Medicaid ACO for children, led by ACH.
Program components	1. **Behavioral health.** Promote family-centered care by providing peer support to parents during and care management after acute behavioral health care episodes. 2. **Complex care.** Provide outpatient care management through a multidisciplinary team and work with inpatient clinical teams to standardize care processes. 3. **Preventing premature births.** Promote delivery of an evidence-based therapy (progesterone) to women at risk for repeat premature delivery by communicating with and providing resources to obstetrics and neonatal care providers. 4. **ACO expansion.** Enroll children eligible for Medicaid based on disability into an existing Medicaid ACO (PFK) and create a new Medicaid ACO in northeast Ohio led by ACH.
Target populations	The program broadly targeted children enrolled in Medicaid, especially MMC. The components of the program targeted the following specific populations: 1. **Behavioral health.** Children with high behavioral health care needs, as identified by admission to an inpatient or crisis-intervention unit, or a provider's referral. 2. **Complex care.** Children with complex care needs, initially defined as those with a neurological condition and a feeding tube and later expanded during the award period to include those with tracheostomies. 3. **Preventing premature births.** Women with a prior history of a premature delivery in Summit County, Ohio. 4. **ACO expansion**. Children eligible for Medicaid based on disability and enrolled in MMC in the 34-county PFK region and children enrolled in MMC in a 12-county region surrounding ACH.
Target impacts on patients' outcomes	1. **Behavioral health** • 85 percent rate of follow-up with behavioral health provider within 30 days of discharge • 10 percent reduction in behavioral health hospital readmissions within 60 days of discharge • 15 percent decrease between discharge and 30 days post-discharge on Columbia Impairment Scale survey among children who received care from an inpatient or crisis-intervention unit

Table II.1 *(continued)*

	2. **Complex care** • 10 percent reduction in hospital days among tube-fed children with a neurological diagnosis • 10 percent increase in the percentage of tube-fed children with a neurological diagnosis who have a healthy weight 3. **Preventing premature births** • 10 percent increase in progesterone use by pregnant women with prior preterm births • 20 percent reduction in preterm birth rate • 10 percent reduction in neonatal intensive care unit days 4. **ACO expansion** • 2 percent reduction in cost of care for children enrolled in Medicaid based on disability • 1 percent reduction in cost of care for other Medicaid-enrolled children
Workforce development	Created 45.2 new FTE positions funded by HCIA, including parent partners, behavioral health and complex care coordinators/care managers and their supervisors, and support staff. Each behavioral health site (NCH and ACH) planned to staff with five parent partners. An outside consultant trained parent partners in approaches to family-centered care, interacting effectively with behavioral health clinicians, and motivational interviewing. Other staff received informal, on-the-job training.
Location	Ohio: 46 counties (of 88 total counties in the state) in the northeastern, central, and southeastern regions of the state (urban, suburban, and rural areas)
Impact evaluation	
Intervention component included in impact evaluation	Behavioral health intervention at ACH only
Core design	Difference-in-differences with matched comparison group
Treatment group	Medicaid FFS and MMC beneficiaries younger than 18 years discharged from an inpatient behavioral health unit at ACH
Comparison group	MMC and Medicaid FFS beneficiaries younger than 18 discharged from 9 inpatient behavioral health units at similar hospitals in Ohio and matched to a beneficiary discharged from ACH
Extent to which the treatment group reflects the awardee's target population (for the component evaluated)	Moderate. Both ACH and NCH delivered the behavioral health component, but we could estimate impacts only for children discharged from ACH. Within the evaluation for ACH, we covered all relevant children because the awardee's target population and the impact evaluation's treatment group both consisted exclusively of pediatric Medicaid beneficiaries who received inpatient behavioral health services from ACH.
Study outcomes, by domain[a]	1. **Quality-of-care processes.** Any outpatient behavioral health follow-up visit within 30 days of discharge 2. **Quality-of-care outcomes.** Any readmission with a behavioral health primary diagnosis within 60 days of discharge 3. **Service use.** Number of outpatient behavioral health ED visits within 90 days of discharge

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

[a] We were unable to evaluate impacts on spending because Ohio Medicaid data did not include payment amounts for managed care encounters. Similarly, we were unable to estimate impacts on children's post-discharge impairment because Columbia Impairment Scale survey results were not available for the comparison group.

ACH = Akron Children's Hospital; ACO = accountable care organization; ED = emergency department; FFS = fee-for-service; FTE = full-time equivalent; HCIA = Health Care Innovation Award; MMC = Medicaid managed care; NCH = Nationwide Children's Hospital; PFK = Partners for Kids.

B. Overview of the impact evaluation

In consultation with CMMI, we chose to focus the impact evaluation on the behavioral health intervention at NCH's partner ACH because it was the only program component to meet two criteria. First, ACH implemented the behavioral health intervention largely as planned and on schedule. In contrast, the prematurity prevention and ACO development intervention were only partly implemented during the award period and so were unlikely to affect study outcomes (Zickafoose et al. 2016). Second, a rigorous impact evaluation for the ACH behavioral health intervention was feasible. This was not true either for the complex care component or for the behavioral health intervention implemented at NCH. Both of these interventions targeted children using strategies that could not be reliably replicated in available claims data; replication in clams is needed to identify a credible comparison group for evaluation.

We estimated program impacts for the behavioral health component at ACH on two of ACH's targeted outcomes: outpatient behavioral health follow-up visits within 30 days of hospital discharge and 60-day behavioral health hospital readmissions.[1] We also estimated impacts on outpatient emergency department (ED) visits for a behavioral health condition in the 90 days after discharge. Although ACH did not include reducing behavioral health outpatient ED visits as a goal, the peer supports and care management services could logically be expected to reduce these ED visits as well as behavioral health hospital readmissions.

We estimated the impacts of the ACH behavioral health intervention using a difference-in-differences design. We first identified Medicaid-enrolled children discharged from ACH's behavioral health unit during the award period (the treatment group) and matched them to beneficiaries discharged from behavioral health units from nine other hospitals in Ohio over the same period (the comparison group). We calculated differences in outpatient follow-up visits, behavioral health readmissions, and behavioral health outpatient ED visits between these two groups. Although differences between these two groups might reflect program impacts, they could also reflect long-standing differences in outcomes for those discharged from ACH versus the nine comparison hospitals that existed regardless of the intervention. To account for this possibility, we also calculated differences in outcomes for Medicaid-enrolled children discharged from ACH and the nine comparison hospitals before the intervention began. We subtracted any differences in outcomes we observed for this pre-intervention group from the differences observed in the post-intervention group. The resulting difference-in-differences estimates then captured our estimate of the program's impacts. We used regression models to adjust for children's characteristics, such as age or chronic conditions, at hospital discharge. These adjustments improved the statistical precision of the impact estimates and accounted for any differences between the treatment and comparison groups that existed despite matching beneficiaries on these characteristics.

[1] Due to limitations in data availability, we were unable to estimate impacts on ACH's third target outcome—impairment as measured by the Columbia Impairment Scale.

III. PROGRAM IMPLEMENTATION

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

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

A. Program design and adaptation

1. Target population and identifying, recruiting, and enrolling patients

Target population. ACH's behavioral health intervention targeted Medicaid-enrolled children younger than 18 hospitalized in the inpatient behavioral health unit at ACH.

Identifying participants. ACH's behavioral health parent partners identified potential participants through a daily review of admissions to the inpatient behavioral health unit.

Recruiting and enrolling participants. ACH's behavioral health parent partners recruited and enrolled participants by meeting in person with caregivers (parents or guardians) of eligible children within 24 hours of their admission to the behavioral health unit. Caregivers had the option of continuing to interact with the parent partner or opting out; caregivers also could request to meet with the parent partner at a later time, even if they initially opted out. Children whose caregivers engaged with the parent partners during hospitalization subsequently were targeted for care management services following discharge. Information about participating caregivers then passed to ACH's behavioral health care manager when the patient was discharged from the inpatient behavioral health unit. The care manager contacted caregivers by telephone following discharge. Consent for both components was informal, based on caregivers' agreement to interact with the parent partners and care manager.

2. Intervention

The ACH behavioral health intervention included two linked intervention components to improve care for children with behavioral health care needs: (1) peer-to-peer support via parent partners for caregivers of children with acute behavioral health care needs and (2) post-hospitalization behavioral health care management.

The peer-to-peer, parent partner intervention component created new, paid positions designed to improve the family-centeredness of care. ACH hired people whose own children had significant behavioral health care needs and who, as a result, had extensive experience with the pediatric behavioral health care system. ACH trained the parent partners to use their own

experiences and knowledge to empathize and effectively communicate with participants' caregivers with the goals of empowering caregivers to engage more actively in their children's care. The parent partners helped to identify important potential barriers to and facilitators of care that otherwise would not be disclosed to providers, and promoting family-centered care by providers. ACH recruited parent partners through contacts in its behavioral health department and through local behavioral health agencies and advocacy groups.

Parent partners met in person with children's caregivers within 24 hours of a child's admission, interacted with caregivers at least daily during the hospitalization, and participated daily in behavioral health team rounds. Parent partners' interactions with caregivers included active listening, sharing their own experiences, modeling positive interactions with providers, motivational interviewing to promote engagement with care and identifying unmet needs, and assistance with navigating the behavioral health care system.

During behavioral health team rounds, parent partners identified opportunities to promote family-centered communication and care planning between providers and caregivers. For example, caregivers might disclose to a parent partner that their reluctance to start a specific medication for a child was related to a prior negative experience of another family member with the same medication; the parent partner could then use this disclosure to encourage the behavioral health provider to address the concern through a more in-depth discussion of the risk of side effects with the medication and alternative choices. ACH aimed to have a parent partner available in person in the behavioral health unit seven days a week. Parent partners continued to follow up with caregivers by telephone as needed following discharge and scheduled calls one and two months after discharge to administer the Columbia Impairment Scale, which assesses a caregiver's ratings of global behavioral health impairment for the child.

ACH also hired a behavioral health care manager to implement the post-hospitalization care management component. This care manager, a licensed clinical social worker, conducted follow-up calls with caregivers one to two days after discharge, a few days before and after the first scheduled outpatient follow-up visit, and as needed thereafter. The care manager sought to identify and address barriers to behavioral health follow-up care (for example, lack of transportation or unstable housing) and connect caregivers to additional social resources as needed (for example, information on accessing the Medicaid transportation benefit and connection to housing programs).

Adaptations. ACH made three significant adaptations to the parent partner intervention component during the award:

1. *It focused on Medicaid-enrolled children only in the inpatient unit after starting with both the inpatient unit and the ED.*
   - The program initially targeted all children regardless of insurance type seen in the ED for a behavioral health concern or admitted to an inpatient behavioral health care unit. In the last quarter of 2013, the program focused on the inpatient unit and, in the first quarter of 2014, it began prioritizing work with children enrolled in Medicaid admitted to the inpatient unit. Awardee leaders made these changes because Medicaid was the primary

target population, staffing constraints made it challenging to include all children in both locations, and they felt the intervention component was likely to have the largest impact on children in the inpatient unit.

2. *ACH provided additional training in motivational interviewing for parent partners.*
   - The training in motivational interviewing occurred in spring 2013, shortly after initial implementation of the award and before the intervention period assessed in our impact evaluation. Intervention component leaders added this training early in the award after recognizing the skills fit well with the goals and roles of the parent partner.
3. *ACH adjusted staffing and personnel policies to meet the needs of parent partners.*
   - In the first half of 2014, ACH began adjusting staffing and personnel policies to provide more flexible use of leave and trading of work responsibilities among parent partners. This was in response to parent partners' need to deal with behavioral health crises among their own children. The awardee reported no significant adaptations to the behavioral health care management intervention.

3. Intervention staff and workforce development

ACH planned to hire five parent partners and one behavioral health care manager. Parent partners were laypeople (that is, nonclinicians) with at least one child with a significant behavioral health condition. The behavioral health care manager was a social worker with experience in the community behavioral health system. The program also provided funding for a full-time supervisor of these intervention staff, who was a licensed professional clinical counselor with extensive experience in providing behavioral health services.

The parent partners received extensive training from outside consultants, including (1) strategies for effective interactions with behavioral health clinicians, including orientation to the roles and hierarchies of hospital staff and participation in multidisciplinary care team discussions; (2) approaches to family-centered care, including building rapport by sharing experiences with their own children and eliciting caregivers' prior experiences with behavioral health care; and (3) motivational interviewing, an evidence-based method that seeks to engage individuals and motivate change by supporting intrinsic motivations. ACH reported that parent partners received 100 hours of motivational interviewing training, weekly supervision meetings, and weekly calls with the consultant. The parent partner supervisor and behavioral health care manager also attended the initial training sessions and weekly calls with the outside consultants. In addition to training its HCIA-funded staff, ACH provided brief informational sessions to its inpatient behavioral health providers, explaining the goals and roles of the parent partners and care manager.

B. Implementation effectiveness

In this section, we examine the evidence on implementation effectiveness—that is, we analyze measures of the intervention delivered and, when possible, compare those measures with the services the awardee intended to deliver. We assess the evidence on implementation effectiveness in five areas: (1) program enrollment, (2) service delivery, (3) staffing, and (4)

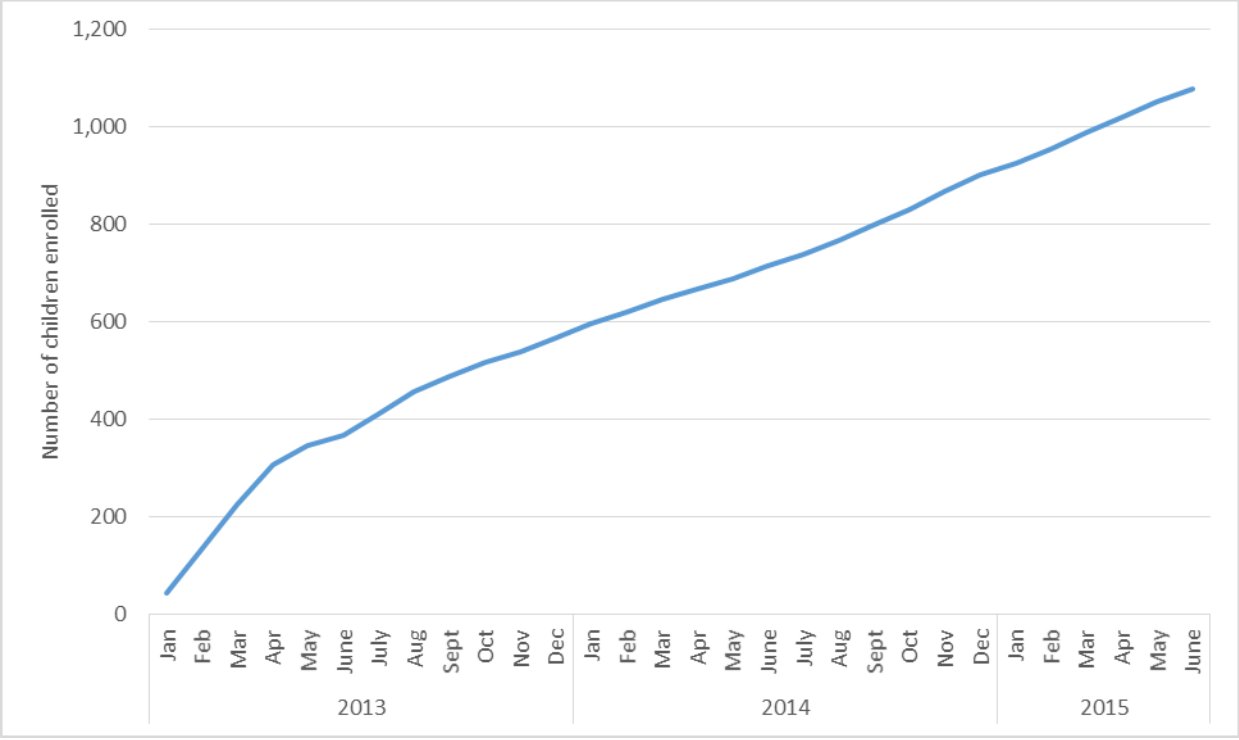
training. To conduct this analysis, we used data from interviews with program administrators and frontline staff, self-reported metrics included in ACH's self-monitoring and measurement reports to CMMI, and data from ACH on patients it enrolled in the interventions. We report metrics through June 2015, the end of the award-funded operating period for the interventions.

1. Program enrollment

ACH enrolled 1,079 Medicaid children (younger than 18 years old) in the parent partner and care manager interventions from January 2013 to June 2015 (Figure III.1). Cumulative enrollment increased throughout the duration of the program, but the rate of increase appeared fastest during the first few months of 2013. This is consistent with the fact that ACH initially enrolled participants through both the ED and inpatient unit before narrowing the focus to only the inpatient unit in mid-2013

ACH did not set specific enrollment targets for the interventions, but did report that more than 96 percent of caregivers of children eligible for the parent partner intervention elected to enroll from August through December 2013 (ACH did not report a participation measure after December 2013).

Figure III.1. Cumulative number of Medicaid children enrolled in the intervention, by month at ACH



Source: Mathematica's analysis of the participant list provided by ACH.
ACH = Akron Children's Hospital.

2. Service-related measures

ACH reported no quantitative measures of service delivery during hospitalization but reported two measures of services after discharge: successful telephone contacts with caregivers 30 and 60 days after discharge to administer the Columbia Impairment Scale. ACH set a goal to reach 40 percent of caregivers of enrolled children at each of these time points. The program reported being at or above this goal for each measure in about two-thirds of the months (15 of 23 months for 30-day follow-up and 16 of 23 months for 60-day follow-up).

Despite not having measures of service delivery in the hospital, ACH program administrators and staff reported in interviews that they believed they delivered the parent partner and care management intervention components successfully for a large majority of the enrolled population. Parent partners noted strong engagement in the intervention by caregivers whose children were enrolled. Program administrators also noted they received many positive write-in comments about parent partners in the standard experience-of-care survey that the hospital administered for all hospitalizations.

3. Staffing measures

ACH intended to staff the parent partner intervention with 5.0 FTEs, but reached this staffing goal only once early in the program (Figure III.2). ACH faced some staff turnover, especially early in the award, which program leaders attributed to hiring people who were unprepared for a position that required ongoing adaptation or who had exaggerated their experience with complex behavioral health needs. For most of the intervention period, there were 3.0 to 4.5 parent partner FTEs, except in the final five months when it dropped to 1.4 FTEs. Program administrators noted that some parent partners left near the end because they expected layoffs when the award funding ended. There was a lack of clarity about the level of support the organization would provide for the intervention after the award.

ACH planned to staff the care manager intervention with 1.0 FTE and maintained this staffing throughout the intervention period. ACH did not report significant changes in staffing for the parent partner supervisor position.

Figure III.2. ACH parent partner staffing

Source: Eighth and 12th quarterly reports to the Centers for Medicare & Medicaid Services.

4. HCIA-funded training

As noted previously, parent partners received extensive training from an outside consultant—both at the beginning of the intervention and then throughout the award, through weekly telephone calls—and received additional training in motivational interviewing. In site visit interviews, ACH parent partners described this training as a key to fulfilling their planned roles.

To assess additional perspectives on training in the program, we administered the HCIA Primary Care Redesign Trainee Survey to NCH and ACH program staff working in all program components from January to March 2015 (24 to 26 months after implementation began). We received responses from three parent partners at ACH. Due to the small number of respondents, we summarize their survey responses without presenting specific counts.

Most or all parent partner respondents rated their overall training as excellent and listed the training in motivational interviewing as most relevant to their roles. Similarly, most or all of the respondents strongly agreed with the following:

- The training clearly defined and met its objectives.
- Topics covered were relevant.

- Content was organized, easy to follow, and delivered at a comfortable pace that allowed for understanding.
- The training experience was useful to the trainees' work and helped to improve their job performance or complete job responsibilities.
- The trainer was knowledgeable about the training topics and well prepared.

Parent partner respondents reported spending time on workday activities in ways consistent with their training and the goals of the intervention. Respondents noted a heavy emphasis on direct communication (Table III.1).

Table III.1. Parent partner activities on a typical work day

On a typical work day, how much time do you spend on each of the following activities?				
None	**Less than 1/2 hour**	**1/2 to 1 hour**	**1 to 2 hours**	**More than 2 hours**
• Email communication with patients and families • Telephone conversations with physicians and other clinicians	• Email communications with physicians and other clinicians	• Following up on transitions of care • Telephone conversations with patients and families • Communication with social support services	• Attending team meetings or care conferences • Face-to-face communication with physicians and other clinicians	• Face-to-face communication with patients and families • Notes and documentation

C. Summary of facilitators of and barriers to implementation

Several factors facilitated implementation of ACH's HCIA-funded behavioral health care intervention, and other factors posed barriers to implementation. We described those factors in detail in the second annual report (Zickafoose et al. 2016), based on data collected during site visits, interviews, and review of awardee documents. Here, we summarize key facilitators and barriers most relevant to the ACH behavioral health intervention, along with any new information since the second annual report.

1. Facilitators of implementation

a. Intervention adaptability

ACH adapted the intervention to different settings and challenges. For example, the parent partner program was based on an outpatient support program at Columbia University for adults with serious mental illness, which ACH adapted for children in their care settings. ACH initially implemented its parent partner program in a hospital ED-based crisis intervention unit and then adapted the model for its own inpatient behavioral health care unit. This helped to more directly address the goal of reducing behavioral health-related readmissions. Another important adaption was allowing peer partners to trade shifts with one another so that they could attend to their own children's behavioral health care needs as they arose.

b. Making peer partners a paid position and providing robust training

Although hiring and retaining experienced peer partners was somewhat of a challenge (as described in Section III.B.3), making the position paid was critical for staffing the positions with qualified people over time. One of the key benefits of peer partners is that they have significant experience with the behavioral health system from the patient's perspective. However, training and supervisory support were essential for making the best use of that experience, translating it into effective strategies for working with patients and providers to improve care. Specifically, ACH invested significant resources into training peer partners in (1) motivational interviewing; (2) the roles and hierarchies of providers within the hospital and how best to communicate with these providers; and (3) approaches to family-centered care, including building rapport by sharing experiences with their own children and eliciting caregivers' prior experiences with behavioral health care. This training occurred through formal training sessions with external consultants and on the job through regular supervision.

2. Barriers to implementation and strategies for addressing them

a. Organizational culture

ACH had early challenges overcoming existing organizational culture to integrate the new interventions. For example, the parent partners and program administrators reported initial skepticism among some behavioral health clinicians about working with parent partners and the care manager. The parent partners and administrators noted that as clinicians worked more frequently with parent partners, they became more accepting and in some cases began to actively seek the opinions of parent partners as members of the care team. However, this challenge continued as the hospital hired new clinicians who were less familiar with the interventions.

b. Obtaining population-level data for self-monitoring

Although ACH obtained and used internal process and billing data to monitor performance, it faced major challenges in obtaining usable state Medicaid data to evaluate program effects on outcomes. For example, due to problems with Medicaid data, ACH was unable to use claims to measure rates of outpatient follow-up after behavioral health-related hospital discharges. To overcome this problem, in early 2013 it began surveying families of children with behavioral health-related admissions after discharge.

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

The available evidence from the awardee's reports and on-site interviews suggests that, over 2.5 years of HCIA funding, ACH largely implemented the behavioral health intervention as intended. Program enrollment increased steadily throughout the award period. In addition, trainee survey findings indicate parent partners undertook activities that would be expected if the intervention was implemented as planned, such as spending most of their time meeting with families and communicating with clinicians. ACH also reported meeting its goals for follow-up at 30 and 60 days post-discharge in about two-thirds of the award months. Although we do not have quantitative measures of some of the services delivered to individual patients—such as the

number of parent partner contacts with patients—program staff said that the parent partners and the care manager delivered services as intended to a large majority of the target population.

That being said, ACH also experienced several implementation challenges. Program staff and administrators noted that behavioral health providers were initially skeptical about working with the parent partners, but over time the providers came to value the contributions of these new staff. The program also faced early and late staffing challenges by not meeting the staffing goal (5.0 parent partner FTEs) for most of the intervention and dropping to only 1.5 FTEs for the final five months. Although enrollment continued at a steady rate during this time, the lower parent partner staffing could have resulted in decreased frequency and duration of interactions between parent partners and caregivers. In response to early staffing challenges, ACH adapted staffing policies to try to retain staff by adjusting work schedules to better accommodate parent partners' needs to care for their own children. ACH also offered more training to support parent partners. Staffing levels throughout the award period suggest this was largely successful, except for the final few months when staff left in anticipation of layoffs at the end of award funding.

IV. PROGRAM IMPACTS ON PATIENTS' OUTCOMES

A. Methods

This section summarizes our methods for estimating impacts of the ACH behavioral health intervention on outcomes for Medicaid-enrolled children. We provide more detail in the supplemental material at the end of this chapter.

1. Design

We estimated the impact of the ACH intervention using a difference-in-differences design. Specifically, we estimated impacts as:

- The difference in post-hospitalization outcomes between Medicaid children discharged from ACH (the treatment group) and a matched set of children discharged from nine comparison hospitals (the comparison group) *after* the intervention began (we call this this *post-intervention cohort*), minus
- The difference in post-hospitalization outcomes between intervention and matched comparison beneficiaries discharged *before* the intervention began (the *pre-intervention cohort*), and
- Adjusting the estimates to account for any baseline differences between the intervention and comparison beneficiaries that remained after matching

In other words, we assumed that any differences between the treatment and comparison groups before the intervention began would have remained the same during the award period, were it not for the intervention. Thus, we attributed any changes observed in the difference between treatment and comparison outcomes during the award period to the effects of the intervention. We used regression adjustment to account for changes that were actually due to small differences in observable characteristics—such as age—between the treatment and comparison beneficiaries.

2. Outcomes

We estimated impacts on three outcomes, which we grouped into three evaluation domains for consistency with the other evaluations in the HCIA Primary Care Redesign portfolio.

1. **Quality-of-care processes.** Whether a beneficiary received an outpatient behavioral health follow-up visit within 30 days of his or her qualifying discharge (where a *qualifying* discharge is the discharge from ACH following the behavioral health admission that made the beneficiary eligible for the intervention)

2. **Quality-of-care outcomes.** Whether a beneficiary had an unplanned readmission for a behavioral health condition within 60 days of the qualifying discharge

3. **Service use.** Number of behavioral health outpatient ED visits (defined as those not ending in a hospital stay)

We selected these outcomes for two reasons. First, they capture two of NCH and ACH's three target outcomes (shown in Table II.1): (1) to increase outpatient behavioral health follow-up visits following discharge and (2) to decrease the rate of behavioral health readmissions. Second, based on the program's design and its anticipated mechanisms to effect change (see supplemental material at the end of this chapter), we thought it reasonable to expect a reduction in behavioral health ED visits, even though this was not an explicit program goal. Appendix 1 describes the outcome construction in detail. We were unable to estimate impacts on ACH's third target outcome—impairment as captured by the Columbia Impairment Scale—because we did not have data on impairment scores for the comparison group members. We were unable to assess impacts on total Medicaid costs because of lack of cost data for the MMC beneficiaries.

3. Treatment group definition

We used an intent-to-treat approach to define the post-intervention treatment group, meaning that we included all beneficiaries who ACH intended to treat even if they did not receive services. Specifically, we defined the group as all Medicaid-enrolled children younger than 18 years who were discharged from ACH's behavioral health ward from August 1, 2013, to May 31, 2015 (Table IV.1). If a child had multiple stays at a behavioral health unit during this period, we selected the first stay. We limited the sample to beneficiaries who were enrolled in Medicaid for all 12 months before their hospital stay to ensure that we could reliably construct baseline characteristics (all defined using claims data) for matching treatment to similar comparison beneficiaries. These sample restrictions resulted in 540 beneficiaries in the post-intervention cohort of the treatment group. Of these, 71 percent actually enrolled in the ACH program, according to ACH records.

We defined the pre-intervention treatment group as all Medicaid-enrolled children younger than 18 years discharged from ACH's behavioral health ward from October 1, 2010, to December 31, 2012, applying the same sample restrictions as for the post-intervention treatment group. This resulted in 605 pre-intervention treatment beneficiaries.

Table IV.1. Calendar dates for enrollment (or pseudo-enrollment) and follow-up

Cohort	Enrollment period	Follow-up period
Pre-intervention[a]	October 1, 2010, to December 31, 2012 (27 months)	October 1, 2010, to March 31, 2013
Post-intervention[b]	August 1, 2013, to May 31, 2015 (22 months)	August 1, 2013, to August 31, 2015

[a] We set the pre-intervention enrollment period to 27 months so that it is of a similar length to the post-intervention enrollment period (22 months). We ended the pre-intervention enrollment period on December 31, 2012, so that everyone in the pre-intervention cohort could be followed for several months before the intervention was fully implemented (by August 2013).

[b] Although the intervention began in January 2013, it took about seven months to fully implement the program. Therefore, we started the initial enrollment date for the post-intervention cohort on August 1, 2013. We ended the post-intervention enrollment period on May 31, 2015, so that each person in the post-intervention treatment group was potentially exposed to the HCIA-funded intervention for at least one month before the funding ended for the behavioral health intervention on June 30, 2015. We ended the follow-up period on August 31, 2015, so that each person could be followed for three months after his or her qualifying discharge.

4. Comparison group definition

We developed two separate matched comparison groups, one for the post- and one for the pre-intervention treatment group. We constructed these groups in three steps.

1. We identified nine comparison hospitals that, like ACH, are located in Ohio and have a high volume of pediatric psychiatric inpatient visits (more than 500 cases during the intervention enrollment period). Table IV.2 lists the hospitals and their locations. We used these nine hospitals to develop both the pre- and post-intervention comparison groups.

2. For both periods (pre- and post-intervention), we defined a pool of potential comparison beneficiaries, identified as everyone who was enrolled in Medicaid, younger than 18, and discharged from an inpatient behavioral health facility at one of the comparison hospitals during the relevant time periods shown in Table IV.1.

3. We matched the treatment beneficiaries to potential comparison beneficiaries, exactly matching on some key characteristics (for example, we required each treatment beneficiary to match only to a comparison beneficiary or beneficiaries of the same sex and with the same primary diagnosis for the admission) and using propensity matching for others. Propensity matching helps to ensure that the treatment and comparison groups have similar characteristics on average, although any given treatment beneficiary might not be matched to a comparison beneficiary with the same or similar value.

Table IV.2 Comparison hospitals and their locations

Hospital	Location
Belmont Pines	Youngstown
Fairview Hospital (Cleveland Clinic)	Cleveland
Windsor Laurelwood Center for Behavioral Medicine	Willoughby
Kettering Medical Center	Kettering
OhioHealth (MedCentral)	Mansfield
Medical University of Ohio at Toledo	Toledo
Toledo Hospital	Toledo
University Hospitals Cleveland Medical Center	Cleveland
Upper Valley Medical Center	Troy

We matched at a ratio of roughly 3.5 comparison beneficiaries to every 1.0 treatment beneficiary in both the pre- and post-intervention cohorts. We did this to increase statistical power for our impact analysis relative to a 1:1 matching ratio.

5. Regression model and statistical tests

To implement the difference-in-differences design, we used a separate linear regression for each of the three outcomes. These regressions estimated the outcome as a function of treatment status (treatment or comparison), time period (pre- or post-intervention cohort), and beneficiary characteristics defined at the time of hospital discharge, such as age and risk scores. The regressions included indicator variables (fixed effects) for each hospital and for each beneficiary matched set. The difference-in-differences impact estimate was captured by the coefficient for the interaction of time period and treatment status. In each model, we gave every treatment beneficiary a weight of one and every comparison beneficiary a weight of one divided by the number of comparison beneficiaries in the matched set. For example, if a single treatment beneficiary was matched to four comparison beneficiaries, each of those comparison beneficiaries would receive a weight of 0.25 in the model. We used bootstrapped standard errors to account for correlation of outcomes among beneficiaries in each matched set, and the hospital-level fixed effects to account for correlation in outcomes among beneficiaries discharged from the same hospital. See Appendix 2 for details about the regression models, including covariate adjustment.

Because one of CMMI's goals for the HCIAs was to identify promising interventions, we conducted one-sided statistical tests (testing only for favorable impacts, not unfavorable ones) and we set our threshold for statistical significance at 0.10, which is less stringent than a more conventional 0.05.

6. Primary tests and robustness checks

Before conducting any analyses, we prespecified primary tests (Table IV.3). The purpose of these tests was to focus the evaluation on the outcomes, populations, and time periods for which we most strongly expected to detect impacts if the program was effective. In other words, the primary tests prespecified the level of evidence that we would need to observe before allowing ourselves to conclude that the program had impacts. Each primary test also specified a threshold for what we call *substantive importance*—that is, a magnitude of effect estimate that is large enough to be substantively interesting to CMMI and other stakeholders even if the results are not statistically significant. We did this because, as noted previously, a major goal of the HCIAs was to identify promising programs, not merely programs with proven impacts. In addition, poor statistical power (due to relatively small sample sizes) might make it difficult to identify statistically significant findings even if the program achieved its intended impacts. We gave both CMMI and the awardee an opportunity to review and comment on the primary tests before we conducted analysis.

In addition to conducting the primary tests, we also conducted a series of secondary, robustness checks (described in the supplemental material at the end of this chapter) and we assessed the plausibility of the primary test results in light of evidence from the implementation evaluation. To draw conclusions about program impacts, we required the primary test results to be consistent with secondary test results and the implementation evidence. Appendix 3 describes our decision rules for drawing conclusions.

Table IV.3. Specification of the primary tests for ACH's behavioral health intervention

Domain	Outcome (units)	Time period (days after hospital discharge)	Population	Substantive threshold (expected direction of the effect)[a, b]
Quality-of-care processes	Any outpatient behavioral health follow-up visit (percentage)	30 days	Medicaid children (ages birth to 17) discharged from ACH's behavioral health unit	7.5% (+)
Quality-of-care outcomes	Any readmission with a behavioral health primary diagnosis (percentage)	60 days	Medicaid children (ages birth to 17) discharged from ACH's behavioral health unit	6.0% (-)
Service use	Number of outpatient behavioral health ED visits (#/1,000 beneficiaries)	90 days	Medicaid children (ages birth to 17) discharged from ACH's behavioral health unit	7.5% (-)

[a] For the 60-day readmissions measure, we set the substantive threshold to 75 percent of ACH's anticipated impact—recognizing that a program might still be effective if the awardee did not fully realize its goal. ACH did not specify by how much it expected to increase 30-day outpatient behavioral health follow-up appointments or reduce behavioral health ED visits. Therefore, for these two outcomes, we based our substantive threshold on results reported in Smith (2014) and clinical expertise.

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

[c] The qualifying stay is the inpatient stay in the hospital's behavioral health ward that led to a beneficiary's assignment to the treatment group.

ACH = Akron Children's Hospital; ED = emergency department; HCIA =Health Care Innovation Award.

7. Data

We constructed all study variables using Medicaid data from the Ohio Colleges of Medicine Government Resource Center. This included information about Medicaid enrollment and health care services received during the pre- and post-intervention periods.

B. Results

1. Baseline characteristics

Post-intervention cohort

Table IV.4 shows characteristics of the post-intervention treatment group (second column). Of the 540 beneficiaries, the average age was 14. Most (72.4 percent) of the children were white, and 35.4 percent were male. Almost all (90.2 percent) were hospitalized for mood disorders (which includes depressive and bipolar disorders), and most (83.3 percent) had also been diagnosed as having attempted suicide or inflicted self-injury (either immediately before the hospital stay or up to two years before). Many of the children were diagnosed with other behavioral health conditions as well, including attention deficit or conduct disorders and adjustment disorders. Given these diagnoses, as well as diagnoses for physical conditions, the children in the post-intervention treatment group were predicted to have very high medical expenses—five times the national average for Medicaid-enrolled children (as indicated by a Chronic Illness & Disability Payment System [CDPS] score of 5.06, compared with a national average of 1.0). The children also had high rates of ED visits in the three months before the qualifying admission, both for behavioral health conditions (247.4 per 1,000 beneficiaries compared with a national average of 2.8) and for other conditions (548.1 versus a national average of 81.7). Almost all of the children (89.8 percent) were enrolled in MMC plans rather than FFS Medicaid.

Table IV.4 also shows that the post-intervention comparison group, which comprised 1,952 beneficiaries, was well matched to the treatment group (within 0.07 standardized differences) on all baseline characteristics except one. Specifically, the treatment group beneficiaries had a higher rate of nonbehavioral health hospital admissions in the three months before their qualifying admission than the comparison beneficiaries (122.2 per 1,000 beneficiaries for the treatment group versus 10.1 for the comparison group).

Pre-intervention cohort

The pre-intervention treatment group included 605 beneficiaries. Table IV.5 shows they were very similar to the post-intervention treatment beneficiaries on most baseline measures with three exceptions. The pre-intervention group was more likely to be male (47.8 versus 35.4 percent in the later period), less likely to be enrolled in MMC (74.5 versus 89.8 percent), and much less likely to be diagnosed as having attempted suicide or inflicted self-harm (33.9 versus 83.3 percent). The lower rates of MMC reflect the fact that Ohio Medicaid shifted more of its beneficiaries to managed care during the study period. The much lower rates of diagnoses for attempted suicide and self-harm are surprising but could reflect changes in billing practices over time, not true changes in underlying conditions.

The pre-intervention comparison group included 2,249 Medicaid children. As in the post-intervention cohort, the pre-intervention comparison group was well matched (within 0.09 standardized differences) to the treatment group on all measures except nonbehavioral health hospital admissions in the three months before enrollment (95.9 per 1,000 beneficiaries for the treatment group versus 28.5 admissions for the comparison group).

Table IV.4. Baseline characteristics of post-intervention cohort

Characteristic	Treatment group (n = 540)	Comparison group (n = 1,952)	Difference	Standardized difference[a]	National benchmark
Reason for hospitalization that qualified the person for the treatment or comparison group					
Adjustment disorders (%)	4.4	4.4	0	0	NA
Anxiety disorders (%)	2.6	2.6	0	0	
Mood disorders (%)	90.2	90.2	0	0	NA
Schizophrenia and other psychotic disorders (%)	1.7	1.7	0	0	NA
Demographic characteristics					
Age (years)	14.0	14.1	0.1	0.050	7.9
Male (%)	35.4	35.4	0	0	49.8
Race: white (%)	72.4	71.9	0.5	0.011	27.0
Ethnicity: Hispanic (%)	0.4	0.4	0.02	0.004	43.7
Medicaid-related characteristics					
Managed care (%)	89.8	89.7	0.07	0.002	86.5
Disabled (%)	10.3	10.9	0.6	0.021	3.4
Health status and chronic conditions					
Chronic Illness and Disability Payment System (CDPS) score	5.06	4.97	0.09	0.029	1.00
Behavioral health conditions measured in prior 24 months (%)					
Adjustment disorders (%)	36.5	35.9	0.6	0.012	NA
Anxiety disorders (%)	83.1	82.5	0.6	0.015	NA
Attention deficit, conduct, and disruptive behavior disorders (%)	73.9	70.8	3.1	0.066	NA
Disorders usually diagnosed in infancy, childhood, or adolescence	12.4	12.8	0.4	0.011	NA
Mood disorders (%)	98.7	98.4	0.3	0.030	NA
Personality disorders (%)	7.4	7.0	0.4	0.017	NA
Schizophrenia and other psychotic disorders (%)	7.8	7.8	0.02	0.001	NA
Alcohol-related disorders (%)	8.7	9.6	0.9	0.030	NA
Substance-related disorders (%)	23.0	23.2	0.23	0.001	NA
Suicide and intentional self-inflicted injury	83.3	84.8	1.5	0.043	NA
Service use in the prior 3 months (#/1,000 beneficiaries)					
Behavioral health hospitalizations	1,020	1,020	0.9	0.006	0.51
Nonbehavioral health hospitalizations	122.2	10.1	112.1	0.626	4.8
Behavioral health outpatient ED visits	257.4	259.9	2.5	0.005	2.8
Nonbehavioral outpatient health ED visits	548.1	546.7	1.5	0.002	81.7
Any outpatient visits for behavioral health in the prior 12 months (%)	86.5	85.6	0.9	0.026	NA

Sources: Analysis by Mathematica of Ohio Medicaid claims data provided by the Ohio Colleges of Medicine Government Resource Center. National benchmarks come from Medicaid and CHIP Payment and Access Commission (2011 and 2012), Healthcare Cost and Utilization Project (2014), Mahajan et al. (2009), Bardach et al. (2013), and the National Center for Health Statistics (2012).

Note: Characteristics are measured at the date of the inpatient discharge from ACH or comparison hospitals that led to a beneficiary's assignment to the treatment or comparison group (the beneficiary's enrollment or pseudo-enrollment date). We used behavioral health Clinical Classifications Software categories from the Healthcare Cost and Utilization Project to define the types of hospital stays and types of behavioral health conditions. The post-intervention cohort included beneficiaries whose enrollment or pseudo-enrollment dates were from August 1, 2013, to May 31, 2015.The comparison group means were weighted based on the number of matched comparisons per treatment beneficiary.

[a] Difference in means between the treatment and comparison groups divided by the standard deviation of the variable, which is pooled across the treatment and comparison groups.

ACH = Akron Children's Hospital; CHIP = Children's Health Insurance Program; ED = emergency department.

NA = not available.

Table IV.5. Baseline characteristics of the pre-intervention cohort

Characteristic	Treatment group (n = 605)	Comparison group (n = 2,249)	Difference	Standardized difference[a]	National benchmark
Reason for hospitalization					
Adjustment disorders (%)	1.5	1.5	0	0	NA
Anxiety disorders (%)	1.7	1.7	0	0	NA
Mood disorders (%)	90.4	90.4	0	0	NA
Schizophrenia and other psychotic disorders (%)	2.6	2.6	0	0	NA
Demographic characteristics					
Age (years)	13.4	13.4	0.05	0.017	7.9
Male (%)	47.8	47.8	0	0	49.8
Race: white (%)	66.6	67.0	0.4	0.009	27.0
Ethnicity: Hispanic (%)	0.3	0.4	0.07	0.011	43.7
Medicaid-related characteristics					
Managed care (%)	75.4	76.6	1.3	0.030	86.5
Disabled (%)	9.1	9.4	0.3	0.010	3.4
Health status and chronic conditions					
Chronic Illness and Disability Payment System (CDPS) score	5.10	5.09	0.010	0.004	1.00
Behavioral health conditions measured in prior 24 months (%)					
Adjustment disorders (%)	30.4	29.7	0.7	0.016	NA
Anxiety disorders (%)	62.8	61.2	1.6	0.032	NA
Attention deficit, conduct, and disruptive behavior disorders (%)	75.4	74.7	0.6	0.014	NA
Disorders usually diagnosed in infancy, childhood, or adolescence	14.9	14.0	0.9	0.025	NA
Mood disorders (%)	98.8	98.0	0.8	0.092	NA
Personality disorders (%)	4.0	3.5	0.5	0.025	NA
Schizophrenia and other psychotic disorders (%)	12.1	12.1	0.1	0.002	NA
Alcohol-related disorders (%)	6.1	6.4	0.3	0.011	NA
Substance-related disorders (%)	20.5	21.3	0.8	0.019	NA
Suicide and intentional self-inflicted injury	33.9	34.1	0.2	0.004	NA
Service use in the prior 3 months (#/1,000 beneficiaries)					
Behavioral health hospitalizations	1,015	1,008	6.4	0.054	0.51
Nonbehavioral health hospitalizations	95.9	28.5	67.4*	0.355	4.8
Behavioral health outpatient ED visits	209.9	209.0	0.9	0.002	2.8
Nonbehavioral outpatient health ED visits	522.3	494.4	27.9	0.034	81.7
Any outpatient visits for behavioral health in the prior 12 months (%)	90.1	90.2	0.1	0.004	NA

Sources: Analysis by Mathematica of Ohio Medicaid claims data provided by the Ohio Colleges of Medicine Government Resource Center. National benchmarks come from Medicaid and CHIP Payment and Access Commission (2011 and 2012), Healthcare Cost and Utilization Project (2014), Mahajan et al. (2009), Bardach et al. (2013), and the National Center for Health Statistics (2012).

Notes: The pre-intervention cohort included beneficiaries whose enrollment or pseudo-enrollment dates were October 1, 2010, to December 31, 2012.

[a] Difference in means between the treatment and comparison groups divided by the standard deviation of the variable, which is pooled across the treatment and comparison groups.

CHIP = Children's Health Insurance Program; ED = emergency department.

NA = not available.

2. Impact estimates

In this section, we present the results of the primary tests, by domain. We also describe what drives these primary test results by comparing the mean outcomes for the treatment and comparison groups in the pre- and post-intervention cohorts. Finally, we draw conclusions about program impacts, following decisions rules laid out in Appendix 3.

Quality-of-care processes (outpatient behavioral health follow-up visit within 30 days of discharge)

The difference-in-differences impact estimate suggests that the treatment group's rate of outpatient follow-up visits (at 84.44 percent), was slightly (3.3 percent, or 2.73 percentage points) higher than it would have been without the intervention (Table IV.6). However, this estimate was not statistically significant ($p = 0.15$) and was smaller than the substantive threshold of 7.5 percent. This small 3.3 percent impact estimate was driven by (1) the post-intervention treatment group's visit rate being slightly higher than comparison group's rate and (2) the pre-intervention treatment group's rate being slightly lower than the comparison group's rate (Table IV.7).

The statistical tests had good statistical power (86.1 percent) to detect impacts as large as the substantive threshold. Therefore, the primary tests indicate that the program likely did not have substantively large effects; if it had, our statistical tests most likely would have detected them. Further, the primary test results passed robustness checks described in the supplemental material at the end of this chapter, increasing confidence in these primary results. Therefore, we conclude that the program **had no substantively large effect on quality-of-care processes.**

Quality-of-care outcomes (60-day behavioral health inpatient readmissions)

The difference-in-differences impact estimate suggests that the treatment group's rate of behavioral health readmissions (at 10.19 percent) was slightly higher (by 4.6 percent, or 0.45 percentage points) than it would have been without the intervention (Table IV.6). However, this estimate was not statistically significant ($p = 0.58$) and was smaller than the substantive threshold of 6 percent. The mean behavioral readmission rates did decline meaningfully (by 16 percent) from the pre- to the post-intervention period for the treatment group (Table IV.7). However, because there was a similar decline for the comparison group, the difference-in-differences estimates showed no substantive impact.

The power to detect substantive impacts was poor, at only 15.3 percent. Therefore, it is possible that the program truly did have an important impact but our statistical tests failed to detect it. As a result, we conclude that the program had an **indeterminate effect on quality-of-care outcomes.**

Service use (outpatient behavioral health ED visits within 90 days of discharge)

The difference-in-differences impact estimate suggests that the treatment group's rate of behavioral health outpatient ED visits (163 per 1,000 beneficiaries) was 14.6 percent lower than it would have been without the intervention. This favorable estimate is not statistically significant ($p = 0.20$), but it is larger than the substantive threshold of 7.5 percent. Two factors

drove the favorable estimate. First, the treatment group rate fell by 12 percent from the pre- to post-intervention cohort. Second, the comparison group rate increased by 25 percent from the pre- to post-intervention cohort. Because the pre-intervention behavioral health ED visit rate was much higher for the treatment group than the comparison group (Table IV.7), these two factors brought the rates for the treatment and comparison groups close to each other in the post-intervention period, though they were still slightly higher for the treatment group (163.0 versus 150.8 visits).

Consistent with our goal to identify promising programs, not only those with definitive evidence for success, we conclude that the program **had a substantively important (but not statistically significant) favorable effect on service use.** We conclude this because (1) the impact estimate is favorable and larger than the substantive threshold and (2) the results are consistent with robustness checks described in the supplemental material to this chapter. However, this impact is estimated imprecisely and—if the program truly had no effect—there is a 20 percent chance ($p = 0.20$) that we would have observed a favorable impact at least as large as the one we did due to chance fluctuations in the data.

Table IV.6. Results of primary tests for ACH's behavioral health intervention

Primary test definition						Results			
Domain	Outcome (units)	Time period (days following qualifying stay)	Population	Substantive threshold (expected direction of the effect)[a]	Statistical power[b]	Treatment group mean	Difference-in-differences estimate, regression adjusted (standard error)	Percent difference	*p*-value
Quality-of-care processes	Any outpatient behavioral health follow-up visit within 30 days of discharge (percentage)	30 days	Medicaid children (ages birth to 17) discharged from ACH's behavioral health unit	7.5% (+)	86.1	84.44	2.73 (2.59)	3.3	0.15
Quality-of-care outcomes	Any readmission with a behavioral health primary diagnosis within 60 days of discharge (percentage)	60 days	Medicaid children (ages birth to 17) discharged from ACH's behavioral health unit	6.0% (-)	15.3	10.19	0.45 (2.26)	4.6	0.58
Service use	Number of outpatient behavioral health ED visits within 90 days of discharge (#/1,000 beneficiaries)	90 days	Medicaid children (ages birth to 17) discharged from ACH's behavioral health unit	7.5% (-)	19.6	163.0	-27.8 (33.7)	-14.6	0.20

Source: Analysis by Mathematica of Ohio Medicaid claims data provided by the Ohio Colleges of Medicine Government Resource Center.

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

[b] Probability of concluding that the program had a statistically significant favorable effect when the true effect was the size of the substantive threshold.

ACH = Akron Children's Hospital; ED = emergency department; HCIA = Health Care Innovation Award.

Table IV.7. Unadjusted mean outcomes, by cohort and treatment status

	Number of Medicaid beneficiaries		Outpatient behavioral health follow-up within 30 days of discharge (percentage of qualifying stays with follow-up[a])			Behavioral health readmission within 60 days of discharge (percentage of qualifying stays followed by readmission[a])			Number of behavioral health outpatient ED visits within 90 days of discharge (#/1,000 beneficiaries)		
Cohort	T	C	T	C	Diff. (%)	T	C	Diff. (%)	T	C	Diff. (%)
Pre-intervention	605	2,249	82.15	84.22	-2.07 (-2.5%)	12.07	12.70	-0.63 (-5.0%)	185.1	120.7	64.4 (53.4%)
Post-intervention	540	1,952	84.44	83.55	0.90 (1.1%)	10.19	11.19	-1.00 (-9.0%)	163.0	150.8	12.1 (8.0%)

Source: Analysis by Mathematica of Ohio Medicaid claims data provided by the Ohio Colleges of Medicine Government Resource Center.

Note: The means are weighted: each treatment beneficiary received a weight of 1; each comparison beneficiary received a weight equal to the reciprocal of the total number of comparison beneficiaries who matched to the same treatment beneficiary.

[a] The qualifying stay is the inpatient stay in the hospital's behavioral health unit that led to a beneficiary being assigned to the treatment or comparison group.

C = comparison group; Diff = difference; ED = emergency department; T = treatment group.

V. DISCUSSION AND CONCLUSIONS

NCH used HCIA funding to implement a range of interventions to improve outcomes for Medicaid-enrolled children in Ohio. We focused our impact evaluation on the behavioral health intervention at NCH's partner hospital, ACH, because a rigorous impact analysis was feasible and because ACH implemented the intervention largely as intended and on schedule. For other program components of the NCH/ACH award, either a rigorous impact design was not feasible, the intervention was implemented partially, or both.

1. Where the ACH model fits in the landscape of behavioral health interventions

The ACH intervention reflects a unique, family-centered approach to improving care for children hospitalized for behavioral health conditions (primarily depression, often accompanied by self-harm). To our knowledge, this is the first peer partner intervention focusing on children, in this case, through the children's caregivers, in an inpatient setting. Earlier peer partner interventions for behavioral health conditions have occurred in outpatient settings, focused on adults, or both (Obrochta et al. 2011; Robins et al. 2008; Chinman et al. 2014; Purington et al. 2016).

ACH hired the peer partners *to provide nonclinical support that might address an important gap in care*. Behavioral health conditions are often stigmatized and, when children are hospitalized, parents or other caregivers can feel overwhelmed or blame themselves in part for their children's condition. By sharing their own experiences, peer partners could normalize the situation and demonstrate that parents are not alone in their circumstances, and might help to give parents confidence to (1) share important but potentially sensitive information with health care professionals and (2) seek appropriate follow-up care. ACH targeted the peer partner intervention to caregivers of children in the hospital, thus reaching parents at a critical time when, potentially, they needed the most support, and when children were at unusually high risk

of the events that ACH sought to reduce, including readmission to the hospital for behavioral health reasons.

2. Summary of impact findings and possible explanations for them

The findings from the impact evaluation were mixed. We found that the program did not attain one of the program goals—to meaningfully increase the rate of behavioral health follow-up visits within 30 days of discharge. We were unable to conclude whether the program reduced 60-day behavioral health readmissions because our tests did not have sufficient statistical power to reliably detect meaningful program impacts. The program appeared to reduce behavioral health outpatient ED visits by an estimated 14.6 percent. Although these results are promising, the impact estimates were not statistically significant ($p = 0.20$), meaning there was a 20 percent chance we would have observed favorable effects at least that large due to chance alone, if the program truly had no impact. Testing this model again with larger samples could provide more confidence in these early promising results.

One potential mechanism for the observed decrease in behavioral health ED use could have been an increased rate of outpatient behavioral health visits after discharge. However, we found no impact on these visits, suggesting that some other mechanism is likely responsible for the favorable impact of decreased ED use. Because we were unable to test other possible mechanisms directly, we can only speculate, based on our implementation study, about other possible mechanisms. For example, motivational interviewing by the parent partners could have led to greater caregiver self-efficacy, more effective family support for the patient, and greater acceptance of medication treatment after discharge; medication is an important modality of treatment. These changes might in turn have led to reduced risk of episodes requiring an ED visit. In addition, post-discharge telephone interactions by parent partners and the care manager might have supported or enhanced problem solving among caregivers and addressed social needs, alleviating environmental stressors that would otherwise exacerbate a patient's condition and lead to acute care need. Finally, it is possible that caregivers were able to discuss medical concerns with parent partners or the care manager and avoid trips to the ED for minor changes in conditions that did not truly require immediate medical attention.

The lack of impacts on outpatient behavioral health follow-up visits could be due, in part, to there being only modest room for improvement. The follow-up visit rate for those discharged from ACH before the intervention began was already high at 82 percent—close to ACH's target of 85 percent.

3. Lessons for hospitals, health systems, or payers interested in pursuing this innovation

Other hospitals, health systems, or health care payers might be interested in implementing or supporting a peer partner and care management intervention such as ACH's. Here we draw some lessons from the ACH experience for such organizations:

- **ACH's approach to targeting—with a target population of all Medicaid-enrolled children discharged from the behavioral health unit—was effective in four important ways.** First, the target population was clearly defined and it identified beneficiaries with similar behavioral health needs. Second, intervention staff could easily implement a method

to identify the target population, specifically by examining daily lists of inpatient admissions. Third, the targeting approach identified a steady flow of participants over time, helping to spread the workload over time. Fourth, the approach identified patients with a high risk of future acute behavioral health events—thus affording the program an opportunity to reduce rates of acute events.

- **Making peer partners a hired position, and then providing robust training, was critical for making sure peer partners could deliver the intervention model as intended.** Through formal training from outside consultants and weekly meetings with a supervisor, ACH trained peer partners in (1) strategies for effective interactions with behavioral health clinicians, including orienting the parent partners to the roles and hierarchies of hospital staff and having the parent partners participate in multidisciplinary care team discussions; (2) approaches to family-centered care, including building rapport by sharing experiences with their own children and eliciting caregivers' prior experiences with behavioral health care; and (3) motivational interviewing, an evidence-based method that seeks to engage individuals and motivate change by supporting intrinsic motivations.
- **Because of access to discharge records, the care manager knew when patients were discharged and could facilitate timely follow up.** Some transitional care programs are hindered by care managers' lack of access to real-time discharge data, meaning that contacts do not always occur soon after discharge.
- **Programs such as this could face implementation challenges and ACH's experience suggests some ways to overcome them.** These experiences include the following:
  - **Difficulties hiring and retaining experienced peer partners over time.** These difficulties can be at least partly addressed through (1) vetting job candidates well to make sure they have had sufficient experience with the behavioral health care system through their own children to effectively engage with caregivers as peers who have had similar experiences; and (2) allowing schedule flexibility so that parents can tend to their own children's needs, while still performing their job duties. In the case of ACH, flexibility included making positions part-time and allowing peer partners to trade shifts with each other. Shift trading, in particular, can require a cultural adaptation in hospitals where shift assignments tend to be rigid.
  - **Difficulties getting hospital providers to understand and buy into the value that peer partners add.** Programs can promote provider buy-in through frequent face-to-face interactions between peer partners and clinicians (for example, through peer partners' participation in hospital rounds), senior leadership championing or explaining the program to providers, and training peer partners about how to interact with medical professionals.
- **Hospitals should assess their patients' needs and organizational capabilities before implementing a new intervention, so that they can set meaningful goals for improvement.** For example, ACH might not have increased the proportion of targeted patients who received timely outpatient follow-up care, in part because ACH was already close to meeting its target for timely follow-up (85 percent of relevant patients) before the intervention began. The ACH staff might not have realized how little room for improvement

existed, however, because ACH was not able to access the Medicaid claims data needed to assess the extent of post-hospitalization outpatient care at baseline.

- **Implementing an intervention like ACH's might be successful in reducing behavioral health outpatient ED visits,** as suggested by the promising findings for ACH.

REFERENCES

Bardach, Naomi S., Tumaini R. Coker, Bonnie T. Zima, J. Michael Murphy, Penelope Knapp, Laura P. Richardson, Glenace Edwall, and Rita Mangione-Smith. "Common and Costly Hospitalizations for Pediatric Mental Health Disorders." *Pediatrics,* vol. 133, no. 602, 2014, pp. 602–611.

Chinman, Matthew, Preethy George, Richard H. Dougherty, Allen S. Daniels, Sushmita Shoma Ghose, Anita Swift, and Miriam E. Delphin-Rittmon. "Peer Support Services for Individuals with Serious Mental Illnesses: Assessing the Evidence." *Psychiatric Services,* vol. 65, no. 4, 2014, pp. 429–441.

Healthcare Cost and Utilization Project Clinical Classifications Software (CCS) for ICD-9-CM. "Healthcare Cost and Utilization Project (HCUP)." Rockville, MD: Agency for Healthcare Research and Quality, March 2017. Available at www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp. Accessed April 5, 2017.

Mahajan, Prashant, Elizabeth R. Alpern, Jackie Grupp-Phelan, James Chamberlain, Lydia Dong, Richard Holubkov, Elizabeth Jacobs, Rachel Stanley, Michael Tunik, Meredith Sonnett, Steve Miller, and George Foltin. "Epidemiology of Psychiatric-Related Visits to Emergency Departments in a Multicenter Collaborative Research Pediatric Network." *Pediatric Emergency Care,* vol. 25, no. 11, 2009, pp. 715–720.

Medicaid and CHIP Payment and Access Commission. "MACStats: Medicaid and CHIP Program Statistics, FY 2011, Table 14." Baltimore, MD: Centers for Medicare & Medicaid Services, 2011. Available at (https://www.macpac.gov/wp-content/uploads/2015/01/2014_06_MACStats.pdf). Accessed April 10, 2017.

Medicaid and CHIP Payment and Access Commission, "MACStats: Medicaid and CHIP Program Statistics, 2010–2012, Table 3." Baltimore, MD: Centers for Medicare & Medicaid Services, 2012. Available at (https://www.macpac.gov/wp-content/uploads/2015/01/2014_06_MACStats.pdf). Accessed April 10, 2017.

National Center for Health Statistics. "National Health Interview Survey, 2012 Person File." Atlanta, GA: Centers for Disease Control and Prevention, 2012. Available at (https://www.cdc.gov/nchs/nhis/nhis_2012_data_release.htm). Accessed April 20, 2017.

Obrochta, C, B. Anthony, M. Armstrong, J. Kallal, J. Hust, and J. Kernan. "Family-to-Family Peer Support: Models and Evaluation." Issue brief. Atlanta, GA: ICF Macro, Outcomes Roundtable for Children and Families, 2011.

Purington, Kitty. "Using Peers to Support Physical and Mental Health Integration for Adults with Serious Mental Illness." Washington, DC: National Academy for State Health Policy, 2016. Available at http://www.nashp.org/wp-content/uploads/2016/01/Peer-Supports.pdf. Accessed April 14, 2017.

Robbins, V., J. Johnston, H. Barnett, W. Hobstetter, K. Kutash, A.J. Duchnowski, and S. Annis. "Parent to Parent: A Synthesis of the Emerging Literature." Tampa, FL: University of South Florida, The Louis de la Parte Florida Mental Health Institute, 2008.

Smith, Mary L. "Summary of Pediatric Quality Measures for Children Enrolled in MaineCare FFY 2009–FFY 2013." Paper 34. *Population Health and Health Policy,* 2014. Available at http://digitalcommons.usm.maine.edu/healthpolicy/34. Accessed July 12, 2016.

Zickafoose, Joseph, Brenda Natzke, Boyd Gilman, Catherine DesRoches, and Lorenzo Moreno. "Findings for Research Institute at Nationwide Children's Hospital." In Moreno, Lorenzo, Boyd Gilman, Greg Peterson, Catherine DesRoches, Sheila Hoag, Linda Barterian, Laura Blue, Katherine Bradley, Emily Ehrlich, Kristin Geonnotti, Lauren Hula, Keith Kranker, Rumin Sarwar, KeriAnn Wells, Joseph Zickafoose, Sandi Nelson, Kate Stewart, and Frank Yoon. "Evaluation of Health Care Innovation Awards (HCIA): Primary Care Redesign Programs." Second Annual Report to CMS. Princeton, NJ: Mathematica Policy Research, March 2016.

CHAPTER 3 SUPPLEMENTAL MATERIAL

METHODS AND SUPPLEMENTAL RESULTS FOR THE EVALUATION OF RESEARCH INSTITUTE AT NATIONWIDE CHILDREN'S HOSPITAL/ AKRON CHILDREN'S HOSPITAL

Eric Lammers, Joe Zickafoose, Keith Kranker, Greg Peterson, Kate Stewart, Laura Blue, Brenda Natzke, Boyd Gilman, and Sheila Hoag

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

I. INTRODUCTION

This supplement to Chapter 3 provides methods detail and additional results for our evaluation of the Health Care Innovation Award (HCIA) won by the Research Institute at Nationwide Children's Hospital (NCH). As we describe in Chapter 3 of this report, we conducted an impact evaluation for one component (behavioral health) of this HCIA-funded program, as implemented at one NCH partner site: Akron Children's Hospital (ACH). We drew conclusions, based on available evidence, about the impacts of ACH's behavioral health intervention in three domains: (1) quality-of-care processes, (2) quality-of-care outcomes, and (3) service use.

According to the NCH/ACH program's design, ACH intended to achieve impacts on patients' outcomes through intermediate impacts on caregivers' behavior (that is, increased caregiver activation leading to more relevant care plans, better outpatient follow-up, and more appropriate use of acute care) and, to a lesser extent, through changes in providers' behavior (better communication leading to more relevant care plans). We do not assess these intermediate impacts, however, due to lack of suitable data. Similarly, we do not assess impacts on spending due to lack of appropriate detail in the Ohio Medicaid data we acquired for this evaluation.

This supplement to Chapter 3 has five sections, including this introduction. Section II describes the program's theory of action—that is, how program activities were expected to improve outcomes examined in the impact evaluation. Section III describes our methods for estimating impacts in detail. Section IV presents results from the secondary quantitative tests (robustness checks) and a discussion of the concordance between primary test results and the implementation evidence. Section V shows how we used our evaluation framework, described in Appendix 3, to draw conclusions about program effectiveness for the NCH/ACH program.

II. THEORY OF ACTION

Based on our review of ACH's program activities and goals, we developed a theory of action to depict the mechanisms through which the behavioral health intervention would be expected to increase behavioral health outpatient follow-up, decrease behavioral health readmissions after hospitalization, and decrease behavioral health ED visits. NCH and ACH expected that the HCIA-funded intervention would improve outcomes for Medicaid participants through two pathways: (1) supporting participants' caregivers and (2) improved interactions between caregivers and behavioral health care providers as those providers formulate and execute plans for care.

Pathway 1: Supporting caregivers to better engage in their children's care and navigate the behavioral health care system. This pathway includes the following planned mechanisms:

1. The program hires and trains parents who have experience with their own children's behavioral health conditions to serve as parent partners.

- Parent partners are trained to use their experiences and knowledge to communicate and empathize with caregivers and promote family-centered care by providers.

2. Parent partners meet with caregivers soon after a child's admission and continue to support caregivers throughout the hospitalization, leading to increased caregiver engagement in care and improved communication between caregivers and providers.
   - These improvements include more active participation of caregivers in daily rounds, care planning, and therapy sessions.
3. Improved communication between caregivers and providers leads to improved, family-centered inpatient care plans for children admitted for behavioral health conditions.
   - Care plans could more comprehensively address caregivers' understanding of conditions and expectations for care, identify families' social resources and needs, and coordinate the activities of participants' existing behavioral and physical health care providers.
4. Improved inpatient care plans result in greater clinical improvement during hospitalization.
   - This could include improved symptoms and increased engagement in care for both participants and caregivers.
5. Improved communication between caregivers and providers and improved inpatient care plans lead to improved discharge care plans.
   - These improvements could be similar to those for inpatient care plans (in item 3).
   - Discharge care plans could also include better understanding of participants' psychosocial needs and could, therefore, better meet the needs and resources of participants and their families.
6. Improved discharge care plans and follow-up calls by the care manager increase the likelihood of outpatient follow-up.
   - These improvements increase the frequency of outpatient behavioral health follow-up by identifying and addressing potential barriers to follow-up care.
7. Improved inpatient and discharge care plans and increased outpatient follow-up leads to reduced behavioral health symptoms and increased caregiver activation.
   - These improvements include decreased impairment due to behavioral health symptoms for participants and greater caregiver ability to manage symptoms and access care as needed.
8. Reduced behavioral health symptoms and increased caregiver activation leads to decreased need for acute behavioral health care.
   - Specifically, these improvements reduce the frequency of emergency department (ED) visits and readmissions for behavioral health needs following discharge by more thoroughly addressing needs during the hospitalization, connecting the participant to appropriate outpatient care, and reducing barriers to outpatient follow-up.

Pathway 2: Promoting family-centered communication and care planning by inpatient behavioral health providers. This pathway includes the following mechanisms:

1. Parent partners contribute to behavioral health team rounds, leading to improved family-centered communication by providers.
   - Providers have more positive interactions with caregivers and improve how they identify participants' psychosocial needs.
2. Improved family-centered communication by providers leads to more family-centered inpatient and discharge care plans.
   - Plans incorporate a more comprehensive approach to the participants' behavioral health needs, such as prior positive or negative experiences with specific providers and medications, and psychosocial needs, such as housing insecurity.
3. Improved inpatient and discharge care plans increase outpatient follow-up and decrease ED visits and readmissions.
   - Because of better communication between caregivers and providers, caregivers become more knowledgeable and activated, which helps them better connect with follow-up care, manage their child's symptoms, and access care as needed. This leads to increased outpatient follow-up and decreased ED visits and readmissions.

III. METHODS FOR ESTIMATING PROGRAM IMPACTS ON PATIENTS' OUTCOMES

A. Overview

We estimated program impacts using a difference-in-differences framework. To implement this framework, we defined two cohorts of Medicaid child beneficiaries, as described in the NCH/ACH chapter: (1) a *post-intervention cohort,* which included beneficiaries discharged from ACH from August 1, 2013, to May 31, 2015, and who met the program eligibility criteria (the post-intervention treatment group) and their matched comparison beneficiaries (the post-intervention comparison group); and (2) a *pre-intervention cohort,* which included beneficiaries discharged from October 1, 2010, to December 31, 2012—that is, at least seven months before the start of the intervention period (August 2013)—but who otherwise met the program eligibility criteria (the pre-intervention treatment group) and their matched comparison beneficiaries (the pre-intervention comparison group). To estimate impacts, we (1) calculated the difference in outcomes between the post-intervention treatment and comparison groups in each outcome-specific time period (that is, 30, 60, or 90 days) following discharge; and (2) subtracted any difference in outcomes between the pre-intervention treatment and comparison groups, using regression adjustment to account for any remaining differences in other characteristics between the groups after matching.

Before conducting the analysis, we prespecified primary tests, describing the evidence we would need to conclude that the program was effective, and the awardee and the Center for Medicare & Medicaid Innovation (CMMI) reviewed these tests. Each test specified a population, outcome, time period, expected direction of effect, and threshold that we counted as

substantively important. The purpose of these primary tests was to focus the impact evaluation on hypotheses that would provide the most robust evidence about program effectiveness. We used the results from the primary tests and secondary tests (robustness checks) described below to draw conclusions about program impacts in each of the three evaluation domains we examined: (1) quality-of-care process, (2) quality-of care outcomes, and (3) service use. The following subsections describe each component of the impact evaluation in more detail.

B. Data

We analyzed Ohio Medicaid enrollment and claims data covering 2008 to 2015. The Ohio Colleges of Medicine Government Resource Center provided these data, as approved by the Ohio Department of Medicaid.

The enrollment files provided information, by month, for children enrolled in Medicaid during the study period, including enrollment in Medicaid managed care (MMC) or fee-for-service (FFS), whether the child had any third-party coverage, whether the child was blind or disabled, whether the child was enrolled in a home and community-based services waiver, and whether the child was dually enrolled in Medicare and Medicaid. The enrollment files also provided basic demographic data and the date, if applicable, a child died.

The claims files contained data on services received that Ohio Medicaid covered (paid by either FFS Medicaid or managed care plans). These included claims from two different database systems: one covering the period before August 2011 and the other covering August 2011 onward. To account for possible overlap in inpatient claims across the two systems and to facilitate our analyses, we constructed stay-level inpatient files using combined data sets of adjudicated inpatient claims. (Appendix 1 provides detail.) We used the resulting stay-level data set to identify admissions that qualified a beneficiary for the treatment or comparison group. Neither database system included payment information for the managed care records, which represent the bulk of the services paid.

C. Treatment group definition

Post-intervention treatment group. The post-intervention treatment group comprised Medicaid beneficiaries (both managed care and FFS) younger than 18 who had an inpatient discharge from ACH's 12-bed psychiatric inpatient unit from August 1, 2013, through May 31, 2015. We ended the enrollment period on May 31, 2015, so that each person in the post-intervention treatment group was potentially exposed to the HCIA-funded intervention for at least one month before the funding ended on June 30, 2015.

We imposed four restrictions on treatment group membership. First, only those whose first inpatient stay in a behavioral health inpatient unit during the intervention enrollment period occurred at ACH qualified for the treatment group. (Those beneficiaries who had a stay at ACH during the intervention enrollment period but who also had an earlier stay during the intervention enrollment period at a comparison hospital were part of the group of potential comparison beneficiaries; see Section III.D.) For each member of the treatment group, we examined outcomes during only the 90-day period after the first inpatient stay at ACH, even if a

beneficiary had more than one inpatient stay during the intervention enrollment period. This approach excluded a few beneficiaries from the treatment group who might have received the treatment through a stay at ACH after an initial stay during the intervention enrollment period at a comparison hospital. It also excluded examining outcomes after second or subsequent visits to a treatment or comparison hospital. Despite excluding a few beneficiaries who might have received the treatment (or cases in which a beneficiary received the treatment for a second time or more), we applied these exclusions to minimize the potential for cross-contamination between the treatment and comparison groups.

Second, we required each treatment beneficiary to be continuously enrolled in Medicaid for the four quarters (that is, 12 months) before the qualifying discharge. This restriction improved the matching of treatment to potential comparison beneficiaries by ensuring we could use a full year of claims to develop baseline indicators of service use and diagnoses for matching.

Third, to be included in the analytic sample during the intervention period (that is, the intervention quarter over which we estimate impacts), each treatment group member had to be observable in Medicaid claims for at least one day during the quarter. We considered beneficiaries to be observable if they were enrolled in Ohio Medicaid.

Finally, we excluded beneficiaries whose stay did not have a behavioral health primary diagnosis.

Pre-intervention treatment group. We defined the pre-intervention group using the same claims-based rules as for the post-intervention group, with one difference: each beneficiary had to have been discharged from ACH from October 1, 2010, to December 31, 2012.

Intent-to-treat criteria. The claims-based rules we used to define the treatment group have two advantages over an alternative definition that includes only those who actually enrolled in ACH's behavioral health intervention. First, because the parent partner and care coordination components targeted all Medicaid patients discharged from the ACH psychiatric inpatient unit, our treatment group definition corresponds to everyone the program intended to treat (that is, the definition follows an intent-to-treat analysis). Notably, the claims-based definition includes Medicaid beneficiaries who parent partners could not contact, possibly due to staff scheduling constraints on some weekends. Second, we can use the same eligibility rule to identify a pre-intervention treatment group, which substantially improves the credibility of the difference-in-differences design. When comparing our treatment group definition to the roster provided by ACH of actual intervention participants, we found that 71 percent of post-intervention treatment group members had enrolled in the parent partner and care coordination intervention components. Therefore, any impacts measured among the full treatment group might understate the impacts among those who enrolled. The awardee listed 639 Medicaid beneficiaries as participants during the period of our analysis (August 1, 2013, to May 31, 2015). Of these, our treatment group included 61 percent. Our treatment group likely did not include the other 39 percent of enrollees due to (1) the restrictions we placed on the sample, such as limiting to those observable in claims for 12 months before enrollment; and (2) potential differences in how beneficiaries were recorded in ACH records and Medicaid claims, which would prevent linking.

D. Comparison group definition

We constructed a comparison group of Ohio Medicaid beneficiaries who were similar to the treatment group beneficiaries. This section describes how we constructed the matched comparison group; the main chapter text presents the balance we achieved between the two groups on the matching variables.

We used the following steps to construct the comparison group:

First, we identified a pool of potential comparison members. This pool consisted of all Ohio Medicaid beneficiaries younger than 18 discharged from August 1, 2013, to May 31, 2015 (for the post-intervention cohort), or October 1, 2010, to December 31, 2012 (for the pre-intervention cohort), from an inpatient behavioral health facility at any of nine comparison hospitals in Ohio: (1) Belmont Pines in Youngstown, (2) Fairview Hospital in Cleveland, (3) Windsor Laurelwood Center for Behavioral Medicine (also known as HHC) in Willoughby, (4) Kettering Medical Center in Kettering, (5) OhioHealth (formerly MedCentral) in Mansfield, (6) Medical University of Ohio at Toledo, (7) Toledo Hospital, (8) University Hospitals of Cleveland Medical Center, and (9) Upper Valley Medical Center in Troy.

These nine hospitals are plausible representations of the counterfactual because each serves a high volume of pediatric psychiatric inpatients (more than 500 cases during the intervention enrollment period).[1] Moreover, five of these are academic hospitals with psychiatric inpatient units, like ACH. Two of the comparison hospitals are freestanding psychiatric hospitals (Belmont and Windsor Laurelwood) and the other two are community hospitals (OhioHealth and Upper Valley). We considered but ultimately excluded a 10th comparison hospital (Cincinnati Children's Hospital) because its volume of behavioral health cases was an outlier (much larger) relative to ACH and the nine comparison hospitals. We also observed a significant increase in behavioral health cases from the pre-intervention to the post-intervention cohorts, suggesting an unusual increase in resources to behavioral health inpatient care over this time. We set the day following hospital discharge as the potential comparison beneficiary's pseudo-enrollment date (that is, the date we assigned a beneficiary to the potential comparison pool).

Second, we used the Ohio Medicaid Enrollment files and each beneficiary's Medicaid claims in the 12 to 24 months before his or her pseudo-enrollment date to develop baseline characteristics for each beneficiary. As with the treatment group definition, to facilitate matching, we required potential comparison beneficiaries to have continuous enrollment in Medicaid for the 12 months before their discharge date.

Finally, we used propensity-score matching and exact matching techniques to limit the potential comparison pool to a list of matched comparison beneficiaries. Matching aims to reduce selection bias in observational studies by selecting comparison beneficiaries from the pool who are roughly equivalent to the treatment group across key baseline characteristics. The goal of matching is to achieve baseline equivalence between the treatment and matched

[1] The counterfactual is the outcome the treatment group would have had in the absence of the intervention.

comparison groups on the variables included in the matching process (Stuart 2010). We matched on primary diagnosis during the enrollment hospitalization, demographic characteristics, health status and chronic conditions, service use 3 months before discharge, and service use 4 to 12 months before discharge.

We used exact matching techniques to ensure that matched comparison group beneficiaries had (1) a qualifying inpatient discharge within 90 days of the treatment beneficiary's enrollment date, (2) the same gender as the treatment beneficiary, and (3) the same reason for the hospitalization as the treatment beneficiary. To identify reason for hospitalization, we used seven categories of psychiatric diagnosis codes sorted into clinically distinct groups as defined by the Healthcare Cost and Utilization Project's Clinical Classification Software [CCS] (Healthcare Cost and Utilization Project 2017).

Applying these exact match restrictions, we implemented a technique called full matching to form matched sets that contain one treatment beneficiary and one or more comparison beneficiaries. Full matching achieves maximum bias reduction on observed matching variables and, subject to this constraint, maximizes the size of the comparison sample (Rosenbaum 1991; Hansen 2004). Each treatment beneficiary was matched to up to five beneficiaries from the potential comparison group. Three of the 543 (0.007 percent) post-intervention treatment beneficiaries were dropped because no comparison beneficiaries could be found within the same strata of exact matching variables (in particular, within the age and date of discharge calipers for these cases). None of the 605 pre-intervention treatment beneficiaries were dropped.

Additional sample restrictions. To be included in the analytic sample, a comparison group beneficiary had to meet four additional criteria. First, like the treatment group members, the comparison beneficiaries had to be observable, with regard to outcomes, in Medicaid claims for at least one day during the quarter. Second, we excluded beneficiaries from the potential comparison group if their qualifying hospital stay involved a transfer in which the originating or middle stay was at ACH or NCH; we did this to avoid contaminating the comparison group with stays that could have been exposed to the intervention. Third, as a further measure to avoid contamination of the comparison group, we excluded beneficiaries who resided in the 34-county service region of NCH/Partners for Kids (PFK). Finally, we excluded beneficiaries whose stay did not have a behavioral health primary diagnosis. Table III.1 depicts the effect of each exclusion on the number of beneficiaries in the sample.

Source of comparison group hospital discharges, by cohort. As shown in Table III.2, all comparison-hospitals contributed to the comparison groups, with no single hospital accounting for more than 30 percent of discharges. The distribution of discharges across the comparison hospitals was similar in the pre- and post-intervention cohorts.

Table III.1. Defining the analytic sample, by cohort

Number of beneficiaries				
Total number of discharges from behavioral health inpatient units at any point from 2010 to 2015	124,594			
As row above, but limited to stays during pre- or post-intervention enrollment period for comparison group	85,158			
Stays for which person is younger than 18 at time of discharge	33,713			
	Pre-intervention cohort		**Post-intervention cohort**	
As above, but broken out by cohort	17,919		15,794	
Stays remaining after selecting the first stay in the enrollment period	12,118		10,934	
Stays flagged as a treatment or potential comparison (that is, at ACH or one of 9 comparison hospitals)	9,138		8,738	
	Treatment	**Potential comparison**	**Treatment**	**Potential comparison**
As above, but broken out by the treatment and potential comparison groups	755	6,571	670	5,609
Stays after removing comparison stays with transfers in which the originating or middle stay is at ACH or NCH	755	6,564	670	5,604
As row above, but limited to cases in which beneficiary is observable for all 12 months before enrollment and at least 1 day during the 3 months after enrollment	618	5,301	544	4,524
As row above, but limited to potential comparison cases in which the beneficiary does not live in one of the 34 NCH or PFK counties	618	5,064	544	4,309
As row above, but excluding stays without a behavioral health primary diagnosis	605	4,876	543	4,237
Final matched treatment and comparison groups	605	2,249	540	1,952

Source: Analysis of Ohio Medicaid claims data provided by the Ohio Colleges of Medicine Government Resource Center.

ACH = Akron Children's Hospital; NCH = Nationwide Children's Hospital; PFK = Partners for Kids.

Table III.2. Comparison hospitals and percentages of all comparison discharges

		Pre-intervention comparison group		Post-intervention comparison group	
Hospital	Location	Count	Percentage	Count	Percentage
Belmont Pines	Youngstown	670	29.8	389	19.9
Fairview Hospital (Cleveland Clinic)	Cleveland	73	3.2	161	8.2
Windsor Laurelwood Center for Behavioral Medicine (HHC)	Willoughby	518	23.0	425	21.8
Kettering Medical Center	Kettering	261	11.6	238	12.2
OhioHealth (MedCentral)	Mansfield	59	2.6	50	2.6
Medical University of Ohio at Toledo	Toledo	227	10.1	218	11.2
Toledo Hospital	Toledo	134	6.0	172	8.8
University Hospitals Cleveland Medical Center	Cleveland	189	8.4	168	8.6
Upper Valley Medical Center	Troy	118	5.2	131	6.7
Totals		**2,249**		**1,952**	

Source: Mathematica's analysis of Ohio Medicaid claims data.

E. Constructing outcomes and covariates

We used Ohio Medicaid claims for beneficiaries assigned to the treatment and comparison groups to develop two types of variables:

1. *Outcomes,* defined for each beneficiary in the quarter during which they were members of the pre-intervention or post-intervention treatment or comparison group.
2. *Covariates,* which describe a beneficiary's characteristics at the time of enrollment or pseudo-enrollment and were used in the regression models for estimating impacts to adjust for differences in these characteristics. Appendix 1 provides details on the methods we used to construct these variables.

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

1. Quality-of-care processes
   - Outpatient behavioral health follow-up visit within 30 days of discharge from the enrollment (or pseudo-enrollment) stay (binary indicator [yes/no] for each beneficiary)
2. Quality-of-care outcomes
   - Unplanned readmission for behavioral health condition within 60 days of discharge from the enrollment (or pseudo-enrollment) stay; calculated as whether a beneficiary had an

inpatient admission with a behavioral health primary diagnosis (identified using *International Classification of Diseases,* 9th edition, codes) within 60 days following the enrollment admission(binary indicator [yes/no] for each beneficiary)

3. Service use
   - Behavioral health-related outpatient ED visit rate; we define outpatient ED visits as ED visits that do not end in a hospital admission, and define behavioral health visits using the same diagnostic criteria as for readmissions (number/beneficiary/quarter)

Covariates. The covariates, defined at the date of enrollment (treatment group) or pseudo-enrollment (comparison group) include the following:

1. Demographics (age at enrollment or pseudo-enrollment, race, and ethnicity)
2. Whether enrolled in Medicaid managed care
3. Whether enrolled in a home and community-based services (HCBS) waiver
4. Disability status
5. Whether there was a diagnosis for substance abuse at the enrollment stay
6. Whether a beneficiary has behavioral health conditions in each of 12 CCS categories created by applying CCS algorithms to claims in the 24 months before each beneficiary's enrollment or pseudo-enrollment date:
   - Adjustment disorders
   - Anxiety disorders
   - Attention deficit, conduct, and disruptive behavior disorders
   - Disorders usually diagnosed in infancy, childhood, or adolescence
   - Impulse control disorders
   - Mood disorders
   - Personality disorders
   - Schizophrenia and other psychotic disorders
   - Alcohol-related disorders
   - Substance-related disorders
   - Suicide and intentional self-inflicted injuries
   - Miscellaneous mental health disorders
7. A claims-based Chronic Illness and Disability Payment System (CDPS) score from the prior 24 months
8. Service use in the prior 3 months and prior 4 to 12 months; service use includes the number of hospitalizations for behavioral health, the number of hospitalizations for other reasons, the number of outpatient ED visits or observation stays for behavioral health, the number of

outpatient ED visits or observation stays for other reasons, any use of hospital outpatient services with a behavioral health primary diagnosis, and any outpatient behavioral health care (hospital and nonhospital)

F. Regression model

We used a regression model to implement the difference-in-differences framework. For each quarter-specific outcome, the model estimates the relationship between the outcome and predictor variables, assuming that each of the predictor variables has a linear (additive) relationship with the outcome. The predictor variables included the following:

1. The beneficiary-level covariates (defined in Section III.E)
2. An interaction of the indicator for enrollment in MMC with an indicator for the post-intervention period (allowing intervention period outcomes to differ by MMC enrollment)
3. An interaction of each CCS category of behavioral health diagnoses in the 24 months before enrollment or pseudo-enrollment with the post-intervention indicator (allowing intervention period outcomes to differ by behavioral health condition)
4. An interaction of the CDPS score in the 24 months before enrollment or pseudo-enrollment with the post-intervention indicator (allowing intervention period outcomes to differ by CDPS score)
5. Indicators for each matched set (that is, fixed effects for a treatment beneficiary plus his or her matched comparison beneficiaries)
6. Whether the beneficiary was assigned to the treatment or comparison group
7. An interaction of a beneficiary's treatment status with an indicator for being in the post-intervention cohort (opposed to the pre-intervention cohort)

Appendix 2 provides details on the regression methods, including descriptions of the weights each beneficiary receives in the model and how the regressions account for correlation in outcomes across beneficiaries in the same matched set.

The estimated relationship between the interaction of treatment status with post-intervention cohort provides the difference-in-differences estimate for that outcome. It measures the average difference between outcomes for post-intervention beneficiaries assigned to the treatment and comparison groups, subtracting out any differences between the pre-intervention treatment and comparison groups. Estimating the model allows for statistical tests that determine whether observed differences are likely due to chance alone.

G. Primary tests

The NCH/ACH chapter lists the primary tests that we used to evaluate impacts of the HCIA-funded program on program impacts. Our rationale for selecting these primary tests is as follows:

Outcomes. The awardee expected impacts would reduce the number of 60-day behavioral health readmissions and increase 30-day outpatient follow-up visits after a behavioral health

hospitalization. We also hypothesized that ACH would reduce the number of outpatient ED visits with a behavioral health primary diagnosis within 90 days of the enrollment stay. ACH expected to affect spending as well; as we noted earlier, however, we were unable to examine the impact of the intervention on spending due to limitations in the Ohio Medicaid data.

Time period. We examined impacts during the first quarter (that is, three months) after the enrollment stay. The intervention focused on the period shortly after discharge, and the awardee expected the program's impacts to be concentrated within these first three months after discharge. We used time periods of less than a full quarter when the awardee explicitly stated an expected period for the impact. For example, for behavioral health readmissions, we examined impacts during the first 60 days following discharge from the enrollment stay and, for outpatient follow-up visits after a behavioral health hospitalization, we examined impacts during the first 30 days following the discharge of the enrollment stay.

Population. ACH expected to affect all Medicaid-enrolled children with behavioral health conditions who had an inpatient stay at ACH in its 12-bed psychiatric unit during the intervention enrollment period.

Direction (sign) of the impact estimate. We tested for a reduction, relative to the estimated counterfactual, for 60-day behavioral health readmissions and 90-day behavioral health ED visits. We tested for an increase, relative to the estimated counterfactual, for 30-day outpatient follow-up visits after a behavioral health hospitalization among treatment group members.

Substantive thresholds. Some impact estimates could be large enough to be substantively interesting to CMMI) and other stakeholders even if they are not statistically significant; for this reason, we specified thresholds for what we call *substantive importance*. ACH expected an 8 percent reduction in the 60-day behavioral health readmission rate. We chose a substantive threshold of 6 percent, or 75 percent of the awardee's target, recognizing that a program could still be considered effective even if it did not have its full anticipated impact. ACH did not specify targets for its goals of (1) increasing 30-day outpatient follow-up visits after a behavioral health hospitalization or (2) reducing behavioral health-related ED visits. Therefore, we selected thresholds based on a prior study of similar interventions (Smith 2014) and clinical expertise. We chose substantive thresholds of a 7.5 percent increase for 30-day outpatient follow-up visits and a 7.5 percent decrease for ED visits.

H. Secondary tests (robustness checks)

We also conducted secondary quantitative tests to help corroborate the findings from the primary tests. This is important because some of the differences observed between the treatment and comparison groups in the primary test results could reflect limitations of the nonexperimental impact evaluation design or random fluctuations in the data. We have greater confidence in the primary results if they are generally consistent with the expected broader pattern of results from the secondary tests. Specifically, we repeated the primary tests described earlier, but excluded from the sample all matched sets of beneficiaries in which at least one treatment or comparison beneficiary had nonbehavioral health hospitalizations during the three months before the qualifying hospitalization. We did this because—as we show in the

NCH/ACH chapter—nonbehavioral health hospitalizations in the prior three months was the one characteristic for which we could not achieve good balance through matching between treatment and comparison groups. For this analysis, we dropped 167 treatment beneficiaries (89 in the pre-intervention cohort and 78 in the post-intervention cohort) and 628 matched comparison beneficiaries (317 and 311, respectively, in the pre- and post-intervention cohorts). These exclusions reduced the size of our sample by 15 percent for the secondary tests relative to the sample size for the primary tests.

I. Synthesizing evidence to draw conclusions

Although we estimated impacts on only one outcome within each of the three evaluation domains (quality-of-care processes, quality-of-care outcomes, and service use), we drew conclusions about program impacts at the domain level; we did this for consistency with the other awardee evaluations within the HCIA Primary Care Redesign portfolio. Within each of the three domains, we could draw one of five conclusions, based on the primary test results, the results of secondary tests, and the plausibility of those findings given the implementation evidence:

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

We could not conclude that a program had a statistically significant unfavorable effect. This is because, in consultation with CMMI, we decided to use one-sided statistical tests (which do not test for evidence of unfavorable effects).

Appendix 3 describes our decision rules for each of the five possible conclusions. In short, we concluded that a program had a statistically significant favorable effect in a domain if the one primary test result in the domain was favorable and statistically significant. We also had to determine that the primary test results were plausible, given the results of the secondary tests and implementation evidence. We concluded that a program had a substantively important favorable effect if the impact estimate in the domain was substantively important but not statistically significant, and if the result was plausible given the secondary tests and implementation evidence. In contrast, if the impact estimate was unfavorable (opposite the hypothesized direction) and larger than the substantive threshold, and unfavorable effects were plausible given the other evidence, we concluded the program had a substantively important unfavorable effect. If the tests in a domain did not meet any of these criteria, we drew one of two conclusions. First, if the test in a domain had sufficient statistical power to detect an impact of the size of the substantive threshold with at least 75 percent probability, we concluded that there was not a substantively large effect, because we are reasonably confident that we would have detected a substantively large effect had there been one. Alternatively, if the power was not sufficient to detect this type of impact, we concluded the impact in the domain was indeterminate.

Indeterminate means either that the program truly did not have effects that were substantively large, or that it did, but our statistical tests were not able to detect them.

IV. SUPPLEMENTAL RESULTS

Before drawing conclusions about program impacts, we assessed the plausibility of the primary test results in light of the secondary test results and implementation evidence. This section presents the secondary test results and relevant implementation evidence.

A. Results of secondary tests

The results of secondary tests for all three outcomes were similar in some respects to their corresponding primary tests. However, the results for two outcomes in particular (outpatient behavioral health follow-up visits and behavioral health readmissions) differed somewhat between their respective primary and secondary tests. The secondary test for the third outcome (outpatient behavioral health ED visits) was consistent with its respective primary test. Table IV.1 summarizes the results.

As described in Section III.H, for the secondary tests, we restricted the sample to matched sets in which no treatment or comparison beneficiary had a nonbehavioral health admission in the three months leading up to the qualifying hospitalization. In the secondary test for outpatient behavioral health follow-up visits within 30 days of discharge we found a treatment group mean of 85.93 percent, which is 5.40 percent (4.37 percentage points) greater than the estimated counterfactual. This was smaller than the substantive threshold of 7.50 percent, similar to the primary test result, but it was statistically significant on a one-sided test ($p = 0.05$). Because the point estimates for the primary and secondary tests were similar (2.73 and 4.37, respectively), we regard the findings as consistent across the two tests for this measure. (We do not consider a statistically significant secondary test result, by itself, to be evidence of program impacts. The purpose of the primary tests was to prespecify the level of evidence we would accept as an indication of impacts.)

In the secondary test for behavioral health readmissions within 60 days of discharge, we found a treatment group mean of 9.74 percent, which is 0.94 percentage points (8.80 percent) lower than the counterfactual implied by the difference-in-differences regression model. Although this result is in the opposite (favorable) direction from the primary test result (of a 4.60 percent unfavorable estimate), we do not consider the secondary test results to contradict the primary test results because of the extremely low statistical power to detect effects (see chapter text)—that is, because of the high degree of statistical uncertainty in the estimates.

Finally, the secondary test result for outpatient behavioral health ED visits was similar to the primary test result for this outcome. Both showed a substantively important decrease in the number of outpatient behavioral health ED visits. Therefore, the primary test result for this measure is plausible given this secondary test result.

Table IV.1. Results of secondary tests for ACH's behavioral health intervention: Robustness checks

Secondary test definition[a] (robustness checks)				Results			
Domain	Outcome (units)	Time period for impacts (days following the qualifying discharge)[b]	Population	Treatment group mean	Difference-in-differences estimate, regression adjusted (standard error)	Percentage difference[c]	*p*-value[d]
Quality-of-care processes	Any outpatient behavioral health follow-up visit within 30 days of discharge (percentage)	30 days	All Medicaid children (ages birth to 17 years) in the treatment group, excluding matched sets with a nonbehavioral health inpatient admission in the 3 months before the qualifying stay[d]	85.93	4.37 (2.65)	5.4	0.05
Quality-of-care outcomes	Any readmission with a behavioral health primary diagnosis within 60 days of discharge (percentage)	60 days	All Medicaid children (ages birth to 17 years) in the treatment group, excluding matched sets with a nonbehavioral health inpatient admission in the 3 months before the qualifying stay[d]	9.74	-0.94 (2.41)	-8.8	0.35
Service use	Number of outpatient behavioral health ED visits within 90 days of discharge (#/1,000 beneficiaries)	90 days	All Medicaid children (ages birth to 17 years) in the treatment group, excluding matched sets with a nonbehavioral health inpatient admission in the 3 months before the qualifying stay[d]	166.7	-24.7 (35.9)	-12.9	0.25

Source: Analysis by Mathematica of Ohio Medicaid claims data provided by the Ohio Colleges of Medicine Government Resource Center.

[a] The secondary test definitions were exactly the same as the primary tests except for the population. In the secondary tests, we excluded from the sample all matched sets of beneficiaries in which at least one treatment or comparison beneficiary had a nonbehavioral health hospitalization during the three months before the qualifying hospitalization.

[b] The qualifying stay is the inpatient stay in the hospital's behavioral health unit that led to a beneficiary being assigned to the treatment group.

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

[d] *p*-values are one-sided, testing for evidence of favorable impacts.

ACH = Akron Children's Hospital; ED = emergency department; HCIA = Health Care Innovation Award.

B. Consistency of impact estimates with implementation findings

Based on the implementation findings, it is plausible that ACH's behavioral health intervention had some, but perhaps not all, of its intended effects on patients' outcomes. ACH implemented the parent partner and care management intervention components mostly as planned, although the program faced some challenges with staffing the parent partner program. The impact estimate in the primary test for one domain, service use (outpatient behavioral health ED visit measure), showed favorable substantively large effects. However, the impact estimates for the measures in the other two domains, quality-of-care processes (outpatient behavioral health follow-up visits within 30 days of discharge) and quality-of-care outcomes (behavioral health readmissions within 60 days of discharge), were neither substantively large nor statistically significant. It is possible that the program was unable to change participant caregiver or provider behaviors in ways that would affect all primary test outcomes.

V. CONCLUSIONS ABOUT PROGRAM IMPACTS, BY DOMAIN

This section provides detail on how we used our evaluation decision rules to draw the conclusions we present about program impacts in the NCH/ACH chapter. Table V.1 summarizes the conclusions and their support.

- **The behavioral health intervention had no substantively large impact on quality-of-care processes, as measured by outpatient behavioral health follow-up visits within 30 days of discharge.** The primary test result neither reached the prespecified threshold for substantive importance nor was it statistically significant, although the test was well powered to detect a substantively large effect. This null finding is plausible given the implementation evidence. It is possible that some of the implementation challenges we identified, including inconsistent staffing levels and staff turnover, contributed to the lack of effects. Alternatively, it is possible the program simply did not have its intended effects despite a generally effective implementation.
- **The behavioral health intervention had an indeterminate effect on quality-of-care outcomes, as measured by behavioral health readmissions within 60 days of discharge.** The primary test result in this domain was neither substantively large nor statistically significant. However, we lacked statistical power to detect an effect the size of the substantive threshold. Thus we cannot be certain whether the program truly had no substantively large effects, or whether it did but our tests failed to detect them.
- **The behavioral health intervention had a substantively important favorable effect on service use, as measured by outpatient behavioral health ED visits within 90 days of discharge.** The primary test result for this domain showed a substantively important decrease in ED visits for behavioral health reasons. The evidence from the secondary test corroborates this conclusion by showing a decrease in this measure of a similar magnitude, and evidence from the implementation analysis further corroborates the finding because the program might have been able to decrease some need for acute care following the initial discharge from the hospital.

Table V.1. Conclusions about the impacts of ACH's behavioral health intervention on patients' outcomes, by domain

		Evidence supporting conclusion		
Domain	Conclusion	Primary test result(s) that supported conclusion	Primary test result plausible given secondary tests?	Primary test result plausible given implementation evidence?
Quality-of-care processes	No substantively large impact	The estimate for outpatient behavioral health follow-up visits within 30 days of discharge is neither statistically significant nor substantively large. The statistical power to detect a substantively large effect is good.	Yes	Yes
Quality-of-care outcomes	Indeterminate	The estimate for 60-day behavioral health readmissions is neither statistically significant nor substantively large. However, the statistical power to detect a substantively large effect is poor.	Yes	Yes
Service use	Substantively important (but not statistically significant) favorable effect	The estimate for outpatient behavioral health ED visits within 90 days of discharge is favorable and substantively large but not statistically significant.	Yes	Yes

Sources: Primary test results, secondary test results, and implementation evaluation evidence.

ACH = Akron Children's Hospital; ED = emergency department.

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

REFERENCES

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

Healthcare Cost and Utilization Project Clinical Classifications Software (CCS) for ICD-9-CM. "Healthcare Cost and Utilization Project (HCUP)." Rockville, MD: Agency for Healthcare Research and Quality, March 2017. Available at www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp. Accessed April 5, 2017.

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

Smith, Mary L. "Summary of Pediatric Quality Measures for Children Enrolled in MaineCare FFY 2009–FFY 2013." Paper 34. *Population Health and Health Policy,* 2014. Available at http://digitalcommons.usm.maine.edu/healthpolicy/34. Accessed July 12, 2016.

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

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

CHAPTER 4

UNIVERSITY HOSPITALS OF CLEVELAND

Andrea Wysocki, Joe Zickafoose, Greg Peterson, Kate Stewart, Laura Blue, Keith Kranker, Brenda Natzke, Boyd Gilman, and Sheila Hoag

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

UNIVERSITY HOSPITALS OF CLEVELAND

CHAPTER SUMMARY

Introduction. University Hospitals of Cleveland (UHC) received a three-year, $12.8 million Health Care Innovation Award (HCIA) to transform the delivery of health care for children enrolled in Medicaid in northeastern Ohio. The program had five clinical components, although the core of the intervention was practice facilitation to promote quality improvement focused on preventive care in 32 primary care practices.

Objectives. This report aims to (1) describe the design and implementation of UHC's intervention, with a focus on two components (primary care practice facilitation and incentives to promote the delivery of preventive care services and avoidance of emergency department [ED] visits) during the intervention period of January 2013 to December 2015; (2) assess impacts of the intervention on Medicaid beneficiaries' quality of care and service use; and (3) use both implementation and impact findings to identify possible explanations for the observed impacts.

Methods. We reviewed UHC's program documents and internal measures of performance and conducted site visits and interviews with UHC program leadership and staff. We used a difference-in-differences design with a matched comparison group to estimate the impacts of the intervention on children (from birth to 18 years) enrolled in Medicaid (fee-for-service [FFS] and managed care). Using Medicaid FFS claims and managed care encounter data, we estimated impacts as the differences in outcomes between Medicaid children served by 37 practice locations (corresponding to the 32 practices in the primary care practice facilitation component) and Medicaid children served by 109 matched comparison practices, minus the differences in outcomes between these groups during a period of one year or six months (depending on data available for the practice) before the intervention began, adjusting for differences in other measured characteristics among the children. Our outcome measures included (1) the proportion of children 2 years of age who had a lead screening test, (2) the proportion of children ages 3 to 6 years with a well-child visit in the past year, (3) the proportion of children ages 3 to 5 years with treatment for dental caries in the past year, and (4) the outpatient ED visit rate among children of all ages. The comparison practices were well matched to treatment group practices on most characteristics at baseline, including rates of lead screening, well-child visits, and outpatient ED visits. We were unable to match well on dental caries treatment rates, which was higher among comparison practices at baseline.

Program design and implementation. In the primary care practice facilitation component, UHC partnered with primary care practices for care improvement activities and a linked financial incentive program. Some practices comprised clinicians employed by an affiliate within UHC's health system whereas others comprised independent clinicians. Practice facilitators from UHC visited the participating practices to assess performance on quality measures through chart audits. They then shared measure results with providers and staff and discussed opportunities for improvement. Practices received semiannual incentive payments if they reached thresholds for the quality measures. UHC began this component in the first year of the program, as planned, and enrolled 32 practices (exceeding its goal of 28) that provided care for more than 71,000 child

Medicaid beneficiaries (exceeding its goal of 68,000). The awardee paid incentives to qualifying practices at six-month intervals, as planned.

Five elements comprised the ED avoidance component:

1. An enhanced nurse-led telephone triage service, which required a consultation between caregivers and a physician for all children triaged to the ED by a nurse and allowed a nurse to call in certain medications for children with low-acuity conditions per standing orders of a physician
2. Community-based telemedicine hubs, at which a medical attendant connected patients remotely to an on-call pediatrician to provide urgent care outside of regular office hours
3. Nurse case managers, who contacted families of children with frequent ED use to identify factors contributing to ED use and potential solutions
4. An after-hours clinic on the main UHC campus
5. Community outreach, such as billboards and bus advertisements, to promote the use of ED alternatives

UHC planned the first two services and added the others to meet its ED avoidance goal. UHC had mixed results in implementing these services. UHC implemented the telephone triage service as planned and reported significant use (more than 90,000 calls during the award). UHC added the ED nurse case manager service in the second year of the award, but did not report any implementation measures for the service. UHC faced significant delays with the telemedicine hubs and reported limited use after implementation in the second and third years of the award (651 visits). In response to this, UHC added the after-hours urgent care clinic in the third year, which saw more use (1,758 visits).

Clinicians' perceptions of intervention effects on the care they provide. In surveys, clinicians reported being engaged in activities consistent with the goals of the program and said they perceived positive impacts of the program on multiple dimensions of care.

Impacts on patients' outcomes. In the service use domain, we found no substantively large effects on outpatient ED visit rates. This means that we saw no evidence of a meaningful decline in ED visits, despite good statistical power to detect such a decline if it existed.

We could not draw definitive conclusions about the quality-of-care processes and outcome domains, as assessed through lead screening, well-child visits, and dental caries treatment. The results from our primary analyses suggested that the intervention decreased the likelihood that children received preventive care services, which we did not consider plausible given the intervention's design and implementation. Rather, we expect that these primary results could be biased due to limitations in the comparison group for these particular measures. Specifically, we anticipate (though were unable to test) that the comparison practices were on a different trajectory for these measures during the baseline period, calling into question a key assumption behind difference-in-differences models. That assumption is that baseline differences in outcomes reflect the differences we would expect to see in the intervention period but for the

effects of the intervention itself. In contrast, we were able to verify that baseline trends for outpatient ED visits were similar for the treatment and comparison groups, which is why we have greater confidence in those results.

Conclusions. Our implementation findings suggest UHC was largely successful in implementing the practice facilitation and ED avoidance interventions, though it faced substantial challenges with the telemedicine hubs. We found the program did not reduce outpatient ED visits and that, despite a rigorous evaluation design, we were unable to draw impact conclusions for the practice facilitation component. We draw two types of lessons from this evaluation: lessons for future evaluations and lessons for ED avoidance interventions.

Lessons for future evaluations include the importance of building in robustness checks, including tests of model assumptions (such as testing for parallel trends in the baseline period in a difference-in-differences model), and of using a clear evaluation framework to synthesize impact and implementation findings before drawing conclusions. There are also a number of considerations to improve impact evaluations that rely on a strong comparison group when factors that influence outcomes, such as practice quality-of-care initiatives, might be unobservable or difficult to measure. Some options for improvements include (1) randomly assigning practices to treatment or control status to ensure there are no systematic difference between practices; (2) using longer baseline periods, if possible, to evaluate trends in outcomes; and (3) developing as comprehensive a set of variables as possible for comparison group selection, including variables that might pick up baseline trends in practice quality-of-care initiatives (such as changes in office hours over time) that could affect outcomes.

The findings from the ED avoidance component also provide some lessons. Although the telemedicine hubs were an innovative way to provide an alternative to ED care in areas with frequent ED use, they saw minimal patient volume. A redesign of the hubs or greater outreach would have to occur for patients in the target population to use them enough to be effective in avoiding an ED visit. Second, we found that the enhancements UHC made to its nurse triage service did not generate meaningful changes in the ED visit rate for the treatment group included in the impact evaluation. Practices could consider ways to make more substantial changes to their triage services from the status quo and/or work to ensure that these new services are widely used among the target population to increase prospects of success.

Policymakers, clinicians, payers, and researchers can use findings from this study to help inform and evaluate efforts to engage primary care practices and develop innovative models to improve health care for children in Medicaid.

Summary of intervention and impact results for UHC

Intervention description			
Awardee description		Academic health system, including a children's hospital and pediatric ambulatory care network, in northeastern Ohio	
Award amount ($ millions)		$12.8	
Award extended beyond June 2015?		Yes (9 months)	
Locations		Eight counties in northeastern Ohio, including urban, suburban, and rural areas	
Target population		All children enrolled in Medicaid who received care at the primary care practices engaged in the practice facilitation program	
Intervention component included in impact evaluation[a]		Practice facilitation and ED avoidance: • For the practice facilitation component, practice facilitators visited participating practices weekly to assess practice performance on quality measures through chart audits, shared measure results with providers and staff, and discussed opportunities for improvement. Practices could receive semiannual incentive payments based on thresholds for the quality measures. • The ED avoidance component included enhanced nurse-led telephone triage, telemedicine hubs, nurse case managers for frequent ED users, an after-hours clinic, and outreach activities.	
Metrics of intervention delivered		• Enrolled 32 practices (exceeding its goal of 28) that provided care for more than 71,000 child Medicaid beneficiaries (exceeding its goal of 68,000) • Provided more than 90,000 nurse telephone triages calls, 651 telemedicine hub visits, and 1,758 urgent care clinic visits over the course of the award	
Impact evaluation methods			
Core design		Difference-in-differences model with matched comparison group	
Treatment group	Definition	Medicaid (FFS and managed care) beneficiaries ages birth to 18 years whom we attributed to the practices participating in the practice facilitation component	
	# of beneficiaries during primary test period[b]	35,773 to 45,459	
Comparison group definition		Medicaid (FFS and managed care) beneficiaries ages birth to 18 whom we attributed to matched comparison practices in other counties in Ohio not considered part of the eight-county UHC service area	
Impact results: Service use domain			
Outpatient ED visits (#/1,000 beneficiaries/quarter)		Comparison mean[c]	165.9
		Impact estimate (% difference)	-2.6 (-1.5%)
Outpatient ED visits (#/1,000 high-risk beneficiaries/quarter)		Comparison mean[c]	198.2
		Impact estimate (% difference)	-2.7 (-1.3%)
Combined impact estimate[d]		(-1.4%)	
Impact conclusion[e]		No substantively large effect	

Notes: See the UHC chapter for details on the intervention, impact methods, and impact results. As explained in the chapter, we did not draw impact conclusions in the two other outcome domains: quality-of-care processes or quality-of-care outcomes.

No difference-in-differences estimates were significantly different from zero. We adjusted the *p*-values for the multiple (two) comparisons made within the service use domain.

[a] Three components of UHC's HCIA-funded program were not included in the impact evaluation: (1) hospital readmission prevention, (2) complex care, and (3) behavioral health.

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

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

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

[e] We drew conclusions at the domain level based on the results of prespecified primary tests, secondary tests (robustness checks), and consistency with implementation evidence. For each domain, we could draw one of five conclusions: (1) statistically significant favorable effect (the highest level of evidence), (2) substantively important (but not statistically significant) favorable effect, (3) substantively important (but not statistically significant) unfavorable effect, (4), no substantively large effect, and (5) indeterminate effect. If the results for the primary tests in a domain were not plausible given the implementation evidence or the secondary, corroborating tests, we did not draw conclusions about program impacts in that domain. Section V.A.9 of this chapter describes the decision rules we used to reach each of these possible conclusions.

ED = emergency department; FFS = fee-for-service; HCIA = Health Care Innovation Award; UHC = University Hospitals of Cleveland.

I. INTRODUCTION

This report presents findings from the evaluation of the Health Care Innovation Award (HCIA) received by the University Hospitals of Cleveland (UHC). Consistent with the evaluations for 12 HCIA in primary care redesign, we designed this impact evaluation to identify promising interventions or intervention components that could be scaled or retested as part of a future model.

Section II of this chapter provides an overview of UHC's HCIA-funded intervention, the design of the impact evaluation, and key findings from the impact evaluation. Section III describes the design and implementation of the intervention. In Section IV, we assess the evidence on the extent to which planned changes in clinicians' behavior occurred. Section V describes our methods for, and results and conclusions from, estimating program impacts on patients' outcomes in three domains: (1) quality-of-care processes, (2) quality-of-care outcomes, and (3) service use. Section VI concludes by synthesizing the impact and implementation findings and drawing lessons from the evaluation for future evaluations and for organizations considering interventions similar to those UHC implemented.

As we will describe in this chapter, we designed our evaluation of impacts on patients' outcomes to cover two of the five components that comprised UHC's HCIA-funded program, and we selected these two components for evaluation because we believed we could test them rigorously and fairly. Specifically, we had planned to assess (1) the impacts of one program component on quality-of-care processes and outcomes, and (2) the impacts of a second program component on service use.

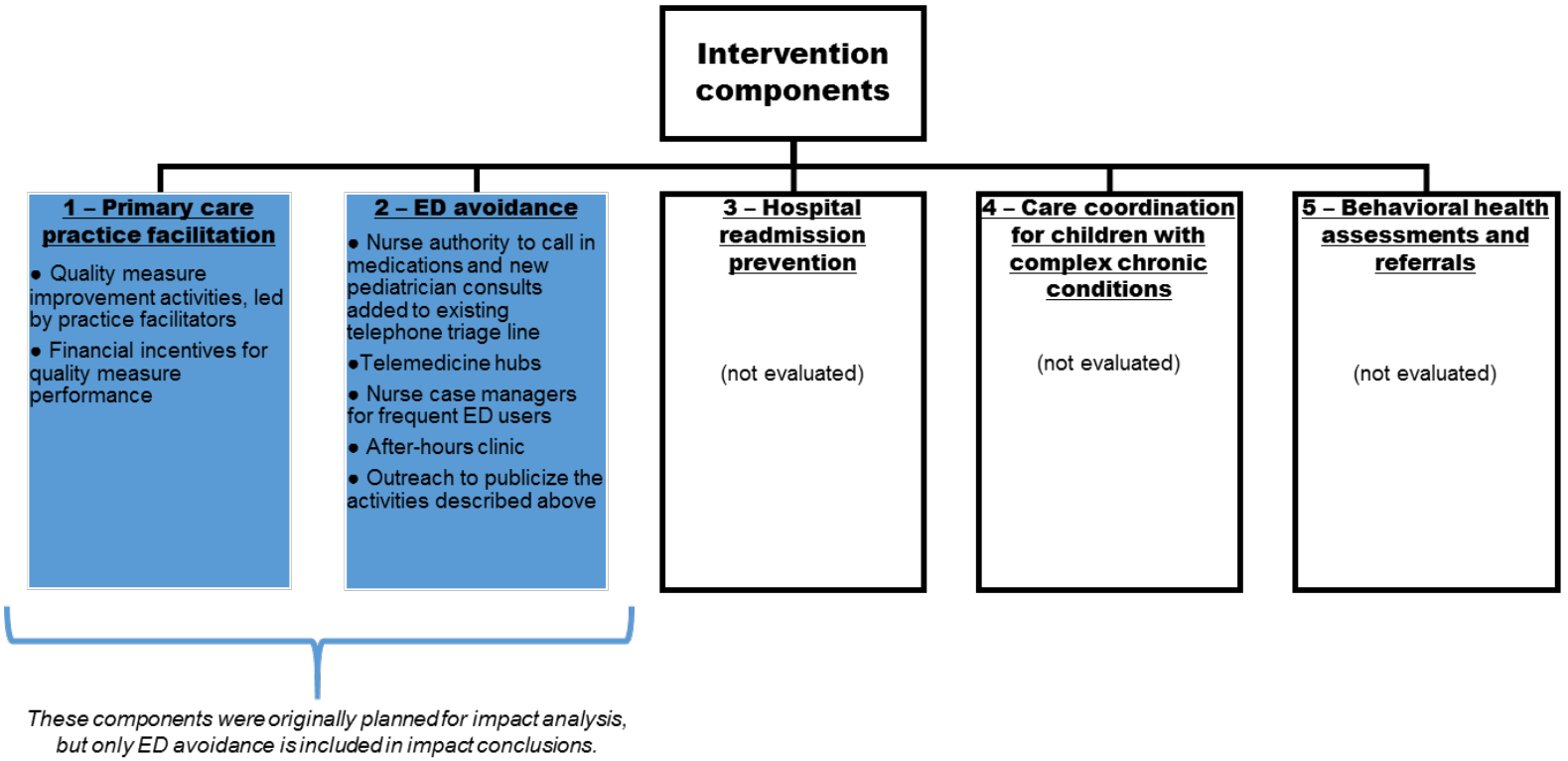
However, due to methodological challenges described in this chapter, we drew conclusions about program impacts only for service use, and not for quality-of-care processes or outcomes. In practice, this also means that our impact estimates apply to only one of the two program components we aimed to evaluate (among the five total components). This chapter nevertheless provides comprehensive findings from our evaluation, including detailed description of program implementation for both program components we aimed to evaluated, and the methods and results from all quantitative analyses—including those for which we did not draw conclusions.

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

A. UHC's HCIA-funded intervention

UHC is an academic medical center, which is part of a large health care system (University Hospitals Health System, Inc.), centered in Cleveland that includes a children's hospital (Rainbow Babies and Children's Hospital) and a network of employed primary care providers. UHC received a three-year, $12.8 million HCIA to transform the delivery of health care for children enrolled in Medicaid in northeastern Ohio. UHC's HCIA program aimed to improve care and health and lower costs for children enrolled in Medicaid. Through multiple clinical and structural program components, UHC worked to build the foundation for a Medicaid accountable care organization (ACO) for children. Figure II.1 and Table II.1 summarize key features of the program.

Figure II.1. UHC intervention components and our evaluation for estimating the intervention's impacts on patients' outcomes



ED = emergency department.

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

Program description	
Award amount	$12,774,935
Award start date	July 2012
Implementation date	January 2013
Award end date	Original: June 2015 With no-cost extension: March 2016
Awardee description	UHC is an academic medical center that is part of a large health system. The health system includes a children's hospital (Rainbow Babies & Children's Hospital) and a pediatric ambulatory care network in northeastern Ohio.
Intervention overview	The UHC intervention aimed to improve care and health and lower costs for children enrolled in Medicaid through multiple program components and build the foundation for a Medicaid accountable care organization.
Intervention components[a]	○ **Primary care practice facilitation.** UHC created a separate legal entity (UHRCC) to contract with primary care practices for care improvement activities and a linked financial incentive program. Practice facilitators visited practices in the program network to assess performance on quality measures through chart audits, shared measure results with providers and staff, and discussed opportunities for improvement. Practices could receive semiannual incentive payments based on thresholds for the quality measures. Over the course of the award, 32 practices participated in the practice facilitation component. ○ **ED avoidance.** UHC undertook several activities to limit unnecessary ED use: • **Enhanced telephone triage.** UHC enhanced an existing after-hours nurse-led telephone triage service by including (1) consultation between the caregiver and a physician for all children triaged by the nurse to the ED and (2) allowing a nurse to call in certain medications for children with low-acuity conditions per a physician's standing order. • **Telemedicine hubs.** UHC implemented telemedicine units in a community center and a storefront location to provide urgent care outside regular office hours; an on-site medical attendant checked families in and then connected patients and their caregivers to an on-call pediatrician through broadband videoconferencing and connected examination tools. • **Nurse case managers.** Case managers contacted families of children with frequent ED use within two days of a child's ED visit to identify factors contributing to frequent ED use, barriers to care, and potential solutions. • **After-hours clinic.** Nurse practitioners provided after-hours care at the main UHC campus. • **Outreach activities.** UHC used billboards, bus advertisements, automated interactive phone calls, and a group of community health workers to engage families about where and when to seek care for their children. ○ **Hospital readmission prevention.** Hospital unit staff developed standardized discharge processes, and a facilitator reviewed patients' charts to determine the units' performance on measures of discharge processes. The facilitator then provided feedback through a unit scorecard and brainstormed with unit leadership about quality improvement. ○ **Complex care.** A multidisciplinary team at UHC conducted a clinical evaluation of medically complex patients and created a standardized portable care plan for each. Intervention staff then provided ongoing care coordination with a child's family and existing providers. ○ **Behavioral health.** Specialized behavioral health social workers managed referrals of children by participating primary care providers to community behavioral health services, provided telephone consultations to clinicians at participating practices, and, when requested, provided initial behavioral health evaluations for children at participating primary care practices. In addition, UHC added behavioral health crisis intervention social workers to its pediatric ED.

Table II.1 *(continued)*

Target population	**Target populations varied by component.**
	Primary care practice facilitation. **All children who received care at the participating primary care practices**
	• **ED avoidance:**
	  ○ **Enhanced telephone triage.** Children who received care at the participating primary care practices
	  ○ **Telemedicine hubs, after-hours clinic, and outreach.** Children living in neighborhoods with high rates of ED use
	  ○ **Nurse case managers.** Children identified by UHC or Medicaid MCOs with frequent ED use
	• **Hospital readmission prevention.** Children hospitalized in medical and surgical units at UHC, except those from hematology/oncology or intensive care units
	• **Complex care.** Children who received care from UHC and (1) had a significant neurocognitive impairment, (2) had three or more body systems impaired, (3) were technology-dependent, or (4) required caregiver assistance with activities of daily living
	• **Behavioral health.** Children who received care at the program's network of primary care practices or in the UHC pediatric ED for behavioral health needs

Target impacts on patients' outcomes	• 75 percent of primary care providers meet quality targets • 15 percent reduction in avoidable ED visits • 2.5 percent reduction in total costs of Medicaid-enrolled children

Workforce development	Hired a total of 54.1 new FTE positions over the course of the award; staffing specific to components covered by the impact evaluation included:[b]
	1. **Primary care practice facilitation.** 5 practice facilitators (4.0 to 8.0 FTE)
	2. **ED avoidance.**
	  ○ **Enhanced telephone triage.** 2 intake specialists (1.80 FTE), 5 nurses (3.60 FTE)
	  ○ **Telemedicine hubs.** 3 medical attendants (0.73 FTE), on-call physicians (1.25 FTE)
	  ○ **Nurse case managers.** 2 registered nurses (2.0 FTE)
	  ○ **After-hours clinic**. 2 nurse practitioners (0.9 FTE), 2 licensed practical nurses (0.9 FTE), 2 receptionists (0.9 FTE)
	  ○ **Outreach.** 1 coordinator (1.0 FTE), 1 physician (0.25 FTE).

Location	UHC is located in Cleveland, Ohio. Participating practices are in eight counties in northeastern Ohio, including urban, suburban, and rural areas.
	Impact evaluation
Core design	Difference-in-differences with matched comparison group
Treatment group	Medicaid (FFS and managed care) beneficiaries ages birth to 18 years whom we attributed to the practices participating in the UHC practice facilitation component
Comparison group	Medicaid (FFS and managed care) beneficiaries ages birth to 18 whom we attributed to matched comparison practices in other counties in Ohio not considered part of the eight-county UHC service area
Intervention component(s) included in impact evaluation	The impact evaluation includes the practice facilitation and ED avoidance program components. Some children in our treatment group were also exposed to the other three components, but these components were not expected to have a measurable impact on the outcomes in this study.
Extent to which the treatment group reflects UHC's target population (for the component[s] evaluated)	**Practice facilitation: High.** The treatment group was defined as children attributed to practices participating in the practice facilitation program component. **ED avoidance: Medium.** The treatment group included all children exposed to the telephone triage invention, but only some children were exposed to the telemedicine hubs, after-hours clinic, and outreach.
Study outcomes, by domain	1. **Quality-of-care processes.** Lead screening for beneficiaries 2 years old; well-child visits for beneficiaries 3 to 6 years old 2. **Quality-of-care outcomes.** Dental caries treatment for beneficiaries 3 to 5 years old 3. **Service use:** Outpatient ED visits

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

Table II.1 *(continued)*

[a] This report focuses on the primary care practice facilitation and ED avoidance components because these were the components included in the impact evaluation.

[b] The remaining FTEs hired under the HCIA included staff for other components and administrative staff.

ED = emergency department; FFS = fee-for-service; FTE = full-time equivalent; HCIA = Health Care Innovation Award; MCO = managed care organization; UHC = University Hospitals of Cleveland; UHRCC = University Hospitals Rainbow Care Connection.

UHC's HCIA-funded program included five clinical components:

1. **Primary care practice facilitation.** UHC created a separate legal entity, University Hospitals Rainbow Care Connection (UHRCC), to support quality measurement and improvement in primary care practices—comprising either clinicians employed by an affiliate within UHC's health system or independent clinicians—and to award financial incentives to providers that met certain quality targets for delivery of preventive care services.
2. **Emergency department avoidance.** UHC enhanced or introduced several services to reduce avoidable emergency department (ED) use, including an after-hours telephone line for nurses to offer advice for urgent concerns, telemedicine hubs, nurse case managers for frequent ED users, and an after-hours clinic.
3. **Hospital readmission prevention**. Hospital staff developed standardized discharge processes, and a facilitator reviewed patients' charts to determine the units' performance on quality-of-care process measures, provided feedback through a unit scorecard, and brainstormed with unit leadership about quality improvement.
4. **Clinical care and care coordination for children with complex chronic conditions.** A multidisciplinary team at UHC identified medically complex children, conducted comprehensive clinical evaluations, created a standardized portable care plan, and provided ongoing care coordination with the children's families and existing providers.
5. **Behavioral health assessment and referral services.** Specialized behavioral health social workers managed referrals of children by participating primary care providers to community behavioral health services, provided telephone consultations to clinicians at participating practices, and, when requested, provided initial behavioral health evaluations for children at participating primary care practices. UHC also added behavioral health crisis intervention social workers to its pediatric ED.

UHC's goals for its program were to (1) have 75 percent of participating pediatricians meet quality targets, (2) reduce avoidable ED visits by 15 percent among Medicaid-enrolled children, and (3) reduce the total cost of care by 2.5 percent for Medicaid-enrolled children.

Broadly, the target population for the UHC program included children from birth to 18 years who were enrolled in Medicaid, although some program components focused on a subset of this population (such as children with complex chronic conditions). Although the program aimed to improve care for Medicaid-enrolled children, program components were open to all children regardless of payer. The practice facilitation component and most of the behavioral health

component focused on children who received care in the participating primary care practices, but the other components were available to children across the region, especially those who had received or were likely to receive care at Rainbow Babies and Children's Hospital.

B. Overview of impact evaluation: design and findings

1. Design

Our impact evaluation focused on two of the five intervention components of the UHC program: (1) primary care practice facilitation and (2) ED avoidance (Figure II.1). We defined the treatment group of our evaluation as Medicaid child beneficiaries (ages birth to 18 years) who were served by the 32 practices participating in UHC's practice facilitation component. We selected the practice facilitation component as the foundation of our evaluation because the awardee considered this to be the core of its HCIA-funded program. In addition to capturing the effects of this component, however, our impact estimates also capture the effects of UHC's other four program components to the extent that they affected Medicaid child beneficiaries at the 32 practices—that is, the members of our evaluation treatment group. In particular, we expect our impact estimates to reflect the effects of the ED avoidance component because the services it provided were available to all children receiving care in practices participating in the practice facilitation component. It is possible the other three program components might also have affected treatment group members, if the children met certain conditions (Figure II.1). However, these components would have touched only a small proportion of the treatment group and, therefore, should have little influence on our impact estimates. The impact design was a difference-in-differences model with a comparison group matched at the practice level. The practices that participated in the primary care practice facilitation component included practices with providers employed by UHC, independent practices, and a federally qualified health center (FQHC). These treatment practices joined the intervention in waves. We matched them to comparison practices in Ohio outside of the eight-county UHC service area and that had similar practices characteristics and characteristics of their pediatric Medicaid populations. We matched to practices outside of the UHC service area because many of the ED avoidance interventions and the hospital readmission prevention component could have spillover effects even for nonparticipating primary care practices, given that these services were available to all children in the region. Both the treatment and comparison groups were defined at the beneficiary level—that is, as children (from birth to 18 years) enrolled in Medicaid fee-for-service (FFS) or managed care and attributed to treatment or comparison practices, respectively.

Using Medicaid data, we estimated impacts on outcomes in three domains: (1) quality-of-care processes, (2) quality-of-care outcomes, and (3) service use. The impact estimates were the differences in outcomes for the treatment and comparison groups during the intervention period minus the average differences in outcomes for these two groups before the intervention began, adjusting for residual differences between the two groups in other observed characteristics.

Before conducting the analysis, we specified primary tests, describing the evidence we would need to conclude that the program was effective, and the awardee and CMMI reviewed these tests. Each test specified a population, outcome, period, expected direction of effect, and threshold that we counted as substantively important. The purpose of these primary tests was to

focus the impact evaluation on prespecified hypotheses that would provide the most robust evidence about program effectiveness. We used the results from the primary tests and robustness checks to draw conclusions about program impacts in each of the three evaluation domains. Because we wanted to identify promising interventions, rather than only those programs with unequivocally demonstrated success, we conducted one-sided statistical tests (that is, testing only for program benefits) and used a threshold for statistical significance of 0.10, which is not as strict as the conventional standard of 0.05.

2. Findings

We had two main findings of our impact evaluation:

First, we found that the program did not reduce outpatient ED visits. The difference-in-differences estimates showed a small favorable effect of -2.6 visits per 1,000 patients per quarter (or -1.5 percent) but this difference was neither substantively large nor statistically significant ($p = 0.27$). The evaluation had good statistical power to detect a substantively large effect if, in fact, the program had one. In interpreting these findings, it is important to note that we assessed the marginal impact of the ED avoidance interventions that UHC implemented using HCIA funding and not the combined effects of the collective efforts that UHC took (with funding from various sources) to reduce outpatient ED visits. For example, we estimated the impacts of HCIA-funded *enhancements* to the after-hours telephone triage line for nurses to offer advice for urgent concerns—not the impact of the telephone line overall. It is possible that, during the baseline period, the triage line already helped to reduce ED visits but that the specific HCIA-funded enhancements did not further reduce ED visit rates during the intervention period. Further, although all four of the clinical services in the ED avoidance program component were open to the treatment group, three of the four services—the telemedicine hubs, nurse case managers for frequent ED users, and after-hours clinic—were likely practically available to only a subset of the children in the treatment group who lived close enough to the location of the services to use them. For this reason, they might not have had a large impact on the treatment group included in the evaluation.

Second, despite using a rigorous design, we were unable to draw definitive conclusions about the impacts of the practice facilitation component. The primary test results suggested that the intervention decreased the likelihood that children received select preventive care services, which was counterintuitive and which we did not consider plausible given the intervention's design and implementation. We expect that limitations in the comparison group for these particular measures could have biased these primary results. Specifically, we anticipate (though were unable to test) that the trajectories of the treatment and comparison practices differed for these measures during the baseline period, calling into a question a key assumption behind difference-in-differences models. That assumption is that baseline differences in outcomes reflect the differences we would expect to see in the intervention period but for the effects of the intervention itself. We were unable to test this assumption because the quality-of-care measures required non-overlapping measurement periods at one- or two-year-long intervals, and we were constrained by data usability for time periods before mid-2011. Therefore, we were able to measure the quality-of-care measures only at one point in the baseline period, making it impossible to evaluate practices' trajectories of these measures before the intervention began. In

contrast, we were able to verify that baseline trends for outpatient ED visits were similar between the treatment and comparison groups because we could measure outpatient ED visits in every quarter of the baseline period. For this reason, we have greater confidence in those results.

In short, our impact evaluation ultimately assessed impacts for one of UHC's five intervention components (Figure II.1). This was a result of our decision to focus the evaluation on two of the five components for which we believed a rigorous impact design was possible, and the challenges in drawing definitive conclusions for the practice facilitation component.

III. PROGRAM IMPLEMENTATION

In this section, we first describe in detail the primary care practice facilitation and ED avoidance components of UHC's HCIA-funded program, including how they evolved over time. Second, we assess the evidence on the extent to which UHC implemented these two program components as planned based on measures of program enrollment, service delivery, staffing, training, and timeliness. Third, we summarize the facilitators and barriers associated with implementation effectiveness.

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

A. Program design and adaptation

1. Target population and enrollment by program component

a. Primary care practice facilitation

The primary care practice facilitation component was delivered at the practice-level, so UHC, through UHRCC, targeted pediatric primary care practices in its eight-county service area in northeastern Ohio, with a goal of recruiting enough practices to join the intervention to reach 68,000 Medicaid-enrolled children. Ultimately, two practices in Summit County, outside of the eight-county service area, also joined the program. Participating practices could either comprise clinicians employed by an affiliate within UHC's health system or independently, but the program did not recruit practices directly affiliated with other health systems in the region. UHRCC targeted practices known to provide care for children enrolled in Medicaid—especially those with Medicaid beneficiaries comprising at least 5 percent of the practice's patient population—and recruited practices by telephone, email, and in-person contact. All interested practices were enrolled and included a mix of practices with clinicians employed by a UHC affiliate, independent practices, and an FQHC. UHRCC continued to enroll new practices throughout the award. For the few practices that declined to participate, UHC described their primary reason for not -participating as feeling overburdened with other initiatives, such as implementing an electronic health record (EHR). Practices had to agree to a contract that included participating in practice facilitation, goals for quality metrics, and the opportunity for

incentive payments based on meeting quality goals. The target population for the intervention was all children who received care at a participating practice, but especially those covered by Medicaid. Patients did not have to opt in or consent to the program.

b. ED avoidance

UHC aimed to reduce unnecessary ED use among children served by the participating primary care practices. Several interventions (described in more detail in Section III.A.2) comprised the ED avoidance component, and the interventions differed slightly in their target populations. The enhanced telephone triage service was available to caregivers of any child served by the participating practices. The telemedicine hubs were located in neighborhoods with high ED visit rates, and an after-hours clinic was located at the UHC main campus. The telemedicine hubs and the after-hours clinic provided services to any child who visited them. Finally, the nurse case manager service targeted all children who, according to UHC internal data, had three or more ED visits in 12 months or had an ED visit in the first 6 months of life. By using the services, patients consented to participate to use the telephone triage service, the telemedicine hubs, and the after-hours clinic. Participants did have to explicitly opt in to the case manager service, however. Participants were enrolled if the case managers contacted a caregiver and the caregiver agreed to discuss the child's care and have follow-up calls.

2. Intervention components

a. Practice facilitation and financial incentives

The practice facilitation component was based on a prior randomized controlled trial (RCT) conducted by UHC that used practice facilitation to promote the delivery of pediatric preventive services in primary care, specifically lead screening, dental fluoride varnish application, and obesity screening and counseling (Meropol et al. 2014).

UHC hired practice facilitators who visited participating practices weekly, performed a small number of chart audits to assess performance on quality measures, shared measure results with providers and other practice staff, and discussed opportunities for improvement. Facilitators provided practice sites with educational materials (such as patient education handouts and posters for exam rooms), clinician decision tools (such as a parent-reported asthma control assessment), and organizational tools (such as binders with drug formularies and desktop document organizers). The facilitators then visited participating practices monthly to perform audits on larger numbers of charts than done at the weekly visits; this was done to assess progress toward quality targets and eligibility for incentive payments. Chart reviews and improvement efforts included all children in the relevant age group for the quality measure receiving care in the practice, regardless of insurance type. Table III.1 lists and defines the quality measures by year of introduction in the program.

At the beginning of the award, UHC established an advisory council composed of providers from the program's participating practices. Throughout the award, the council provided feedback to UHC on the clinical components of the program, particularly practice facilitation and the quality improvement goals. Thus, all changes to the program were at least informed by, and at times directly suggested by, the advisory council. As planned, the program increased the number

of measures over time (3 in the first year, 7 in the second year, and 11 in the third year) and, in the third year of the award, adapted the frequency of facilitator visits to every other week for practices that were performing well. During the no-cost extension period (June 2015 to March 2016), UHC added an additional 4 measures, combined 2 of the previous well-child care measures, and retired 3 of the measures used earlier in the award period.

The program's participation agreement with practices included incentive payments for reaching specified thresholds for each measure. There were no penalties. The program made semiannual incentive payments to practices that met measure thresholds based on an independent audit of practices' quality measure results by UHC's health system's internal audit department. As an additional incentive in all program years, UHC offered credit for participating physicians for the quality improvement activities required to maintain board certification.

Table III.1. Quality measures introduced in the initial award period (January 2013 to June 2015) for primary care practices participating in the UHC HCIA program

		Program year			
Quality measure	Definition	Year 1	Year 2	Year 3	Year 4
Fluoride varnish application	Apply varnish every 6 months after tooth eruption (age 12 to 35 months)	X	X	X	X
Obesity screening	Calculate BMI and percentile, diagnose weight, counsel if overweight or obese	X	X	X	X
Lead screening	Order appropriate testing at 12 and 24 months	X	X	X	R
Asthma management	Assess control of asthma using a standardized assessment		X	X	X
Drug formulary	Prescribed following drug formulary		X	X	X
Appropriate URI treatment	Do not use antibiotics for URI diagnosis		X	X	R
Well-child care ages 3 to 6	Schedule patients ages 3 to 6 years for follow-up well visit at time of sick visit, if needed		X	X	C
ADHD management	Document follow-up care for patients ages 6 to 17 years with a diagnosis of ADHD and prescription for a stimulant medication			X	X
Adolescent vaccine (Tdap, meningococcal, HPV)	Document age-appropriate doses of meningococcal, Tdap, and HPV vaccines			X	X
Pharyngitis	Provide appropriate evaluation and treatment for patients ages 2 to 18 years diagnosed with pharyngitis			X	R
Well-child care ages 13 to 17	Schedule patients ages 13 to 17 years for follow-up well visit at time of sick visit, if needed			X	C

Table III.1 *(continued)*

Sources: Interviews from second site visit, March 2015; document review, March 2015; and review of final progress report.

ADHD = attention deficit hyperactivity disorder; BMI = body mass index; C = combined (the two well-child care metrics were combined in Year 4 and were measured together); HCIA = Health Care Innovation Award; HPV = human papilloma virus; Tdap = tetanus, diphtheria, and acellular pertussis; UHC = University Hospitals of Cleveland; URI = upper respiratory infection.

R = retired in mid-2015 due to high levels of attainment by all practices, allowing four new metrics to be added in Year 4 (the no-cost extension period). The four new measures included promoting early childhood literacy, assessing safe sleep habits for infants, maternal depression, and early childhood education. These four measures were not evaluated.

b. ED avoidance

The ED avoidance component included several related services to reduce preventable ED use. From the beginning of the program, UHC offered an enhanced after-hours nurse telephone triage service to participating practices. The intervention built on an existing service by requiring the telephone triage nurses to consult with an on-call physician for all children triaged to the ED and allowing nurses to call in certain medications for children with low-acuity conditions per a physician's standing order.

In the second year of the program, UHC also began operating two telemedicine hubs in two neighborhoods with high rates of ED use by Medicaid-enrolled children. The telemedicine hubs, or kiosks, included high resolution video conferencing and examination tools that transmitted audio and video, such as a stethoscope and an otoscope for examining a child's ears. An in-person medical attendant and an on-call physician available through a high-speed Internet connection staffed the hubs. The medical attendant checked patients in, obtained vital signs, and assisted the on-call physician by operating the examination tools and performing point-of-care diagnostic tests, such as urinalysis and rapid streptococcal infection testing. The on-call physician provided evaluation and care from off-site using the telemedicine tools.[1]

Also in the second year of the program, UHC used unspent funding from the first award year to hire nurse case managers to work with families of children with frequent ED use. This intervention was not part of the program originally proposed to CMMI and was intended to provide additional support to meet UHC's goals of reducing avoidable ED visits. The UHC pediatric ED referred children with four or more ED visits in the prior year or any avoidable ED visit in the first six months of life to the program. The case managers contacted families within two days of their child's most recent ED visit to identify factors contributing to frequent ED use, barriers to care, and potential solutions and followed up with families at regular intervals for up to three months after the initial call.

[1] UHC also worked with a large retail pharmacy chain toward the end of the no-cost extension (April to June 2015) to begin providing telemedicine urgent care through other hubs owned by the chain and located in individual pharmacy locations. This ended when the telemedicine hub company ceased operations.

In the third year of the program, UHC again used carry-over funding to support its goal for reducing ED use by also opening an after-hours clinic on its main hospital campus to provide an option for in-person urgent care. A nurse practitioner staffed the clinic.

Throughout the award, UHC promoted these alternatives for ED care through outreach, including billboards, bus advertisements, automated interactive calls, and a group of community health workers hired using HCIA funds to engage with families door-to-door and at community events about where and when to seek care for their children.

3. Intervention staff and workforce development

Each component of the UHC program required creating new frontline staff positions with new responsibilities (Table III.2). These staff received a general orientation provided to all new employees in UHC's organization, an orientation to the goals and components of the UHC HCIA program, and, as needed, training for their program-specific roles through shadowing more seasoned staff.

For the practice facilitation component, UHC hired practice facilitators who had master's-level training and/or other significant experience with clinical research or quality improvement. In addition, facilitators received about one month of training, shadowing other facilitators who had worked in the preceding RCT of practice facilitation, or in the HCIA. About halfway through the award period, a practice facilitator with significant self-training in quality improvement was promoted to manager of the practice facilitation component and underwent formal quality improvement training through a distance learning program from another children's hospital. At first, there was also a separate role for a chart reviewer who performed monthly reviews to measure whether practices met thresholds for incentive payments. This role merged with the practice facilitator's role in mid-2014, and the separate position was eliminated.

At the administrative level, a medical director designed intervention activities, recruited and managed relationships with participating practices, provided advice and feedback to staff, and served as the public face of the program. At first, a senior program director managed the day-to-day operations of the program and staff, established staff and managerial roles, and created program databases for and information feedback mechanisms to staff. After the program administrative director left the program in June 2014, the medical director, the manager of the practice facilitation and ED avoidance components, and the manager of the telephone triage program shared these responsibilities. The manager of the practice facilitation component had master's-level training and, as noted, had previously served as a practice facilitator. The director of the UHC health system-affiliated Medicare and commercial ACOs also assumed some of the responsibilities for managing contracts with participating practices, managed care contracts, and population health data. A data analyst managed the program databases and data reports.

In addition to training its program staff, UHC also provided formal educational events for providers and staff in the program's participating primary care practices. During the award, UHC held seven events that qualified for continuing medical education credits. The events drew 90 to 157 attendees and covered topics such as pediatric headaches, asthma care, cost-conscious drug prescribing, a toolkit for attention-deficit/hyperactivity disorder, and obesity care.

Table III.2. UHC frontline staff qualifications and adaptations

Intervention component	Staff member	Qualifications	Adaptations during the award
Primary care practice facilitation	Practice facilitator	Master's degree in public health or epidemiology or significant experience in research or quality improvement	The practice facilitator role and a chart reviewer role merged in mid-2014 following cross-training of staff to perform both roles.
ED avoidance	Telephone triage nurse	Registered nurse	None noted
	Telehealth attendants	Medical assistant or emergency medicine technician	None noted
	Telehealth physicians	Pediatrician	None noted
	ED nurse case managers	Registered nurse	None noted
	Outreach activities coordinator	Experience in community relations	None noted
	After-hours clinic clinician	Nurse practitioner	None noted

Sources: Interviews and document review.

ED = emergency department; UHC = University Hospitals of Cleveland.

B. Implementation effectiveness

1. Program enrollment

UHC set specific enrollment goals only for the practice facilitation component. UHC enrolled 32 practices (representing 49 practice sites) including 192 clinicians over the course of the award, exceeding its goal of 28 practices. This total included several practices that sought to join the program during the second and third years of the award after declining to participate in the first year. Of the 32 practices that joined, 13 practices (or 20 practice sites) participated in the prior RCT pilot program. UHC's health system employed 60 percent of the clinicians participating in the program through physician practice subsidiaries, and the remaining 40 percent of the clinicians in the program were independent. The treatment practices provided care for more than 71,000 children in Medicaid, exceeding the goal of 68,000 children. The practices also provided care for about 140,000 privately insured children, representing a potential spillover population for the program's interventions well beyond the target of children in Medicaid.

2. Service-related measures

a. Practice facilitation and incentives

The awardee did not formally report measures of service delivery for the practice facilitators but, during site visit interviews, program and practice staff described success in reaching the goal of weekly visits and feedback to practices by the facilitators. In its final report, the awardee described reaching its goal of completing chart audits and making incentive payments to practices every six months throughout the award. UHC reported that in the final program quarter, 30 of 31 participating practices received an incentive payment (two practices merged in July 2015, reducing the overall number from 32 to 31). UHC considered the thresholds and the values

of the incentive payments to be proprietary, but reported paying nearly $1.9 million in incentive payments in total over the course of award, ranging from $4,000 to $32,000 per full-time equivalent (FTE) physician, with a mean total payment of about $19,000 per FTE physician over the 3.75 years of the initial award and extension period.

b. ED avoidance

UHC did not set specific goals for process or service delivery measures related to ED avoidance. The measures the awardee reported for its ED alternatives showed varied uptake of specific interventions. For the approximately 71,000 Medicaid-enrolled and 140,000 privately insured children in the program's participating practices, UHC reported a significant call volume for enhanced telephone triage. Specifically, UHC provided more than 30,000 calls per award year. About 13 percent of calls were referred to the on-call physician by the nurse for potentially needing referral to the ED, and 4 percent of calls were ultimately referred to the ED by the on-call physician. Another 3 percent of calls were triaged to nurses who, based on authority from physicians' standing orders, called in medications for children with low-acuity conditions.

The UHC telemedicine hubs had relatively few users—only 651 child visits after beginning operation in the second year of the award—whereas the after-hours urgent care clinic had 1,758 visits from the 10th quarter of the award (October to December 2014) through the end of no-cost extension (March 2016). UHC did not report measures for the nurse case manager services targeting frequent ED users.

3. Staffing measures

UHC met its overall goal for new hires into the program and staffed the specific components according to its plans (Table III.3). Cumulatively, it hired 54.1 new FTEs by the end of the 12th program quarter (June 2015) and 55.6 FTEs by the end of the 15th program quarter (March 2016), compared with a goal of 53.1 FTEs. When UHC lost staff, it either hired new staff or redistributed responsibilities among existing staff. Notable changes in staffing occurred in the practice facilitation program component, as described previously. The practice facilitation component began with 8.0 full-time staff members, which declined to 4.0 by the end of the award after consolidating the practice facilitator and chart reviewer roles and other staff attrition.

Table III.3. UHC HCIA staffing

Intervention component	Role	Average number of positions during award	Number of FTEs (average unless otherwise noted)
Primary care practice facilitation	Practice facilitators	5 positions	Started with 8.0 and decreased to 4.0 by award's end
ED avoidance	Telephone triage	2 intake specialists 5 nurses	1.8 intake specialists 3.6 nurses
	Telehealth attendants	3 positions	0.7 attendants
	Telehealth physicians	Used as needed	Budgeted for 1.25 FTE
	ED nurse case managers	2 positions	2
	Outreach activities	1 coordinator 1 lead MD	1 coordinator 0.25 lead MD
	After-hours clinic	2 LPNs 2 nurse practitioners 2 receptionists	0.9 LPNs 0.9 nurse practitioners 0.9 receptionists

Source: Correspondence with awardee.

ED = emergency department; FTE = full-time equivalent; HCIA = Health Care Innovation Award; MD = doctor of medicine; LPN = licensed practical nurse; UHC = University Hospitals of Cleveland.

4. HCIA-funded training

In interviews, UHC leaders described primarily informal training for program staff and more formal educational events for clinicians in participating practices. They noted that the interventions in the program were new, so their approach was to hire motivated staff, orient them to the goals of the program, and allow them to define their roles based on the goals. The main exception to this was the practice facilitation component, which was based on protocols from a prior randomized trial. In interviews, practice facilitation staff described an informal training process based primarily on shadowing staff with more experience.

To better understand the experiences of UHC staff with training, we surveyed staff in all components from January to March 2015, about 24 months after the program began; 49 staff members (83 percent) responded. In this section, we focus on responses from staff members whom we could clearly identify as associated with the program components included in our impact evaluation, including the practice facilitation component (practice-tailored facilitators, quality staff, and manager) and nurses associated with the telephone triage service of the ED avoidance component. We were unable to differentiate responses for staff associated with other interventions in the ED avoidance component (care management, telemedicine hubs, and outreach). As a result, the survey findings we present are from about 45 percent of the respondents (7 for practice facilitation and 15 for telephone triage), and we present results qualitatively due to small numbers.

Among the 22 respondents for this analysis, nearly all practice facilitation respondents and about half of telephone triage nurses reported receiving informal training. UHC delivered this

informal training primarily through on-on-one work with a supervisor or other staff, written materials, staff meetings, and field training. Few practice facilitators or telephone triage nurses reported receiving formal training. The types of formal training the staff in both components reported reflected general orientation to the program, including new-hire training from UHC and concepts in quality improvement. Of those who reported receiving formal training, all rated it as good to excellent.

The 15 telephone triage nurses who responded to the survey described spending their time in ways consistent with the goals of the program, including calling participants' families to discuss symptoms and medications (two hours or more per day), educating participants' families on managing their care (two hours or more per day), and executing standing orders for medications (one to two hours per day). Questions on daily activities focused on clinical tasks, so were irrelevant to practice facilitation staff.

5. Program timeline

UHC received HCIA funding in July 2012 and began implementing intervention services in January 2013 (Table III.4). The practice facilitation component began in February 2013, and all the practices participating at that time completed a run-in phase of orientation to the intervention and baseline metric measurement by May 2013. The program continued on steady pace with semiannual payment of practice incentives and annual changes to quality metrics, despite the intermittent addition of new practices to the program. In the ED avoidance component, enhanced telephone triage services began in January 2013 and continued throughout the award, but the planned telemedicine services were implemented later (first hub: October to December 2013; second hub: January to March 2015). UHC implemented ED nurse case management (October to December 2013) and the after-hours urgent care clinic (October to December 2014) with unused funds from prior years.

Table III.4. UHC HCIA timeline

Quarter and date	Activity
Quarter 1: July 2012	UHC's HCIA funding began.
Quarters 1 & 2: July to December 2012	Infrastructure development began, including hiring and training staff and contracting with practices.
Quarter 3: January to March 2013	Practice facilitation and enhanced telephone triage services began, serving a program network of 22 pediatric primary care practices. Facilitators began working in practices on February 1, 2013, using three metrics.
Quarter 4: April to June 2013	Practice facilitation continued rolling out to practices; all initial practice sites completed the run-in phase by the end of May; and the first round of incentive payments for participating practices occurred.
Quarter 5: July to September 2013	The program expanded to 28 practices; it increased the number of quality metrics to 7.
Quarter 6: October to December 2013	The first telemedicine hub and ED nurse case management began operations. UHC made the second round of incentive payments for participating practices.
Quarter 7: January to March 2014	No significant events reported.
Quarter 8: April to June 2014	The program team restructured and combined the roles of facilitator and chart reviewer; UHC made the third round of incentive payments for participating practices.
Quarter 9: July to September 2014	The program increased to 32 practices. The number of quality metrics increased to 11.
Quarter 10: October to December 2014	The after-hours urgent care clinic opened on UHC's campus. UHC made the fourth round of incentive payments for participating practices.
Quarter 11: January to March 2015	The second telemedicine hub began operations.
Quarter 12: April to June 2015	UHC began telemedicine care through hubs owned by and located in retail pharmacy locations. UHC made the fifth round of incentive payments for participating practices.
Quarter 13: July to September 2015 (no-cost extension)	The program retired 3 metrics with high levels of achievement in all practices, added 4 new metrics, and combined the two well–child care metrics (Table III.1).
Quarter 14: October to December 2015 (no-cost extension)	No significant events reported.
Quarter 15: January to March 2016 (no-cost extension)	UHC made the sixth round of incentive payments for participating practices (30 of 31 practices received an incentive payment).

Sources: Interviews and document review.

ED = emergency department; HCIA = Health Care Innovation Award; UHC = University Hospitals of Cleveland.

C. Summary of facilitators of and barriers to implementation

There were many facilitators of and barriers to implementation of UHC's HCIA-funded intervention, and we described those factors in detail in the second annual report (Zickafoose et al. 2015). Here we summarize key facilitators and barriers most related to the program components included in this study, along with any new information since the second annual report related to those facilitators or barriers (Table III.5).

Table III.5. Summary of key facilitators of and barriers to implementing UHC's HCIA-funded initiative

Item	Description based on findings in the second annual report	Additional supporting data not available in the second annual report, if applicable
	Facilitators	
Adaptability of the program to meet patients' and providers' needs	Multiple leaders and staff noted that the program aimed to be innovative and that changes to the operational plan were expected and necessary for effective implementation. Although the core components of the program remained unchanged, administrators added subcomponents, such as adding nurse case managers to decrease unnecessary ED visits. The program also made changes to improve efficiency, such as merging the practice facilitator and chart reviewer roles and decreasing the amount of detail in quality measure data collected for the practice facilitation component compared with the original research-based protocols.	No new data
Self-monitoring and quality improvement	Program administrators and staff established a data infrastructure for the program that facilitated implementation by enabling them to track workflows, measure intermediate outcomes, and begin to assess proxies for core outcomes, such as costs of care. The practice facilitation component was built on quality improvement principles with monitoring and feedback to practices. Staff in the ED avoidance component used internal hospital and Medicaid MCO data to identify children with frequent ED visits and identify neighborhoods with high rates of ED use to determine where to place the telemedicine hubs. Within the program, staff noted that program administrators promoted a culture of continuous quality improvement in which staff described frequently looking for ways to improve their work through the use of new information and small changes.	No new data
Stakeholder engagement	Program leaders and staff engaged a broad group of stakeholders to support program implementation. The practice facilitation component engaged practice leaders and staff through initial meetings and continued engagement through direct practice facilitation, a quality-based incentive plan, provider advisory group meetings, and continuing medical education events. Staff and providers at all the practices we visited during site visits noted that the practice facilitators were easy to work with, provided valuable information, and were responsive to requests.	Awardee's Quarter 15 report to CMMI: "Provider feedback and satisfaction was an essential facet to the success of the [program network]."

Table III.5 *(continued)*

Item	Description based on findings in the second annual report	Additional supporting data not available in the second annual report, if applicable
Team characteristics	Program staff and leaders described the structure and functioning of teams as a facilitator to implementation within and across components. The program leaders considered the role of teamwork during hiring throughout the award, emphasizing hiring people who would be proactive and work well in teams. Shared roles and responsibilities, ad hoc peer consultation, and formal team meetings all contributed to strong team functioning within components. In spring 2014, nearly all of the program staff (except the staff in the complex care program component) relocated to a single physical location after previously working in separate sites, which they noted promoted collaboration across components and improved understanding of the overall goals of the program.	No new data
	Barriers	
Provider engagement	Although the program exceeded its targeted number of practices to recruit, program administrators noted that some providers were slow to embrace efforts to reduce unnecessary utilization and costs. For example, several providers were reluctant to implement changes that might reduce the number of FFS visits, such as physician standing orders allowing telephone triage nurses to call in medications for children with low-risk acute conditions, because the providers were concerned about a negative financial impact on their practices due to decreased volume of patients with these common conditions. Program administrators tried to address these concerns by educating concerned providers about the health and other ancillary benefits to their patients in this model through the medical advisory council and individual interactions with providers. In addition, ongoing positive interactions with practice facilitators helped providers identify more closely with the goals of the program and engage with the interventions.	No new data
Prior history	Limited prior experience implementing similar interventions among UHC administrators and staff was also a barrier to program implementation. Although the primary care practice facilitation model was built upon a prior randomized trial, the other components of the program were built from the ground up. As a result, staff underestimated, for example, the challenges of identifying and contracting for locations for the telemedicine hubs, and limited acceptance of the telemedicine hubs.	No new data

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

CMMI = Center for Medicare & Medicaid Innovation; ED = emergency department; FFS = fee-for-service; HCIA = Health Care Innovation Award; MCO = managed care organization; UHC = University Hospitals of Cleveland.

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

Based on these findings, we believe that UHC implemented its HCIA-funded program largely as planned. It implemented all interventions in the components in this study on the planned timeline, with the exception of the telemedicine hubs. The program modestly adapted planned intervention services for efficiency and added additional services to assist with meeting its ED avoidance goals, using unspent carry-over funding from the first years of the award.

Although the program faced some initial barriers engaging providers and with limited experience implementing similar innovations in care, interviews with program leaders and staff and participating providers suggested that these barriers were largely overcome through program leaders' and staff's willingness to adapt and continue to engage with stakeholders.

For the practice facilitation component, UHC began delivering services in the first year of the program and exceeded its goals for the number of practices participating in its program network and the number of Medicaid beneficiaries served by the these practices. Further, UHC implemented the practice facilitation model with modest adaptations over time to staffing and the frequency of interactions with practices for efficiency, and made incentive payments at the planned six-month intervals. Based on the limited survey data available among staff in this component, the practice facilitators reported receiving training consistent with UHC's plans.

For the ED avoidance component, UHC implemented the planned telephone triage intervention on time but faced significant delays in implementing the telemedicine hubs. Based on the limited survey data available for staff in this component, the telephone triage nurses reported receiving training and spending their time in ways consistent with UHC's plans. The program did not set goals for service provision, but reported a significant volume of use of the telephone triage service (more than 30,000 telephone triage calls per program year). However, UHC also reported a very low volume (651 users) at the telemedicine hubs after implementation in the second year of the award. Partly in response to this low use of telemedicine hubs, UHC added the after-hours urgent care clinic in the third year of the award and had a substantially higher volume of visits there. In addition, UHC used carry-over funding to add the ED nurse case manager service in the second year of the award, but did not report volume or other implementation measures for the service to allow an assessment of its implementation.

IV. INTERMEDIATE PROGRAM EFFECTS ON CLINICIANS' BEHAVIOR

This section describes the available evidence on the extent to which UHC's program had its intended effects on changing primary care providers' behavior as a way to achieve desired impacts on patients' outcomes. We use data from two rounds of the HCIA Primary Care Redesign Clinician Survey and from UHC's self-monitoring metrics on care processes and providers' satisfaction to assess providers' engagement with the program and changes in behavior.

A. Clinicians' perceptions of program effects on the care they provided

1. Survey methods

We administered a clinician survey in two rounds, 21 to 23 months after program implementation (September to November 2014) and 29 to 31 months after program implementation (May to July 2015). We sent the survey to all clinicians involved in the practice facilitation component. We did not survey the small number of clinicians involved in the ED avoidance services. A total of 88 and 118 clinicians responded to the first and second rounds of the survey, respectively, with response rates of 64 percent in Round 1 and 73 percent in Round 2. Of the respondents in Round 1, 87 percent were physicians and 13 percent were nurse practitioners; these proportions were similar in Round 2.

2. Survey results

In Round 1 of the survey, nearly all (92 percent) of the clinicians we surveyed were very or somewhat familiar with the UHC program. However, this dropped to fewer than half (44 percent) in Round 2. We were unable to identify a clear explanation for this finding, but suspect it might be a consequence of wording in the survey. The first round used UHC's specific name for the program, but the second round referred generically to the "HCIA initiative/CMMI grant." UHC made a strong effort to brand its program over the course of the award and it is possible that, by the second round of the survey in the final year of the program, many clinicians identified the program by its brand rather than the generic name of the "HCIA initiative/CMMI grant." Of those clinicians who reported familiarity with the program, a large majority reported receiving training related to the program (91 percent in Round 1 and 79 percent in Round 2), and these clinicians reported receiving about 10 hours of program-related training.

Most clinicians who responded to the survey had engaged in quality improvement activities in the previous two years (92 percent in Round 1 and 88 percent in Round 2), training on quality improvements and tools (88 percent in Round 1 and 89 percent in Round 2), and at least one clinical audit of care that their patients received (80 percent in Round 1 and 77 percent in Round 2).

More than three-quarters of clinicians who were familiar with the program believed it would have a positive effect on the quality of the care they provided (Table IV.1). About half of clinicians in each round reported they believed the program had a positive impact on their ability to respond to patients' needs in a timely way, safety, the patient-centeredness of care, and equity of care for all patients. The only item for which more than a few clinicians perceived a negative impact was on the efficiency of care (12 percent in Round 1 and 17 percent in Round 2), but substantially larger proportions of clinicians perceived positive or no effects on efficiency.

Table IV.1. Clinicians' perceptions of effects of the UHC program on patients' care

	Percentage reporting positive impact		Percentage reporting negative impact		Percentage reporting no impact		Percentage reporting too soon to tell	
Survey item	Round 1	Round 2	Round 1	Round 2	Round 1	Round 2	Round 1	Round 2
Quality of care	75	81	0	0	12	12	11	6
Ability to respond in a timely way to patients' needs	51	40	1	0	33	44	12	12
Efficiency	33	42	12	17	42	27	11	12
Safety	43	48	0	0	44	40	11	10
Patient-centeredness	51	48	1	0	33	42	11	6
Equity of care for all patients	46	58	2	0	40	33	10	8

Source: HCIA Primary Care Redesign Clinician Survey, 2014 and 2015 (N = 88 for Round 1 and N = 118 for Round 2).

Note: Clinicians were asked about the perceived effect of the program only if they reported being at least somewhat familiar with the program. The question was worded as follows: "Please indicate if you believe the HCIA initiative/CMMI award has had a positive impact, negative impact, or no impact on the following aspects of the care you provide to patients enrolled at this practice location over the last year." The first column of the table shows verbatim response options.

HCIA = Health Care Innovation Award; UHC = University Hospitals of Cleveland.

B. Awardee data on clinicians' behavior

In its final awardee report to CMMI, UHC reported that 98 percent of practices met targets on nine or more of the program's quality metrics in the final year, exceeding the program goal of 75 percent of practices. UHC reported substantial increases in scores on two of the three process measures related to measures in our impact evaluation: fluoride varnish application and ordering lead screening (Table IV.2). In addition, UHC reported that, in its own surveys of providers' satisfaction with the program, 80 percent or more of providers rated their satisfaction as 4 or 5 on a 5-point scale during each of the program years.

Table IV.2. UHC's HCIA program report of primary care practices' performance on quality measures

	Met measure goal[a] (percentage of practices)	
Quality measure	Baseline	Award end
Fluoride varnish application	0	90
Lead screening orders[b]	32	91
Well-child care scheduled, ages 3 to 6[c]	94	90

Source: Awardee's final progress report to CMMI.

[a] UHC considered its measure goals to be a proprietary part of its contracts with practices and so we do not report them here.

[b] Measure represented ordering of lead screening for indicated children, not completion of lead testing as in the claims-based analysis in this study.

[c] Measure represented scheduling of a well-child visit if a child presented for an acute visit and was overdue for a well-child visit, not completion of a well-child visit as in the claims-based analysis in this study.

CMMI = Center for Medicare & Medicaid Innovation; HCIA = Health Care Innovation Award; UHC = University Hospitals of Cleveland.

C. Conclusions about intermediate program effects on clinicians' behavior

Based on available information, the UHC program appears to have had its intended effects on primary care clinicians' delivery of care to children. In our survey, clinicians reported engaging in activities consistent with the goals of the program and perceived positive impacts of the program on multiple dimensions of care, such as quality of care, safety, and equity of care for all patients. UHC's self-reported quality metrics and clinician survey results also support the conclusion that clinicians were engaged in the program and actively improving care.

V. PROGRAM IMPACTS ON PATIENTS' OUTCOMES

In this section of the report, we present the quantitative analysis we conducted to assess impacts of UHC's HCIA program on patients' outcomes in three domains: quality-of-care processes, quality-of-care outcomes, and service use. We first describe the methods for estimating impacts (Section V.A) and then the characteristics of the treatment and comparison practices at baseline (Section V.B). In Sections V.C, V.D. and V.E, we describe the quantitative impact estimates and their plausibility given implementation findings. Finally, in Section V.F, we describe the conclusions we could draw about program impacts.

A. Methods

1. Overview

We estimated program impacts on patients' outcomes as the difference in outcomes for Medicaid children served by 37 treatment practice locations participating in the primary care practice facilitation component (described later in this section) and those served by 109 matched comparison practices, subtracting observed differences in outcomes between these groups during the year before the intervention began, and regression-adjusting for differences in other practice

and beneficiary characteristics that remained after matching. Before conducting the analysis, we prespecified primary tests, describing the evidence we would need to conclude that the program was effective, and UHC and CMMI reviewed these tests. Each test specified a population, outcome, time period, expected direction of effect, and threshold that we count as substantively important. The purpose of these primary tests was to focus the impact evaluation on hypotheses that would provide the most robust impact evidence about program effectiveness. We used the results from the primary and secondary tests (robustness checks) to attempt to draw conclusions about program impacts in each of the three evaluation domains.

2. Medicaid data

We analyzed Ohio Medicaid enrollment and claims data covering the period 2009 to 2015. The Ohio Colleges of Medicine Government Resource Center provided these data, as approved by the Ohio Department of Medicaid.

3. Treatment group definition

The treatment group was defined as children (from birth to 18 years) enrolled in Medicaid (FFS or managed care) whom we attributed to practices participating in UHC's primary care practice facilitation component during the HCIA funding period and/or during a 6- to 12-month baseline period before the intervention began. To define the treatment beneficiaries in both periods, we first identified the providers who worked at participating practices and then attributed beneficiaries to practices based on the beneficiaries' visits with those providers.

To identify providers in all practices (except for the FQHC locations), we purchased data from SK&A, a private vendor of U.S. health care reference information that collects national provider data. We opted to use SK&A to identify the participating providers instead of using information provided directly by UHC so that we could standardize the process to identify providers for the treatment and comparison groups—eliminating a possible source of bias in our evaluation.

We identified the treatment practices' locations in the SK&A data and collapsed the 45 non-FQHC practice sites into 33 sites that we used in the impact analysis. We used the rule that if two or more practice sites had 50 percent or more of their providers working at both (or all) of the sites, we combined the sites into one. This facilitated attributing beneficiaries to practices. For the four FQHC locations that participated in the program, we used the sites' organizational National Provider Identifiers (NPIs) to identify the sites in the Ohio Medicaid claims data. This left us with a total of 37 treatment practice sites used in the impact analysis. For simplicity, we refer to these in the rest of the report as *practices,* although they do not correspond exactly to the 32 participating practices, as defined by UHC.

Because practices joined the primary care practice facilitation component in waves, we organized the practices into four cohorts for the impact evaluation, summarized in Table V.1. For each cohort, we also defined a baseline (pre-intervention) period relative to the cohort's intervention start date.

Having identified the intervention practices, we constructed the treatment group of beneficiaries by attributing children to practices. Further details about the treatment group definition are available in the supplemental material at the end of this chapter.

Table V.1. Baseline and intervention dates for practice cohorts

Cohort	Number of practices	Baseline start[a]	Baseline end	Intervention start	Intervention end
1	23	8/1/2012[b]	1/31/2013	2/1/2013	10/31/2015
2	7	8/1/2012	7/31/2013	8/1/2013	10/31/2015
3	4	1/1/2013	12/31/2013	1/1/2014	12/31/2015
4	3	7/1/2013	6/30/2014	7/1/2014	12/31/2015

Note: The last full quarter for Cohorts 1 and 2 ended on October 1, 2015. The next full quarter to measure outcomes for these cohorts would have ended on January 31, 2016, but our data ran only through December 31, 2015.

[a] We defined the 23 practices that joined by the start of the intervention period in January 2013 as Cohort 1 practices, and set the intervention start date for them to the start date of the practice facilitators—February 1, 2013. We defined the 7 practices that joined the intervention in July, August, or September 2013 as Cohort 2 practices and set the intervention start date for them to August 1, 2013. We defined the 4 practices that joined the intervention in January 2014 as Cohort 3 practices and set the intervention start date for them to January 1, 2014, and we defined the 3 practices that joined the intervention in July 2014 as Cohort 4 practices and set the intervention start date for them practices to July 1, 2014.

[b] Cohort 1's baseline period was only six months due to data limitations that did not enable us to calculate practice- or provider-level measures for the period before August 2011. The earliest baseline month for which we could attribute children based on a one-year service period was August 2012, using data from August 2011 to July 2012.

4. Comparison group definition

The comparison group was defined as children (ages birth to 18 years) enrolled in Medicaid (FFS or managed care) and attributed to matched comparison practices in each of the baseline and intervention quarters. We selected comparison practices located in Ohio but outside the counties where the intervention practices were located. We made this choice out of concern that selecting comparison practices from within the targeted counties could bias impact estimates downward (toward no impacts) because some of the intervention components could affect all practices in the targeted outcome. For example, UHC's community outreach about alternatives to ED visits could potentially reach all patients in the intervention counties. We selected comparison practices that were similar to the treatment practices during the baseline period on factors that can influence patients' outcomes and the decision to participate in the program. See the supplemental material at the end of this chapter for details about identifying the 109 matched comparison practices, and see Section V.B to assess balance between treatment and comparison groups after matching.

5. Constructing outcomes and covariates

a. Outcomes

For each beneficiary, we calculated four outcomes that we grouped into three domains:

1. Quality-of-care processes

a. Lead screening (binary); calculated as whether a child 2 years old (24 to 35 months) had one or more capillary or venous lead blood tests within the previous two years.

b. Well-child care ages 3 to 6 (binary); calculated as whether a child 3 to 6 years old (36 to 72 months) had one or more well-child visits with a primary care provider in the previous year.

2. Quality-of-care outcomes

a. Dental caries treatment (binary); calculated as whether a child 3 to 5 years old (36 to 59 months) received any dental caries treatment in the previous year. We set the age criteria for this outcome to 36 to 59 months of age so that it measured outcomes at the end of the intervention period among the children targeted for the fluoride varnish application at the beginning of the intervention period (children 12 to 35 months of age).

3. Service use

a. Outpatient ED visit rate (number per quarter); outpatient ED visits are defined as ED visits or observational stays that do not end in a hospital admission.

The number of times that we measured these outcomes differed by the measure. For the quality-of-care processes and outcomes, we measured the outcomes just twice: once during the baseline period and once during the intervention period. This is because these measures are defined over one- or two-year periods. Specifically, for the two measures that assess services delivered in the past year (well-child visits and dental caries), we calculated the measures over a full year: that is, once over the 12 months before the start of the intervention, and once over the last four intervention quarters. We calculated the lead screening measure over a two-year period: once over the 24 months before the start of the intervention and once in the intervention period.[2] We avoided calculating the measures for overlapping periods, meaning that no measurement period included services provided in another measurement period.

In contrast, the measure of outpatient ED visits was defined as a quarterly measure—that is, assessed over a period of three months. This means we measured the outpatient ED visit rate 4 times in the year-long baseline period (or twice for Cohort 1, which had a six-month baseline period; see supplemental material at the end of this chapter) and up to 11 times in the intervention period.

An important implication of the number of measurements is that we could assess whether outcomes followed a similar trend for the treatment and comparison groups in the baseline period for the ED visit rate but not for the quality-of-care processes or outcomes measures. Being able to assess trends is helpful because verifying parallel trends in the baseline period can give greater confidence in the key assumptions of the difference-in-differences model.

[2] The intervention period covered eight intervention quarters. This measure applied only to Cohorts 1, 2, and 3, because we did not observe beneficiaries in Cohort 4 for eight intervention quarters.

We did not estimate impacts on spending because the Ohio Medicaid data did not contain expenditure data for children enrolled in managed care, which represented most children in the study.

b. Covariates

For each beneficiary, we calculated the following covariates:

1. Pediatric Medical Complexity Algorithm complexity groupings (Simon et al. 2014)
2. Demographics (age, gender, and race or ethnicity)
3. Blind or disabled categorical eligibility
4. Managed care enrollment

We defined demographic covariates as of the start of each quarter and defined all remaining covariates as of the start of the relevant period (baseline or intervention) for children who were at least six months old at the start of each period. For children younger than six months at the start of a period, we calculated these covariates as of their sixth month of life (to obtain a more stable picture of these covariates). We did not measure the nondemographic covariates for children born late in either period who did not reach six months of age during the period. For these infants, we created a series of indicator variables to denote missing covariate values, so as to avoid dropping them from the regression model.

6. Regression model

We used a regression model to implement the difference-in-differences design for estimating impacts. For each outcome, the model estimates the relationship between the outcome and a series of predictor variables, assuming that each of the predictor variables has a linear (additive) relationship with the outcome. The predictor variables include the beneficiary-level covariates; beneficiary assignment (treatment or comparison practice); an indicator for each practice (a fixed effect that accounts for differences between practices in their patients' outcomes at baseline); indicators for each post-intervention measurement period (quarter or year); and an interaction of a beneficiary's treatment status with one or more post-intervention indicator variables, depending on the post-intervention measurement period included in the primary tests.

The estimated relationship between the interaction term and the outcome in a given quarter or year (depending on the outcome measure) is the impact estimate for that interval. It measures the average difference between outcomes for beneficiaries assigned to the treatment and comparison practices during that period, subtracting out any differences between these groups during the baseline period. The model uses robust standard errors to account for clustering of outcomes across measurement intervals for the same beneficiary and a dummy variable for each practice to account for clustering of outcomes for beneficiaries assigned to the same practice. Appendix 2 provides details on the regression models.

7. Primary tests

Table V.2 shows the five primary tests for UHC, across three domains.

Outcomes. UHC aimed to reduce avoidable outpatient ED visits and to improve 11 quality-of-care processes, of which we examined lead screening and well-child visits. (The other processes were not easily measured in claims data.) Because the program aimed to improve fluoride varnish application, a process that we cannot measure reliably in claims, we expected that the intervention would also eventually reduce the need for, and receipt of, dental caries treatment (which can be measured in claims).

Time period. UHC focused on different outcomes at different points in the award period (Table III.1). To provide time for practices to implement the intervention in response to UHC's focus areas, we chose to analyze impacts on ED visits starting one quarter after each practice's intervention start date through the end of the intervention period. For lead screening, we chose to analyze impacts in the last two years of the intervention. For well-child visits, we chose to analyze impacts in the last year of the intervention. Finally, for dental caries treatment, we chose to analyze outcomes only in the last year of the program to provide time for the process of fluoride varnish application to affect the outcome of caries among eligible children.

Substantive thresholds. Some impact estimates could be large enough to be policy relevant (to CMMI and other stakeholders) even if they are not statistically significant; for this reason, we have prespecified thresholds for what we call *substantive importance*. We express the threshold as a percentage change from the counterfactual. UHC expected a 15 percent reduction in the *avoidable* ED visit rate. Based on estimates in the awardee's application and existing research literature (Christensen et al. 2015), we estimated that about 70 percent of ED visits for Medicaid children are for not urgent, potentially avoidable, or primary care-treatable conditions. Thus a 15 percent decline in avoidable ED visits would translate into a 10.5 percent reduction in the all-cause ED visit rate that we assess in our evaluation. We chose a substantive threshold for this outcome of 7.9 percent (for both the full evaluation population and the high-risk subgroup) because it is 75 percent of UHC's expected impact. (We use 75 percent recognizing that UHC could still be considered successful if it approached, but did not fully achieve, its anticipated effects.) Because UHC did not state measure-specific goals for the quality-of-care processes and outcomes measures in the study, we based our substantive thresholds on prior studies of similar interventions (Meropol et al. 2014; Pahel et al. 2011). We chose a substantive threshold of 15 percent for lead screening and well-child visits and a substantive threshold of 10 percent for dental caries treatment.

Table V.2. Specification of the primary tests for UHC

Domain (number of tests in the domain)[a]	Outcome (units)	Time period for impacts	Population	Substantive threshold (expected direction of effect)[c]
Quality-of-care processes (2)	Lead screening (binary [yes or no]/beneficiary/two years)	Two-year period ending in I11 for Cohort 1, I9 for Cohort 2, and I8 for Cohort 3	All observable Medicaid beneficiaries age 2 years (24 to 35 months) assigned to treatment practices	15.0% (+)
	Well-child visit (binary [yes or no]/beneficiary/year)	One-year period ending in I11 for Cohort 1, I9 for Cohort 2, I8 for Cohort 3, and I6 for Cohort 4	All observable Medicaid beneficiaries ages 3 to 6 years (36 to 72 months) assigned to treatment practices	15.0% (+)
Quality-of-care outcomes (1)	Dental caries treatment (binary [yes or no]/beneficiary/year)	One-year period ending in I11 for Cohort 1	All observable Medicaid beneficiaries ages 3 to 5 years (36 to 59 months) assigned to treatment practices	10.0% (-)
Service use (2)	Outpatient ED visits (#/1,000 beneficiaries/quarter)	I2 through I11 for Cohort 1, I2 through I9 for Cohort 2, I2 through I8 for Cohort 3, and I2 through I6 for Cohort 4	All observable Medicaid beneficiaries attributed to treatment practices	7.9% (-)
	Outpatient ED visits (#/1,000 beneficiaries/quarter)	I2 through I11 for Cohort 1, I2 through I9 for Cohort 2, I2 through I8 for Cohort 3, and I2 through I6 for Cohort 4	All observable high-risk Medicaid beneficiaries attributed to treatment practices[d]	7.9% (-)

Note: For all primary tests, the expected direction of effect is a decrease relative to the counterfactual with the exception of lead screening and well-child visits, which are expected to increase.

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

[b] The regression models controlled for differences between the treatment and comparison groups during the baseline year when estimating program impacts.

[c] The substantive threshold is the impact as a percentage of the counterfactual. The counterfactual is the outcome the treatment group would have had in the absence of the HCIA-funded intervention. For the outpatient ED visit rate, we set the substantive threshold to 75 percent of UHC's anticipated impact. For the other outcomes, we set the substantive threshold based on evidence from the literature.

[d] High-risk beneficiaries are defined as assigned beneficiaries classified as "children with complex chronic disease" or "children with noncomplex chronic disease" based on the Pediatric Medical Complexity Algorithm at the start of the baseline and/or intervention periods.

ED = emergency department; HCIA = Health Care Innovation Award; I = intervention quarter; UHC = University Hospitals of Cleveland.

8. Secondary tests (robustness checks)

The evaluation also included secondary quantitative tests to help corroborate the findings from the primary tests. This is important because some of the differences observed between the treatment and comparison groups in the primary test results could reflect limitations of the study design or random fluctuations in the data. We have greater confidence in the primary results if they are generally consistent with the expected broader pattern of results from the secondary tests.

We conducted two secondary tests for UHC. First, we reran the primary tests for well-child visits, dental caries treatment, ED visits among the full population, and ED visits among the high-risk population, limiting the sample only to beneficiaries assigned to the treatment and comparison groups at the start of the period, either baseline or intervention.[3] It is possible that differences in sample addition between the treatment and comparison groups could bias the impact results to some degree if the sample members added over time differ from earlier sample members (if, for example, they are healthier); this could create differences in mean outcomes between the treatment and comparison groups that are unrelated to the HCIA intervention. We explored this possibility in particular because the practice facilitation program component aimed to increase well-child visits among children, but we used well-child visits to attribute beneficiaries to practices. Thus if the program succeeded increasing well-child visit rates, it could have led, on average, to assigning healthier beneficiaries (who had well-child visits) to the treatment practices (relative to beneficiaries assigned to the comparison practices), making it appear that outcomes for the treatment group were better than those for the comparison group. We allowed newborns into the sample because the program is less likely to affect their well-child visit rate.

Second, we conducted two-sided statistical tests for lead screening, well-child visits, and dental caries treatment after reviewing initial primary test results. As we describe in the next section (V.A.9), all of our primary tests are one-sided tests, testing only for favorable impacts. However, as we show in Section V.C, many of our findings suggested the program could be having an *unfavorable* impact, which is counterintuitive. Because we have no reason to believe that the intervention would have led to significantly worse outcomes for the treatment group in these domains, we added these two-sided tests as robustness checks. Significantly unfavorable effects in these post hoc secondary tests would suggest there was a problem with the comparison group or data, and that the primary test results are suspect.

9. Drawing impact conclusions based on quantitative tests and implementation evidence

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

1. Statistically significant favorable effect (the highest level of evidence)

[3] The one exception to this sample restriction, which prevents new sample addition during either the baseline or intervention periods, was that we allowed beneficiaries who were born during the period to be added to the sample.

2. Substantively important (but not statistically significant) favorable effect
3. Substantively important (but not statistically significant) unfavorable effect
4. No substantively large effect
5. Indeterminate effect

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

Appendix 3 describes our decision rules for each of the five possible conclusions. In short, we concluded that a program had a statistically significant favorable effect in a domain if it met the following criteria:

1. At least one primary test result in the domain was favorable and statistically significant, after adjusting the statistical tests to account for multiple tests (if applicable) within a domain
2. The average impact estimate across all primary tests in the domain was favorable and statistically significant

In both cases, we also had to determine that the primary test results were plausible given the results of the secondary tests and implementation evidence. We concluded that a program had a substantively important favorable effect if the average impact estimate in the domain was substantively important but not statistically significant, and if the result was plausible given the secondary tests and implementation evidence.

In contrast, if the average impact estimate was unfavorable (opposite the hypothesized direction) and larger than the substantive threshold, and unfavorable effects were plausible given the other evidence, we concluded the program had a substantively important unfavorable effect. If the tests in a domain did not meet any of these criteria, we instead used the following rules.

First, if the tests for at least one outcome in the domain (or all outcomes in the domain together) had sufficient statistical power to detect an impact of the size of the substantive threshold with at least 75 percent probability, we concluded that there was not a substantively large effect because we are reasonably confident that we would have detected such a large effect had there been one.

Second, if the power was not sufficient (less than 75 percent) to detect this type of impact, we concluded the impact in the domain was indeterminate. Indeterminate means either that the program truly did not have effects that were substantively large or that it did, but our statistical tests were unable to detect them.

Finally, if the results for the primary tests in a domain were not plausible given the implementation evidence or the secondary, corroborating tests, we drew no conclusions about program impacts in that domain.

B. Baseline characteristics of the treatment and comparison practices

1. Treatment practices

Almost all 37 treatment practices (97 percent) were located in an urban zip code (Table V.3). The SK&A data classified about 78 percent of non-FQHC treatment practices as being owned by a hospital or health system. The non-FQHC practices had an average of 4.9 providers, with about 87 percent of providers having a pediatric primary care specialty. Almost one-half (49 percent) of non-FQHC practices had at least one provider who received payments from the Centers for Medicare & Medicaid Services for Medicaid meaningful use of EHRs in the baseline period.

The treatment practices had an average of 925 assigned Medicaid child beneficiaries during the baseline period. The mean age of beneficiaries attributed to treatment practices during the baseline period was 7.0 years. A majority (90 percent) of beneficiaries attributed to treatment practices were enrolled in Medicaid managed care during the baseline period. Slightly more than one-third (35 percent) of attributed beneficiaries were classified as high-risk patients.

About 29 percent of attributed beneficiaries who were 2 years old by the end of the baseline period had received a lead screening test in the previous two years. Among the children ages 3 to 6 years by the end of the baseline period, about three-quarters (78.0 percent) had a well-child visit in the previous year. This was higher than the national average of 67.1 percent. About 8.0 percent of beneficiaries ages 3 to 5 years by the end of the baseline period received dental caries treatment in the previous year. The mean outpatient ED visit rate was 178 per 1,000 patients per quarter among all treatment group beneficiaries (modestly higher than the national average of 165.3) and 216 per 1,000 among high-risk treatment group beneficiaries.

2. Treatment and comparison group balance

The comparison practices were similar (within 0.25 standardized differences) to the intervention practices on many characteristics (Table V.3). In particular, the two groups of practices were well balanced on baseline measures of three study outcomes: outpatient ED visit rate (overall and for high-risk beneficiaries), well-child visit rates, and lead screening. In addition to being balanced on the ED visit rate level, the two groups had similar trends in ED visit rates during the baseline period (as we discuss in Section V.F).

Five characteristics were outside of our target of 0.25 standardized differences. This included one of the study outcomes—dental caries treatment (7.9 percent of eligible patients received caries treatment in the intervention group versus 9.6 for the comparison group). The treatment practices had a higher proportion of black beneficiaries than the matched comparison practices (34.6 versus 19.8 percent), and a slightly lower proportion of high-risk beneficiaries than the matched comparison practices (35.1 versus 37.1 percent, respectively). The proportion

of treatment practices with providers who attested to meaningful use in the Medicaid EHR Incentive Program differed notably (48.7 and 30.3 percent, respectively).

We attempted a number of matching iterations in an attempt to improve the balance on baseline performance of the outcome measures. In particular, we focused on prioritizing the dental caries treatment outcome in the matching model because we were outside of our target of 0.25 standardized differences. However, when we prioritized the dental caries treatment outcome, we achieved substantially worse balance on several other characteristics, including the other outcome measures. Therefore, we had to accept imbalance on the dental caries outcome to maintain balance within our targets for the other outcomes. The imbalance among our treatment and matched comparison groups on several characteristics reflects the trade-offs that we faced with matching—we had a relatively small number of practices in our potential comparison pool and we could not identify a final matched comparison group that was balanced on all characteristics.

Table V.3. Baseline characteristics of the treatment and comparison practices

Characteristic of practice	Treatment practices (n = 37)	Matched comparison group (n = 109)	Absolute difference	Standard-ized difference[a]	National Medicaid average
Characteristics of all beneficiaries attributed to practices during the baseline period					
Number of assigned Medicaid beneficiaries	924.9	795.6	129.3	0.23	n.a.
Outpatient ED visits for all assigned beneficiaries[b]	178.1	183.7	-5.66	-0.14	165.3
Outpatient ED visits for high-risk beneficiaries[b]	215.6	210.6	4.93	0.10	n.a.
Proportion age 2 years with appropriate lead screening (%/2 years)	28.8	28.9	-0.15	-0.01	n.a.
Proportion ages 3 to 6 with a well-child visit (%/year)	77.8	75.8	2.01	0.17	67.1
Proportion ages 3 to 5 receiving treatment for dental caries (%/year)	7.9	9.6	-1.67	-0.49	n.a.
Age in years (mean)	7.0	6.6	0.47	0.50	7.9
Female (%)	51.3	51.5	-0.26	-0.09	50.2
Race: black (%)	34.6	19.8	14.83	0.69	23.7
Ethnicity: Hispanic (%)	0.6	0.8	-0.13	-0.14	43.7
Blind or disabled categorical eligibility (%)	2.3	2.0	0.03	0.13	n.a.
Enrolled in managed care (%)	89.7	89.5	0.25	0.04	86.5[k]
High-risk (%)	35.1	37.1	-1.99	-0.29	n.a.
Characteristics of the practices					
FQHC (%)	10.8	10.8	0	0	n.a.
Providers in practice (#)[c]	4.9	5.4	-0.51	-0.11	n.a.
Urban location (%)	97.3	99.5	-2.16	-0.19	n.a.

Table V.3 *(continued)*

Characteristic of practice	Treatment practices (n = 37)	Matched comparison group (n = 109)	Absolute difference	Standard-ized difference[a]	National Medicaid average
Median household income in the zip code ($)	54,489.4	55,772.8	-1,283.4	-0.07	n.a.
Pediatric providers in practice (%)[c]	86.9	89.4	-2.56	-0.11	n.a.
Owned by a hospital or health system[c] (%)	78.4	69.3	9.11	0.19	n.a.
Medicaid meaningful use of EHRs (%)[c,d]	48.7	30.3	18.34	0.37	n.a.
Certified as a patient-centered medical home by NCQA (%)[c]	0.0	0.5	-0.54	-0.07	n.a.

Sources: Analysis of the Ohio Medicaid eligibility and claims data obtained from the Ohio Colleges of Medicine Government Resource Center. Zip code data merged from the 2012 Five-Year American Community Survey Zip Code Characteristics and from the Ohio Department of Health. Benchmark data are from the U.S. Department of Health and Human Services (2016); Centers for Disease Control and Prevention, National Health Interview Survey, 2012 Data Release; Medicaid and CHIP Payment and Access Commission (2014).

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

[a] The standardized difference is the difference in means between the treatment and comparison groups divided by the standard deviation of the matching variable, which is pooled across the treatment and comparison groups.

[b] # of visits/1,000 beneficiaries/quarter.

[c] Measured for all practices except FQHCs. Due to data limitations, these variables were not available for FQHCs.

[d] Percentage of practices with at least one provider (NPI) working in the practice who attested to meaningful use in the Medicaid EHR Incentive Program during the baseline period. Applies only to non-FQHC practices.

ED = emergency department; EHR = electronic health record; FQHC = federally qualified health center; NCQA = National Committee for Quality Assurance; NPI = National Provider Identifier.

n.a. = not applicable.

C. Primary test results

In this section, we present the results of the primary tests, by domain. We also describe what drives these primary test results by comparing the difference in mean outcomes for the treatment and comparison groups in the baseline and intervention periods.

1. Quality-of-care processes

The proportions of the eligible populations receiving lead screening or a well-child visit were 14.9 and 5.5 percent lower, respectively, for the treatment group (unfavorable estimates) than the estimated counterfactual.[4] The combined impact estimate across these two measures in the domain was -10.2 percent, also an unfavorable estimate. The statistical power values in Table V.4 show that our analysis was well powered (with roughly 99 percent power) to detect differences the size of the substantive threshold in both lead screening and well-child visits, if there had been true impacts this large.

[4] Our estimated counterfactual—the outcome the treatment group members would have had in the absence of the HCIA intervention—is the treatment group mean minus the difference-in-differences estimate.

For well-child visits, a reduction in well-child visits from the baseline to the intervention period for the treatment group (from 77.1 to 63.3 percent) that was larger than the reduction for the comparison group (from 75.6 to 65.4 percent) (Table V.5) drove these primary test results. For lead screening, the percentage of children who received recommended screening declined for the treatment group (from 31.5 in the baseline period to 27.1 in the intervention period), whereas it increased slightly for the comparison group (from 23.0 to 25.4 percent).

The declines in well-child visits from baseline to intervention for both the intervention and comparison groups are likely due, at least in part, to how we defined the sample—and might not reflect true declines in quality of care. We used an intent-to-treat sample definition, as described in the supplemental material at the end of this chapter. This meant that, in the baseline and intervention periods, we assigned beneficiaries to the first intervention or comparison practice they were attributed to during the period, and then continued to assign the beneficiary to that practice for all remaining quarters in the period, whether or not they continued to visit that practice over time.

By definition, when a beneficiary is first assigned to a practice, he or she must have had at least one evaluation and management (E&M) visit in the past year. But, over time, beneficiaries will continue to be assigned to the practice, whether they continue to have any evaluation and management visits. So, beneficiaries in later quarters do not, by definition, need to have an E&M visit in the prior year, which—all else equal—should reduce the percentage of those in the group in these later quarters who had a well-child visit in the prior year. This effect should be stronger in the intervention period than the baseline period, because the intervention period is longer (up to 11 quarter versus the 4 baseline quarters). This should not bias the impact estimates, however, because the influence of sample definition on outcome levels should be the same for the intervention and comparison groups.

2. Quality-of-care outcomes

The proportion of eligible beneficiaries receiving dental caries treatment was 24.9 percent higher for the treatment group (an unfavorable estimate) than the estimated counterfactual (Table V.4). This estimate is greater than the substantive threshold of 10.0 percent. We had poor statistical power (22.4 percent) to detect true impacts the size of the substantive threshold.

We observed this unfavorable primary test estimate because—although the rate of dental caries treatment declined over time for both the treatment and the comparison groups—the reduction from the baseline to the intervention period was smaller for the treatment group (8.0 to 7.7 percent) than for the comparison group (10.7 to 8.1 percent) (Table V.5).

3. Service use

The treatment group's average number of outpatient ED visits was 163.3 per 1,000 beneficiaries per quarter during the primary test period. This was 2.6 fewer visits per 1,000 beneficiaries per quarter than the estimated counterfactual, a 1.5 percent favorable difference. This estimate was not statistically significant. We do not consider this point estimate to be substantively large because it is smaller than the substantive threshold of 7.9 percent. Among the high-risk subgroup, the treatment group's average number of outpatient ED visits was 195.5 per

1,000 beneficiaries per quarter during the primary test period. This was 2.7 fewer visits per 1,000 high-risk beneficiaries than the estimated counterfactual, which is a 1.3 percent favorable difference. As with the full population, this estimate was neither statistically significant nor substantively large. The combined estimate for the two outcomes in this domain is a -1.4 percent favorable difference, but this is also not statistically significant nor substantively large. Statistical power to detect effects for the service use domain was good (99.9 percent for the combined effect) (Table V.4).

In both the intervention and comparison groups, the outpatient ED visit rates declined from the baseline to the intervention period (Table V.6). On average, the decline was slightly larger for the intervention group than the comparison group, which drove the small—though statistically not significant—favorable point estimate for the primary test result.

Table V.4. Results of primary tests for UHC

Primary test definition					Statistical power[a] to detect an effect that is		Results			
Domain (# of tests in domain)	Outcome (units)	Time period for impacts	Population	Substantive threshold (expected direction of effect)[b]	Size of the substantive threshold	Twice the size of the substantive threshold[c]	Treatment group mean	Regression-adjusted difference between the treatment group and estimated counterfactual (standard error)[b]	Percentage difference[d]	*p*-value[e]
Quality-of-care processes (2)	Lead screening (binary [yes or no]/beneficiary/two years)	Two-year period ending in I11 for Cohort 1, I9 for Cohort 2, and I8 for Cohort 3	All observable Medicaid beneficiaries age 2 years (24 to 35 months) assigned to treatment practices	15.0% (+)	98.6%	>99.9%	27.1	-4.8 (1.4)	-14.9%	>0.99
	Well-child visit (binary [yes or no]/beneficiary/year)	One-year period ending in I11 for Cohort 1, I9 for Cohort 2, I8 for Cohort 3, and I6 for Cohort 4	All observable Medicaid beneficiaries ages 3 to 6 years (36 to 72 months) assigned to treatment practices	15.0% (+)	>99.9%	>99.9%	63.3	-3.7 (1.0)	-5.5%	>0.99
	Combined (%)	Varies by test	Varies by test	15.0% (+)	>99.9%	>99.9%	n.a.	n.a.	-10.2%	>0.99
Quality-of-care outcomes (1)	Dental caries treatment (binary [yes or no]/beneficiary/year)	One-year period ending in I11 for Cohort 1	All observable Medicaid beneficiaries ages 3 to 5 years (36 to 59 months) assigned to treatment practices	10.0% (-)	22.4%	40.7%	7.7	1.5 (1.2)	24.9%	0.90
Service use (2)	Outpatient ED visits (#/1,000 beneficiaries /quarter)	I2 to I11 for Cohort 1, I2 to I9 for Cohort 2, I2 to I8 for Cohort 3, and I2 to I6 for Cohort 4	All observable Medicaid beneficiaries attributed to treatment practices	7.9% (-)	>99.9%	>99.9%	163.3	-2.6 (2.8)	-1.5%	0.27

Table V.4 *(continued)*

Primary test definition					Statistical power[a] to detect an effect that is		Results			
Domain (# of tests in domain)	Outcome (units)	Time period for impacts	Population	Substantive threshold (expected direction of effect)[b]	Size of the substantive threshold	Twice the size of the substantive threshold[c]	Treatment group mean	Regression-adjusted difference between the treatment group and estimated counterfactual (standard error)[b]	Percentage difference[d]	*p*-value[e]
	Outpatient ED visits (#/1,000 beneficiaries /quarter)	I2 to I11 for Cohort 1, I2 to I9 for Cohort 2, I2 to I8 for Cohort 3, and I2 to I6 for Cohort 4	All observable high-risk Medicaid beneficiaries attributed to treatment practices	7.9% (-)	97.4%	>99.9%	195.5	-2.7 (4.9)	-1.3%	0.40
	Combined (%)	I2 to I11 for Cohort 1, I2 to I9 for Cohort 2, I2 to I8 for Cohort 3, and I2 to I6 for Cohort 4our	Varies by test	7.9% (-)	99.9%	>99.9%	n.a.	n.a.	-1.4%	0.22

Source: Analysis of the Ohio Medicaid claims received from the Ohio Colleges of Medicine Government Resource Center.

Note: The results for each outcome are based on a difference-in-differences regression model, as described in the text. Estimates are calculated for Medicaid beneficiaries who are observable in the relevant time period: that is, beneficiaries who were alive, enrolled in Ohio Medicaid, had no third-party coverage, and were ages birth to 18 years. Additional sample restrictions apply to the quality-of-care process measures; see text for details. We defined *high-risk* as beneficiaries who were classified as "children with complex chronic disease" or "children with noncomplex chronic disease" based on the Pediatric Medical Complexity Algorithm (Simon et al. 2014).

[a] The power calculation is based on actual standard errors from the analysis. For example, in the first row of the service use domain, a 7.9 percent effect on outpatient ED visits for all Medicaid beneficiaries (from the counterfactual of 163.3 + 2.6 = 165.9) would be a change of 13.1 ED visits. Given the standard error of 2.8 from the regression model, we would be able to detect a statistically significant result more than 99.9 percent of the time if the impact was truly -13.1, assuming a one-sided statistical test at the $p = 0.10$ significance level.

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

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

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

[e] *p*-values test the null hypothesis that the regression-adjusted difference-in-differences estimate is greater than or equal to zero for outcomes in the quality-of-care processes domain, or less than or equal to zero in all other domains (a one-sided test). We adjusted the *p*-values for the multiple (two) comparisons made within the quality-of-care processes domain, and (separately) for the two comparisons made within the service use domain.

ED = emergency department; HCIA = Health Care Innovation Award; I = intervention quarter; UHC = University Hospitals of Cleveland.

n.a. = not applicable

Table V.5. Unadjusted mean outcomes (quality-of-care processes and outcomes) observed among select Medicaid beneficiaries, by treatment status and time period

			Number of Medicaid beneficiaries (practices)			Mean outcomes		
Outcome	Period	Time period	T	C (not weighted)	C (weighted)	T	C	Difference (%)
Received lead screening	Baseline	Two-year period[a]	2,542 (34)	6,338 (99)	2,713	31.5%	23.0%	8.6 p.p. (37.4%)
(Among those age 2 years [24 to 35 months]/beneficiary/two years)	Intervention	Two-year period[b]	3,423 (34)	7,882 (99)	3,562	27.1%	25.4%	1.6 p.p. (6.5%)
Received well-child visit	Baseline	One-year period[c]	7,617 (37)	19,459 (109)	8,179	77.1%	75.6%	1.5 p.p. (2.0%)
(Among those ages 3 to 6 years/beneficiary/year)	Intervention	One-year period[d]	9,848 (37)	23,269 (109)	9,820	63.3%	65.4%	-2.2 p.p. (-3.3%)
Received dental caries treatment	Baseline	One-year period[e]	3,440 (23)	6,241 (56)	3,669	8.0%	10.7%	-2.8 p.p. (-25.7%)
(Among those ages 36 to 59 months/beneficiary/ year)	Intervention	One-year period[f]	4,744 (23)	7,616 (56)	4,379	7.7%	8.1%	-0.4 p.p. (-5.4%)

Sources: Analysis of Ohio Medicaid claims data received from the Ohio Colleges of Medicine Government Resource Center.

Notes: The baseline quarters are measured relative to the start of the baseline period on August 1, 2012, for Cohorts 1 and 2; January 1, 2013, for Cohort 3; and July 1, 2013, for Cohort 4. For example, the first baseline quarter (B3) for Cohort 1 runs from August 1, 2012, to October 31, 2012. Cohort 1 has only two baseline quarters, whereas Cohorts 2 through 4 have four baseline quarters (see text for further explanation). The intervention quarters are measured relative to the start of the intervention period on February 1, 2013, for Cohort 1; August 1, 2013, for Cohort 2; January 1, 2014, for Cohort 3; and July 1, 2014, for Cohort 4. For example, the first intervention quarter (I1) for Cohort 1 runs from February 1, 2013, to April 30, 2013. In each period (baseline or intervention), the treatment group each quarter includes all beneficiaries assigned to a treatment practice by the start of the quarter and who met other sample criteria—that is, they were alive, enrolled in Ohio Medicaid, had no third-party coverage, and were ages birth to 18 years. In each period, the comparison group includes all beneficiaries assigned to a comparison practice by the start of the quarter and who met the other sample criteria. See text for details.

Table V.5 *(continued)*

The outcome means were weighted such that (1) each treatment beneficiary gets a weight of 1; and (2) each comparison beneficiary gets a weight that is the product of two weights: (a) a matching weight, equal to the reciprocal of the total number of comparison practices matched to the same treatment practice as the beneficiary's assigned practice, and (b) a practice size weight, which equals the average number of beneficiaries assigned to the matched treatment practice during the two (for Cohort 1) or four (for Cohorts 2 through 4) baseline quarters divided by the average number of beneficiaries assigned to the beneficiary's comparison practice over those quarters. The difference between the treatment and comparison groups in a quarter is calculated by subtracting the mean outcome for the comparison group from the mean outcome for the treatment group. The percentage difference equals that difference divided by the mean outcome for the comparison group.

[a] Two-year period ending B4.

[b] Two-year period ending I11 for Cohort 1, I9 for Cohort 2, and I8 for Cohort 3.

[c] One-year period ending B4.

[d] One-year period ending I11 for Cohort 1, I9 for Cohort 2, I8 for Cohort 3, and I6 for Cohort 4.

[e] One-year period ending B4.

[f] One-year period ending I11 for Cohort 1.

B = baseline; C = comparison; I = intervention; p.p. = percentage point; T = treatment.

Table V.6. Unadjusted mean outcomes (service use) measured for Medicaid beneficiaries, by treatment status and quarter

		Number of Medicaid beneficiaries (practices)			Number of high-risk Medicaid beneficiaries (practices)			Outpatient ED visit rate (#/1,000 beneficiaries/quarter)			Outpatient ED visit rate (#/1,000 high-risk beneficiaries/quarter)		
	Q	T	C (no wgt)	C (wgt)	T	C (no wgt)	C (wgt)	T	C	Diff (%)	T	C	Diff (%)
Baseline period[a]	B1	9,700 (14)	36,437 (53)	9,857	3,647 (14)	14,187 (53)	3,732	196.9	200.1	-3.2 (-1.6%)	241.7	224.9	16.8 (7.5%)
	B2	10,651 (14)	39,796 (53)	10,740	3,866 (14)	14,884 (53)	3,888	207.6	228.7	-21.1 (-9.2%)	230.0	253.8	-23.9 (-9.4%)
	B3	33,330 (37)	81,213 (109)	33,631	11,516 (37)	30,032 (109)	11,642	169.0	171.5	-2.5 (-1.5%)	202.1	200.8	1.3 (0.7%)
	B4	36,545 (37)	86,509 (109)	35,998	12,163 (37)	30,685 (109)	11,930	182.2	191.6	-9.3 (-4.9%)	210.1	212.0	-1.8 (-0.9%)
Intervention period[b]	I1	32,263 (37)	77,031 (109)	32,548	11,731 (37)	30,366 (109)	11,898	167.1	176.4	-9.3 (-5.3%)	201.5	209.7	-8.2 (-3.9%)
	I2	36,146 (37)	83,700 (109)	35,336	12,660 (37)	31,666 (109)	12,323	170.6	177.6	-7.0 (-3.9%)	202.8	200.0	2.8 (1.4%)
	I3	38,801 (37)	88,072 (109)	37,270	13,227 (37)	32,289 (109)	12,568	166.4	174.8	-8.4 (-4.8%)	199.6	209.3	-9.7 (-4.6%)
	I4	41,508 (37)	93,321 (109)	39,647	13,727 (37)	33,065 (109)	12,874	165.7	173.1	-7.4 (-4.3%)	195.6	192.8	2.8 (1.5%)
	I5	43,372 (37)	96,847 (109)	41,192	14,017 (37)	33,445 (109)	13,012	162.2	175.9	-13.6 (-7.8%)	193.4	198.8	-5.5 (-2.8%)
	I6	44,685 (37)	100,830 (109)	42,941	14,096 (37)	33,802 (109)	13,153	167.6	176.7	-9.1 (-5.1%)	195.0	197.3	-2.3 (-1.2%)
	I7	43,882 (34)	94,426 (99)	42,236	13,556 (34)	30,866 (99)	12,615	168.1	177.6	-9.5 (-5.3%)	206.4	205.1	1.2 (0.6%)
	I8	45,459 (34)	98,151 (99)	44,253	13,687 (34)	31,131 (99)	12,763	179.2	190.8	-11.5 (-6.0%)	209.6	216.0	-6.3 (-2.9%)
	I9	43,591 (30)	84,796 (79)	42,891	12,873 (30)	26,193 (79)	12,018	161.1	171.0	-9.9 (-5.8%)	193.6	193.1	0.5 (0.3%)
	I10	35,665 (23)	57,067 (56)	33,718	10,116 (23)	17,106 (56)	8,986	154.8	157.6	-2.7 (-1.7%)	192.6	184.1	8.5 (4.6%)
	I11	35,773 (23)	57,180 (56)	33,703	9,899 (23)	16,767 (56)	8,814	137.6	162.5	-24.9 (-15.3%)	166.7	193.3	-26.6 (-13.8%)

Source: Analysis of the Ohio Medicaid claims received from the Ohio Colleges of Medicine Government Resource Center.

Notes: The baseline quarters are measured relative to the start of the baseline period on August 1, 2012, for Cohorts 1 and 2; January 1, 2013, for Cohort 3; and July 1, 2013, for Cohort 4. For example, the first baseline quarter (B3) for Cohort 1 runs from August 1, 2012, to October 31, 2012. Cohort 1 has only two baseline quarters, whereas Cohorts 2 through 4 have four baseline quarters (see text for further explanation). The intervention quarters are measured relative to the start of the intervention period on February 1, 2013, for Cohort 1; August 1, 2013, for Cohort 2; January 1, 2014, for Cohort 3;

Table V.6 *(continued)*

and July 1, 2014, for Cohort 4. For example, the first intervention quarter for Cohort 1 (I1) runs from February 1, 2013, to April 30, 2013. In each period (baseline or intervention), the treatment group each quarter includes all beneficiaries assigned to a treatment practice by the start of the quarter and who met other sample criteria—that is, they were alive, enrolled in Ohio Medicaid, had no third-party coverage, and were ages birth to 18 years. In each period, the comparison group includes all beneficiaries assigned to a comparison practice by the start of the quarter and who met the other sample criteria. See text for details.

The outcome means were weighted such that (1) each treatment beneficiary gets a weight of 1; and (2) each comparison beneficiary gets a weight that is the product of two weights: (a) a matching weight, equal to the reciprocal of the total number of comparison practices matched to the same treatment practice as the beneficiary's assigned practice, and (b) a practice size weight, which equals the average number of beneficiaries assigned to the matched treatment practice during the two (for Cohort 1) or four (for Cohorts 2 through 4) baseline quarters divided by the average number of beneficiaries assigned to the beneficiary's comparison practice over those quarters. The difference between the treatment and comparison groups in a quarter is calculated by subtracting the mean outcome for the comparison group from the mean outcome for the treatment group. The percentage difference equals that difference divided by the mean outcome for the comparison group.

We defined *high-risk* as beneficiaries who were classified as "children with complex chronic disease" or "children with noncomplex chronic disease" based on the Pediatric Medical Complexity Algorithm (Simon et al. 2014).

[a] The baseline period is August 1, 2012, to January 31, 2013 for Cohort 1; August 1, 2012, to July 31, 2012 for Cohort 2; January 1, 2013, to December 31, 2013, for Cohort 3; and July 1, 2013, to June 30, 2014, for Cohort 4.

[b] The intervention period is February 1, 2013, to October 31, 2015, for Cohort 1; August 1, 2013, to October 31, 2015, for Cohort 2; January 1, 2014, to December 31, 2015, for Cohort 3; and July 1, 2014, to December 31, 2015, for Cohort 4.

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

D. Secondary test results

1. Estimates limiting the sample to prevent sample addition, except for children born during the relevant period

The secondary test results—in which we limited the sample to beneficiaries attributed at the start of the baseline or intervention period (except for children born during the period, whom we do allow to enter the sample)—are generally consistent with the primary test results (Table V.7). As with the primary tests, we found estimates in the unfavorable direction for well-child visits and dental caries treatment. We found favorable estimates for ED visits among the full and high-risk populations, similar to the estimates from the primary tests, but the magnitude of the estimates is slightly larger (although still under our substantive threshold). The estimate for ED visits among the full population is statistically significant. However, the *p*-values from the secondary test results were not adjusted for multiple comparisons within the domain, unlike the primary test results. Although there are a few differences in the estimates from this set of secondary tests and the primary tests, the results do not contradict the primary test results.

2. Estimates using two-sided statistical tests

As noted in Section V.A.8, we conducted two-sided statistical tests for the outcomes in the quality-of-care processes and outcomes domain to determine if the unfavorable impact estimates from the primary tests were statistically significant. The unfavorable impact estimates for lead screening (-14.9 percent) and well-child visits (-5.5 percent) were statistically significant in a two-sided test with a threshold of significance of either 0.1 or 0.05 ($p = 0.001$ and < 0.001, respectively), but the unfavorable impact estimate for dental caries treatment (24.9 percent) was not statistically significant at a conventional level of significance ($p = 0.193$). Given the unexpected findings, we need to find a plausible mechanism through which the intervention could have worsened quality-of-care processes, or else question whether there might be a limitation in the comparison group. We discuss this issue in the next section (V.E).

3. Exploratory test dropping Cohort 4 practices from ED visit sample.

The estimates for ED visits after dropping Cohort 4 practices were similar to the estimates from the primary tests. We found a 1.7 percent favorable difference for ED visits among the full population and a 1.8 percent favorable difference for ED visits among the high-risk population, broadly consistent with the results for the primary tests. This suggests that, even though some ED avoidance intervention activities occurred during the Cohort 4 baseline period, including the Cohort 4 practices did not bias the primary test results.

Table V.7. Results of secondary tests for UHC

Secondary test definition				Results			
Domain (# of tests in domain)	Outcome (units)	Time period for impacts	Population	Treatment group mean	Regression-adjusted difference between the treatment and estimated counterfactual (standard error)[a]	Percentage difference[a]	*p*-value[b]
Estimates limiting the sample to prevent sample addition, except for newborns, after the first baseline or intervention quarter[c]							
Quality-of-care processes (1)	Well-child visit (binary [yes or no]/beneficiary/year)	One-year period ending in I11 for Cohort 1, I9 for Cohort 2, I8 for Cohort 3, and I6 for Cohort 4	All observable Medicaid beneficiaries ages 3 to 6 years (36 to 72 months) assigned to treatment practices	60.6	-4.6 (1.1)	-7.0%	>0.99
Quality-of-care outcomes (1)	Dental caries treatment (binary [yes or no]/beneficiary/year)	One-year period ending in I11 for Cohort 1	All observable Medicaid beneficiaries ages 3 to 5 years (36 to 59 months) assigned to treatment practices	7.8	1.7 (1.3)	26.8%	0.90
Service use (2)	Outpatient ED visits (#/1,000 beneficiaries/quarter)	I2 to I11 for Cohort 1, I2 to I9 for Cohort 2, I2 to I8 for Cohort 3, and I2 to I6 for Cohort 4	All observable Medicaid beneficiaries attributed to treatment practices	171.6	-4.5 (2.9)	-2.6%	0.06
	Outpatient ED visits (#/1,000 beneficiaries/quarter)	I2 to I11 for Cohort 1, I2 to I9 for Cohort 2, I2 to I8 for Cohort 3, and I2 to I6 for Cohort 4	All observable high-risk Medicaid beneficiaries attributed to treatment practices	197.3	-6.0 (5.1)	-2.9%	0.12
Estimates using two-sided statistical tests							
Quality-of-care processes (2)	Lead screening (binary [yes or no]/beneficiary/two years)	Two-year period ending in I11 for Cohort 1, I9 for Cohort 2, and I8 for Cohort 3	All observable Medicaid beneficiaries age 2 years (24 to 35 months) assigned to treatment practices	27.1	-4.8 (1.4)	-14.9%	0.001

Table V.7 *(continued)*

Secondary test definition				Results			
Domain (# of tests in domain)	Outcome (units)	Time period for impacts	Population	Treatment group mean	Regression-adjusted difference between the treatment and estimated counterfactual (standard error)[a]	Percentage difference[a]	*p*-value[b]
	Well-child visit (binary [yes or no]/beneficiary/year)	One-year period ending in I11 for Cohort 1, I9 for Cohort 2, I8 for Cohort 3, and I6 for Cohort 4	All observable Medicaid beneficiaries ages 3 to 6 years (36 to 72 months) assigned to treatment practices	63.3	-3.7 (1.0)	-5.5%	<0.001
Quality-of-care outcomes (1)	Dental caries treatment (binary [yes or no]/beneficiary/year)	One-year period ending in I11 for Cohort 1	All observable Medicaid beneficiaries ages 3 to 5 years (36 to 59 months) assigned to treatment practices	7.7	1.5 (1.2)	24.9%	0.193
Estimates dropping Cohort 4 practices							
Service use (2)	Outpatient ED visits (#/1,000 beneficiaries/quarter)	I2 to I11 for Cohort 1, I2 to I9 for Cohort 2, and I2 to I8 for Cohort 3	All observable Medicaid beneficiaries attributed to treatment practices	163.4	-2.9 (2.9)	-1.7%	0.16
	Outpatient ED visits (#/1,000 beneficiaries/quarter)	I2 to I11 for Cohort 1, I2 to I9 for Cohort 2, and I2 to I8 for Cohort 3	All observable high-risk Medicaid beneficiaries attributed to treatment practices	195.6	-3.5 (5.2)	-1.8%	0.25

Sources: Analysis of the Ohio Medicaid claims received from the Ohio Colleges of Medicine Government Resource Center.

Notes: The results for each outcome are based on a difference-in-differences regression model, as described in the text. Estimates are calculated for Medicaid beneficiaries who are observable in the relevant time period: that is, beneficiaries who were alive, enrolled in Ohio Medicaid, had no third-party coverage, and were ages birth to 18 years. We defined *high-risk* as beneficiaries who were classified as "children with complex chronic disease" or "children with noncomplex chronic disease" based on the Pediatric Medical Complexity Algorithm (Simon et al. 2014).

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

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

[c] We allowed sample addition for beneficiaries born during the baseline or intervention periods.

ED = emergency department; HCIA = Health Care Innovation Award; I = intervention quarter; UHC = University Hospitals of Cleveland.

E. Consistency of impact estimates with implementation findings

Based on implementation findings, it is plausible that the program had no impact on the quality-of-care processes in our study. UHC reported that, by the end of the award, 90 percent or more of participating practices had met their goals for ordering lead tests and scheduling well-child visits for children ages 3 to 6, but UHC did not report its thresholds for meeting goals on these measures. It did report that most practices (94 percent) already met the threshold for well-child visits before the intervention, suggesting very little room for improvement for this measure during the intervention period.

In addition, the measures used, and reported, by UHC were defined only as ordering and scheduling these services regardless of whether they were completed. In contrast, the measures we used for our impact analysis, which were constructed using claims and encounter data, were defined as completing the services. There are many additional steps and potential barriers to completing a screening test or attending a well-child visit after simply ordering or scheduling those services; thus, improvements in UHC's own self-monitoring process measures might not have been detected as improvements in completing lead testing or well-child visits in our impact analysis due to issues such as dropped appointments and tests not completed (and thus not billed).

Nevertheless, the secondary tests suggested that the unfavorable impact estimates on the two quality-of-care process measures—although not substantively large—were statistically significant. Although the implementation evidence provides many plausible explanations for why the program might not have affected completion of lead testing or well-child visits at all, we see no plausible mechanism through which the intervention could have *worsened* performance on these measures. We therefore have a dilemma because we find that the quantitative estimates are inconsistent with the implementation evidence.

Similarly, we do not believe the primary test result for the quality-of-care outcome (a greater decline in dental caries treatment among comparison practices than treatment practices) is plausible based on the implementation findings. UHC reported that no practices met its goal for fluoride varnish application when programs began to participate in the award, but that this had increased to 90 percent of practices by the end of the award. Although UHC did not report the threshold for meeting its program goal, this result still suggested a substantial increase in fluoride varnish application for children who would have been included in our outcome measure. Although it is plausible that at the time of fluoride varnish application providers might detect more caries and refer children for treatment (thus counteracting other reductions in the rate of dental caries treatment), our impact estimate assessed dental caries treatment in an older age group—at least one to two years *after* they were eligible to receive the fluoride varnish treatment. This makes it unlikely that our estimates represent the effect of fluoride varnish application leading to the detection of caries in children.

Finally, we believe that the impact estimates for service use, specifically ED visits, are plausible given the implementation evidence. That is, it is plausible that UHC did not have substantively large or statistically significant reductions in ED visits for children in our study. Enhanced telephone triage was the only ED avoidance program service available to children in

participating practices throughout the award period. Although UHC reported a substantial volume of users for this service (30,000 calls per award year), we do not know what percentage of those calls were for children enrolled in Medicaid because the service was available to all children in the practice, and UHC was unable to report volumes based on insurance type. In addition, UHC noted that most participating practices already had some form of after-hours telephone triage before they participated, so the new enhancements to telephone triage services might not have been a significant enough change to produce an impact. There were significant delays in implementing the telemedicine hubs and, after implementation, volume for the hubs was low and not specific to children at participating practices. UHC also implemented two additional services to reduce ED visits: nurse case managers and an after-hours urgent care clinic. These services were also not specific to the participating practices, and their later implementation might have limited any potential impact on findings in our study.

F. Conclusions about program impacts, by domain

Based on all evidence currently available, we determined that we could not draw definitive conclusions about program impacts on patients' outcomes for the quality-of-care processes or quality-of-care outcomes domains. We conclude that the program did not have a substantively large impact on service use. Table V.8 summarizes our conclusions and their support.

For the two quality-of-care process measures of lead screening and well-child visits, the primary test results were neither significant (testing for favorable effects) nor substantively large. However, when we used two-sided statistical tests, we found statistically significant unfavorable impacts for both measures. There are a number of concerns with these results. First, we do not believe the program could have genuinely worsened performance on these measures. Second, as is clear from the unadjusted means (Table V.5), even though we matched on practice-level means of the quality-of-care processes during the baseline period, there was some imbalance between the treatment and comparison groups at the beneficiary level at baseline for lead screening. This could occur, for example, because of differences in the proportion of beneficiaries at each practice who were in the eligible population for the measure. This imbalance raised concerns about the comparison group for lead screening specifically but also suggests that the treatment and comparison groups could have had underlying unobservable differences related to quality of care processes more generally, and we might not have adequately captured these differences with our observable measures when we matched at the practice level. Third, we were unable to confirm a key assumption of the difference-in-differences model—namely, that the trends in the outcomes during the baseline period were similar between the treatment and comparison groups. We could not assess this because the quality-of-care process measures are estimated over one- or two-year periods rather than over distinct quarters as is the case for the ED visit rates, so we have only one baseline measurement for each quality-of-care process measure—making it impossible to calculate a trend for either measure. All of these concerns lead us to question whether the comparison group adequately reflects the counterfactual for these outcomes—that is, whether the comparison group reflects the experience the treatment group would have had without HCIA funding—and we are unable to draw conclusions for the quality-of-care processes domain.

The single estimate for dental caries treatment in the quality-of-care outcomes domain was substantively large in the unfavorable direction. When we used two-sided statistical tests, we did not find a statistically significant unfavorable impact. However, this measure shares many of the same concerns as the lead screening and well-child visit measures: (1) the implementation evidence does not suggest a plausible mechanism by which the program could worsen the outcome (that is, causing the rate of dental caries treatment to decline more slowly in the treatment group than the comparison group); (2) the treatment and comparison groups were imbalanced at baseline on the outcome, both at the beneficiary level (Table V.5) and, in this case, at the practice level (Table V.3); and (3) we were unable to confirm assumptions about the trends in the rates during the baseline period. These concerns lead us to believe the comparison group might not adequately reflect the counterfactual for this outcome either. Therefore, we are unable to draw definitive conclusions for the quality-of-care outcomes domain.

For service use, which we measured through the outpatient ED visit rate, we found no substantively large or statistically significant effects for the individual or combined estimates in the domain. The secondary test results supported our findings from the primary tests, and the implementation findings supported the primary test results. In addition, the analytic challenges we experienced for the other two evaluation domains did not affect the service use domain. Specifically, the treatment and comparison groups were well balanced at baseline on this outcome, both on rates and trends (Tables V.3 and V.6) in the baseline period, which supported a key assumption of our difference-in-differences model. We believe that the comparison is a valid counterfactual for the service use domain because the types of initiatives that might affect ED visits (for example, extended office hours or care management for frequent users of the ED) are largely distinct from the types of interventions that would affect the quality-of-care measures (for example, EHR prompts for overdue well-child visits or screenings). Therefore, even though we believe the comparison group is limited for the quality-of-care measures, we believe the comparison group is valid for service use and we can draw definitive conclusions about service use. Our conclusion for this domain is that the program had no substantively large effect on service use.

Table V.8. Conclusions about the impacts of UHC's HCIA program on patients' outcomes, by domain

Domain (individual outcomes within domain)	Conclusion for domain	Evidence supporting conclusion					
		Primary test result(s)	Statistically significant (one-tailed test)?	Substantively large?	Power to detect a substantively large difference	Primary test result(s) plausible given secondary tests?	Primary test result(s) plausible given implementation evidence?
Quality-of-care processes	No conclusion						
• Lead screening		-14.9%	No	No	98.6%	No	No
• Well-child visits		-5.5%	No	No	>99.9%	No	No
• Combined		-10.2%	No	No	>99.9%	No	No
Quality-of-care outcomes	No conclusion						
• Dental caries treatment		24.9%	No	Yes	22.4%	Yes	No
Service use	No substantively large effect						
• Outpatient ED visits (all)		-1.5%	No	No	>99.9%	Yes	Yes
• Outpatient ED visits (high risk)		-1.3%	No	No	97.4%	Yes	Yes
• Combined		-1.4%	No	No	99.9%	Yes	Yes

ED = emergency department; HCIA = Health Care Innovation Award; UHC = University Hospitals of Cleveland.

VI. DISCUSSION AND CONCLUSIONS

UHC used its $12.8 million HCIA to implement multiple program components with an overall goal to transform the delivery of health care for children enrolled in Medicaid in northeastern Ohio. This included interventions to increase use of preventive care through quality improvement at primary care practices, reduce avoidable outpatient ED visits, reduce hospital readmissions, coordinate care for children with complex health needs, and integrate behavioral health services within primary care practices. We chose to evaluate two components—practice facilitation to increase use of preventive services and efforts to reduce ED visits—because a rigorous impacts design was feasible for these components and because UCH considered practice facilitation to be its core intervention.

A. Conclusions about program implementation

For the primary care practice facilitation component, UHC worked with primary care practices, comprising both UHC-employed and independent clinicians, for care improvement activities and a linked financial incentive program. Practices focused on improving 11 quality measures over the course of the award. UHC began this component in the first year of the program, as planned, and, through UHRCC, enrolled 32 practices (exceeding its goal of 28) that provided care for more than 71,000 child Medicaid beneficiaries (exceeding its goal of 68,000). UHC made incentive payments, earned by the practices, at six-month intervals as planned.

UHC had mixed results in implementing the different services that comprised the ED avoidance component. UHC implemented the telephone triage service as planned and reported significant use (more than 90,000 calls during the award). UHC then added the ED nurse case manager service in the second year of the award, but did not report implementation measures for the service. UHC faced significant delays with the telemedicine hubs and reported use after implementation in the second and third years of the award (651 visits). In response to this, UHC added the after-hours urgent care clinic in the third year, which saw more use (1,758 visits).

B. Conclusions about program impacts

Although we used a rigorous design to estimate the impacts of both the practice facilitation and ED avoidance components, we were able to draw definitive conclusions only for the ED avoidance component. For this component, we found that the program did not have a substantively large impact on outpatient ED visits (a measure that we grouped into the broader domain of service use to facilitate comparisons across HCIA awardees at the domain level). For the practice facilitation component, our primary quantitative results suggested that the UHC program had unfavorable impacts on quality-of-care measures—reducing the rates at which Medicaid-enrolled children (within select age groups) received well-child visits and lead screening, relative to a matched comparison group, and achieving smaller reductions than the comparison group in the rate of treatment for dental caries. We viewed these findings as implausible because, based on the program design and implementation, we see no reason to expect the program could decrease use of preventive care or increase the likelihood of needing treatment for cavities (the program aimed to increase use of fluoride varnish which, in theory, should reduce cavities). Rather, we expect that these primary estimates are biased, reflecting

limitations in the comparison group for these select measures. Specifically, we anticipate (though were unable to test due to data limitations) that the comparison and treatment practices were on different trajectories during the baseline period for these measures, undermining the key assumption behind the difference-in-differences model. That model assumes that baseline differences reflect the differences we would expect to see in the intervention period but for the effects of the intervention itself. In contrast, we were able to verify this parallel trends assumption for outpatient ED visits, which is why we have greater confidence in those results.

Our interpretation of the results suggests that the comparison group represents the counterfactual for service use but not for the quality-of-care measures; that is, the comparison group experience reflects what would have happened to the treatment group without HCIA funding for outpatient ED visits, but not for the other outcomes because, for them, the treatment and comparison baseline trajectories might have differed. We believe this interpretation is reasonable because the types of initiatives (such as extended office hours or care management for frequent users of the ED) that might affect ED visits are largely distinct from the types of interventions that would affect the quality-of-care measures (such as EHR prompts for overdue well-child visits or screenings).

C. Possible explanations for why UHC's program did not reduce ED visits

Although UHC added multiple services to meet its goal of reducing outpatient ED visits, we have identified five possible explanations for why the UHC program did not have substantively large reductions in outpatient ED visits over the time period examined.

1. **The enhanced nurse triage system might not have been sufficiently distinct from earlier triage systems, or used extensively enough, to make a difference in ED visit rates.** Most practices likely had some form of after-hours phone advice that existed before the intervention. The HCIA paid for two enhancements in particular: (1) allowing a nurse to call in certain medications for children with low-acuity conditions per standing orders of a physician and (2) requiring nurses to consult with a pediatrician before referring a patient to the ED. These changes might have had only modest impacts if, for example, the pediatricians did not refer substantially fewer patients to the ED than the nurses would have on their own, or if some practices had previous after-hours phone lines—replaced by the UHC triage line during the award period—that were already staffed by physicians. In addition, the nurses' authority to call in medications might have had limited impact if the covered low-acuity conditions represented only a small proportion of ED visits, or, even if the covered low-acuity conditions did represent a large share of ED *visits*, if those conditions were not equally represented in phone calls to the triage line because patients and their caregivers were unaware that they could get a prescription by phone. Only a small fraction (less than 3 percent) of calls to the triage system resulted in prescriptions over the phone.
2. **The after-hours clinic was not open for most of the primary test period.** The after-hours clinic received a significant volume of patients when it opened in late 2014. However, the clinic was open for only 20 of the 35 months of the primary test period. It is possible that the longer outcome period diluted any impact the after-hours clinic had on ED visits was diluted

by. In contrast, the other intervention components were active for all or most of the primary test period. In addition, the after-hours clinic was available to any patient and was not limited to the patient population included in the treatment group for our evaluation. Therefore, some proportion of patients in our treatment group might not have been exposed to this intervention component, so our estimates capture only the effects of the children in the treatment group who were exposed to the after-hours clinic.

3. **The telemedicine hubs saw minimal patient volume.** The telemedicine hubs were an innovative approach to reducing ED visits, but in practice they were not used very much and therefore could not have been expected to substantially reduce ED visits. UHC faced a variety of challenges in establishing these telemedicine hubs. First, it took longer than expected to find suitable physical locations for the hubs in neighborhoods with high ED visit rates. Second, when the hubs were established, UHC had difficulties gaining community interest in using them. The hubs represented a very different method of interacting with primary care providers and there was little community member buy-in, despite UHC's outreach efforts. Third, the company that operated the telehealth hubs ceased operations during the intervention period. UHC had briefly had some success in higher visit rates to the hubs after the telehealth company placed hubs in RiteAid convenience stores rather than stand-alone hubs. However, UHC was not able to carry through with this approach for long because the telehealth company ceased operations.

4. **The case managers might not have had a sufficiently large patient volume to meaningfully influence care patterns.** Unfortunately, no metrics on patient volume for this component were available. It is also possible that only a small portion of the treatment group included in the impact evaluation might have been exposed to this component.

5. **Patients in some of the treatment practices could not be exposed to some of the intervention components.** The case management services were offered to patients who went to the ED at the main UHC campus, and the after-hours clinic was also opened on the main campus. Therefore, patients going to practices located far from the UHC campus could not reasonably be expected to be exposed to these interventions. Similarly, the two telemedicine hubs were opened in neighborhoods with high ED utilization rates and that lacked access to other ED alternatives, such as urgent care clinics. These hubs, even if they had received high volume, would not likely have affected ED visit rates for patients living outside of these neighborhoods.

D. Lessons for future impact evaluations

Although we were unable to draw definitive conclusions about the practice facilitation component, there are several lessons for future evaluations gained from our experience with trying to evaluate this component. Our findings highlight the importance of building in robustness checks, including tests of model assumptions (such as testing for parallel trends in the baseline period in a difference-in-differences model), and of using a clear evaluation framework (see Appendix 3) to synthesize impact and implementation findings before drawing conclusions. In this evaluation, adhering to our prespecified decision rules for estimating impacts enabled us to separate credible findings from those that could not be relied upon because they did not pass robustness checks. Although we could test the parallel trends assumption for one key outcome

measure—outpatient ED visits—we were unfortunately unable to do so, due to the length of the measurement periods and constraints in the data, for other measures. The key constraint was that examining trends in quality of care measures that are defined over a year or more requires multiple years of baseline to identify trends. We were unable to use the Medicaid claims over a sufficiently long baseline period to do that type of trend analysis. We were able to conduct trend analyses for outpatient ED visits because we could define those measures quarterly and so examine whether the outcomes moved in similar directions over the four baseline quarters (which they did).

Our findings also highlight common challenges identifying strong comparison groups when factors that influence outcomes might not be observable. This is particularly true for measures defined over long time periods (for example, one or two years) because establishing a baseline trend can be difficult or, with limited baseline data, impossible. Some options for improving impact evaluation in situations like this include (1) randomly assigning practices as treatment or control to ensure there are no systematic difference between practices; (2) using longer baseline periods, if possible; and (3) developing as comprehensive a set of variables as possible for matching, including variables that might pick up baseline trends in practice quality-of-care initiatives (such as changes in office hours over time) that could affect outcomes.

E. Lessons for organizations considering similar ED avoidance interventions

In today's health care environment, with many initiatives and new incentives to improve the value of care over the volume of care, other provider organizations serving pediatric populations might be interested in implementing interventions similar to those UHC used to try to reduce outpatient ED visits. We draw two lessons from the impact and implementation findings for service use that can inform future program design and evaluation. First, the telemedicine hubs sought to provide an alternative to ED care in neighborhoods with high ED use, but the hubs saw minimal patient volume. Patients who are already leaving their homes to visit an outside care setting—as patients had to in order to visit the UHC telemedicine hubs—might prefer to see a health care professional in person.

Second, we found that the nurse triage service did not generate meaningful changes for ED visit rates for the treatment group included in the impact evaluation. Practices could consider ways to make larger changes to their triage services from the status quo and/or work to ensure that these new services are widely used among the target population to increase prospects of success.

In conclusion, our implementation findings suggest UHC largely succeeded in implementing the components of its program included in our study. The lack of observed impacts on outpatient ED visits illustrates the challenges inherent in efforts to change long-standing patterns of service use and to implement innovative programs to address those patterns. Policymakers, clinicians, and payers can use these findings to inform efforts to improve health care for children in Medicaid.

REFERENCES

Centers for Disease Control and Prevention. "National Health Interview Survey. 2012 Data Release, Person File." Atlanta, GA: CDC, 2014. Available at https://www.cdc.gov/nchs/nhis/nhis_2012_data_release.htm. Accessed March 24, 2017.

Christensen, A.L., J.S. Zickafoose, B. Natzke, S. McMorrow, and H.T. Ireys. "Associations Between Medical Homeness and Health Care Utilization Among Publicly Insured Children." *Academic Pediatrics,* vol. 15, 2015, pp. 267–274.

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

Mathews Burwell, Sylvia. "2015 Annual Report on the Quality of Care for Children in Medicaid and CHIP." Washington, DC: U.S. Department of Health and Human Services, February 2016. Available at https://www.medicaid.gov/medicaid/quality-of-care/downloads/2015-child-sec-rept.pdf. Accessed April 18, 2017.

Medicaid and CHIP Payment and Access Commission. "MACStats: Medicaid and CHIP Program Statistics. June 2014." Baltimore, MD: Centers for Medicare & Medicaid Services, 2014. Available at https://www.macpac.gov/wp-content/uploads/2015/01/2014_06_MACStats.pdf. Accessed March 24, 2017.

Meropol, Sharon B., Nicholas K. Schiltz, Abdus Sattar, Kurt C. Stange, Ann H. Nevar, Christina Davey, Gerald A. Ferretti, Diana E. Howell, Robyn Strosaker, Pamela Vavrek, Samantha Bader, Mary C. Ruhe, and Leona Cuttler. "Practice-Tailored Facilitation to Improve Pediatric Preventive Care Delivery: A Randomized Trial." *Pediatrics,* vol. 133, no. 6, 2014, pp. e1664–e1675.

Pahel, Bhavna T., R. Gary Rozier, Sally C. Stearns, and Rocio B. Quiñonez. "Effectiveness of Preventive Dental Treatments by Physicians for Young Medicaid Enrollees." *Pediatrics,* vol. 127, no. 3, 2011, pp. e682–e689.

Simon, Tamara D., Mary Lawrence Cawthon, Susan Stanford, Jean Popalisky, Dorothy Lyons, Peter Woodcox, Margaret Hood, Alex Y. Chen, and Rita Mangione-Smith. "Pediatric Medical Complexity Algorithm: A New Method to Stratify Children by Medical Complexity." *Pediatrics,* vol. 133, no. 6, 2014, pp. e1647–e1654.

Zickafoose, Joseph S., Lisa R. DeCamp, and Lisa A. Prosser. "Association Between Enhanced Access Services in Pediatric Primary Care and Emergency Department Utilization: A National Parent Survey." *Journal of Pediatrics,* vol. 163, no. 5, 2013, pp. 1389–1395.

Zickafoose, Joseph, Brenda Natzke, Boyd Gilman, Catherine DesRoches, and Lorenzo Moreno. "Findings for University Hospitals of Cleveland Rainbow Babies and Children's Hospital." In Moreno, Lorenzo, Boyd Gilman, Greg Peterson, Catherine DesRoches, Sheila Hoag, Linda Barterian, Laura Blue, Katherine Bradley, Emily Ehrlich, Kristin Geonnotti, Lauren Hula, Keith Kranker, Rumin Sarwar, KeriAnn Wells, Joseph Zickafoose, Sandi Nelson, Kate Stewart, and Frank Yoon. "Evaluation of Health Care Innovation Awards (HCIA): Primary Care Redesign Programs." Second Annual Report to CMS. Princeton, NJ: Mathematica Policy Research, March 2016.

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

CHAPTER 4 SUPPLEMENTAL MATERIAL

METHODS FOR THE EVALUATION OF UNIVERSITY HOSPITALS OF CLEVELAND

Andrea Wysocki, Joe Zickafoose, Greg Peterson, Kate Stewart, Laura Blue, Keith Kranker, Brenda Natzke, Boyd Gilman, and Sheila Hoag

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

I. INTRODUCTION

This supplement to Chapter 4 provides methods detail for our evaluation of the Health Care Innovation Award (HCIA) received by the University Hospitals of Cleveland (UHC). As we describe in Chapter 4 of this report, we conducted an impact evaluation for two components of this HCIA-funded program: primary care practice facilitation and emergency department (ED) avoidance. In this supplemental material, we describe in detail how we constructed the treatment and comparison groups. We also describe the program's theory of action—that is, how the program could be expected to affect the study outcomes.

II. METHODS

A. Treatment group definition

The treatment group was defined as children (from birth to 18 years) enrolled in Medicaid (fee-for-service [FFS] or managed care) whom we attributed to practices participating in UHC's primary care practice facilitation component during the HCIA funding period and/or during a 6- to 12-month baseline period before the intervention began. To define the treatment beneficiaries in both periods, we first identified the providers who worked at participating practices and then attributed beneficiaries to practices based on the beneficiaries' visits with those providers.

1. Identifying providers at program-participating practices

To identify providers in all practices (except for the federally qualified health center [FQHC] locations), we purchased data from SK&A, a private vendor that collects national provider data. These data organize providers into practices, which was not feasible for us to do with only the Ohio claims data. We purchased SK&A data from 2012 to 2015 that covered practices in Ohio with at least one pediatric primary care provider. Because the SK&A data defined practices at the practice site level based on address, we used site-level information (including address) from UHC to flag the 45 non-FQHC treatment practice sites in the SK&A data. These 45 sites corresponded to 31 of the 32 practices UHC called its participants—all but the FQHC, which had 4 sites. We then identified pediatric providers who worked in these 45 locations from 2012 to 2015 using the SK&A data. We limited our list of providers to those with a pediatric primary care specialty. We opted to use SK&A to identify the participating providers instead of using information provided directly by UHC so that we could standardize the process to identify providers for both the treatment and comparison groups—eliminating a possible source of bias in our evaluation.

Having identified the participating practices' locations in the SK&A data, we collapsed the 45 non-FQHC practice sites into 33 sites that we used in the impact analysis. We used the rule that if two or more practice sites had 50 percent or more of their providers working at both (or all) of the sites, we combined the sites into one. This facilitated attributing beneficiaries, described later in this section.

For the four FQHC locations that participated in the program, we used the sites' organizational National Provider Identifiers (NPIs) to identify the sites in the Ohio Medicaid claims data. This left us with a total of 37 treatment practice sites used in the impact analysis. For

simplicity, we refer to these in the rest of the report as *practices,* although they do not correspond exactly to the 32 participating practices, as defined by UHC.

2. Defining cohorts of practice participants

As mentioned previously, practices joined the primary care practice facilitation component in waves. For the impact evaluation, we organized the practices into four cohorts, summarized in Table 1. We defined the 23 practices that joined by the start of the intervention period in January 2013 as Cohort 1 practices, and set the intervention start date for them to the start date of the practice facilitators—February 1, 2013. We defined the 7 practices that joined the intervention in July, August, or September 2013 as Cohort 2 practices and set the intervention start date for them to August 1, 2013. We defined the 4 practices that joined the intervention in January 2014 as Cohort 3 practices and set the intervention start date for them to January 1, 2014, and we defined the 3 practices that joined the intervention in July 2014 as Cohort 4 practices and set the intervention start date for them practices to July 1, 2014.

For each cohort, we defined a baseline (pre-intervention) period relative to the cohort's intervention start date. For Cohort 1, the baseline period was the 6 months before the practices joined the intervention, and for Cohorts 2, 3, and 4, the baseline period was the 12 months before the practices joined the intervention. Cohort 1's baseline period was only 6 months because the Ohio Medicaid data did not enable us to calculate practice- or provider-level measures for the period before August 2011, so we used Ohio Medicaid data starting in August 2011 for attribution. Thus, the earliest baseline month for which we could attribute children based on a one-year service period was August 2012, using data from August 2011 to July 2012.

Although the awardee received a no-cost extension to operate the program through March 2016, we ended the analysis period by December 31, 2015, the last date for which we had claims data. Specifically, the intervention analysis period was defined as the time from the date the practice joined the intervention to either October 31, 2015 (for Cohorts 1 and 2), or December 31, 2015 (for Cohorts 3 and 4). For Cohorts 1 and 2, October 31, 2015, was the end date for the last full quarter of data available (because the intervention start dates were in February and August, meaning that the next quarter would have ended January 31, 2016—and we did not have data beyond December 2015.) For Cohorts 3 and 4, December 31, 2015, was the end of the last full quarter of available data.

Table II.1. Baseline and intervention dates for practice cohorts

Cohort	Number of practices	Baseline start	Baseline end	Intervention start	Intervention end
1	23	8/1/2012[a]	1/31/2013	2/1/2013	10/31/2015
2	7	8/1/2012	7/31/2013	8/1/2013	10/31/2015
3	4	1/1/2013	12/31/2013	1/1/2014	12/31/2015
4	3	7/1/2013	6/30/2014	7/1/2014	12/31/2015

Note: The last full quarter for Cohorts 1 and 2 ended on October 1, 2015. The next full quarter to measure outcomes for these cohorts would have ended on January 31, 2016, but our data ran only through December 31, 2015.

[a] Cohort 1's baseline period was only six months due to data limitations that did not allow us to calculate practice- or provider-level measures for the period before August 2011. The earliest baseline month for which we could attribute children based on a one-year service period was August 2012, using data from August 2011 to July 2012.

3. Assigning beneficiaries to practices

Having identified the intervention practices, we constructed the treatment group of beneficiaries in three steps:

1. In each baseline and intervention month, we attributed children using a two-stage attribution approach. In the first stage, we attributed children to the primary care practice whose pediatric providers provided the plurality of their well-child visits in the past 12 months. If there was not a plurality, we attributed children to the practice at which they received their most recent well-child visit. In the second stage, for children who did not have any well-child visits in the past 12 months, we attributed children to the practice that provided the plurality of their visits for other evaluation and management codes in the past 12 months. If there was not a plurality, we attributed children to the practice at which they received their most recent evaluation and management visit. We were unable to attribute children who did not have any well-child or evaluation and management visits to a practice in the 12 months before the measurement quarter. We used the list of pediatric providers in practices from the SK&A data to attribute children to non-FQHC treatment practices, and used organizational NPIs to attribute children to the FQHC treatment practices.
2. In each baseline and intervention period, we assigned each child to the first treatment practice to which he or she was attributed in that period (baseline or intervention), and continued to assign him or her to that practice for all quarters in the period. This rule followed an intent-to treat framework and ensured that, during the intervention period, beneficiaries did not exit the treatment or comparison groups at different rates, which might bias impact estimates. The definition for the baseline period corresponds to that of the intervention period so that, across the two periods, interpretation of the population changes over time should be comparable.
3. We applied additional restrictions to refine the analysis sample in each quarter. A beneficiary assigned to a treatment practice in a quarter was included in the analysis sample for that quarter if he or she was alive, enrolled in Ohio Medicaid, had no third-party coverage, and was age birth to 18 years.

B. Comparison group definition

The comparison group was defined as children (from birth to 18 years) enrolled in Medicaid (FFS or managed care) and attributed to matched comparison practices in each of the baseline and intervention quarters. We selected comparison practices that were similar to the treatment practices during the baseline period on factors that can influence patients' outcomes and the decision to participate in the program.

We identified the 109 matched comparison practices in five steps:

1. Define a pool of potential comparison practices.

We used SK&A data to develop a list of potential non-FQHC comparison practices in Ohio. We also developed a list of organizational NPIs for FQHCs in Ohio from the Centers for Medicare & Medicaid Services (CMS) Integrated Data Repository and from the Ohio Medicaid claims data and verified in the publicly available National Plan and Provider Enumeration System NPI Registry. After developing the list of non-FQHC and FQHC practices throughout Ohio, we identified a pool of potential comparison practices in counties not considered part of the eight-county UHC service area. We excluded practices in the UHC service area because many of UHC's ED avoidance services were available to all children regardless of their primary care provider. Thus, these services were likely to have spillover effects for primary care practices in the area even if those practices did not participate in the program. Figure 1 shows the eight counties in the UHC service area (shaded dark blue) and the counties that we used to draw potential comparison practices (shaded light blue). We limited the potential comparison pool to practices within Ohio (rather than going out of state) because Medicaid policies—and how they have changed over time—vary substantially across states.

Importantly, we excluded the 34 counties considered part of the service area for the Research Institute at Nationwide Children's Hospital (NCH) HCIA program (shaded medium blue). Before the HCIA, NCH had established an accountable care organization (ACO) and engaged extensively with primary care providers in the region on quality improvement projects analogous to UHC's. A similar structure did not exist in the UHC service area before the HCIA-funded program began. Because one goal of UHC's award was to lay the foundation for a Medicaid ACO, practices in the NCH counties would not reflect the counterfactual—that is, the outcome the treatment group would have had in the absence of the HCIA-funded program—for the UHC treatment practices.

We did include comparison practices in the service area covered by the Akron Children's Hospital (ACH), a partner in NCH's HCIA. The NCH HCIA included developing a pediatric Medicaid ACO led by ACH, but this component of the award did not affect the delivery of primary care services during the award period. The other components of the NCH award implemented at ACH would not be expected to have significant impacts at the primary care practice level on the outcomes we used in this study. In addition, the service area for ACH is similar to the UHC service area across many other characteristics, including the location in Ohio, demographics of Medicaid beneficiaries, an affiliation with a large health system among many physician practices, and lack of a large-scale ACO for children in the region at the start of the program. Although two UHC treatment practices were located in Summit

County, which was considered part of the ACH service area, we still included this county in the potential comparison pool because the possibility of spillover from UHC's other intervention components is low in this county due to distance from UHC and proximity to ACH. We dropped practices located in the Summit County zip codes in which the treatment practices were located from our pool, but otherwise included practices located in other zip codes in Summit County.

We chose to include counties in the southwestern part of Ohio that are in the primary service area of Cincinnati Children's Hospital Medical Center (CCHMC). We considered excluding counties in this region because CCHMC began developing an ACO structure during the period of UHC's HCIA. However, we felt there was a low risk of biasing our impact estimates by including practices from this region because the CCHMC ACO developed in parallel to UHC's intervention and was not yet mature like the NCH-led ACO. In addition, we did not have a sufficient number of potential comparison practices if we excluded those counties in the large metropolitan Cincinnati region, which presented a greater risk to our design than including the counties to develop our potential comparison practice pool.

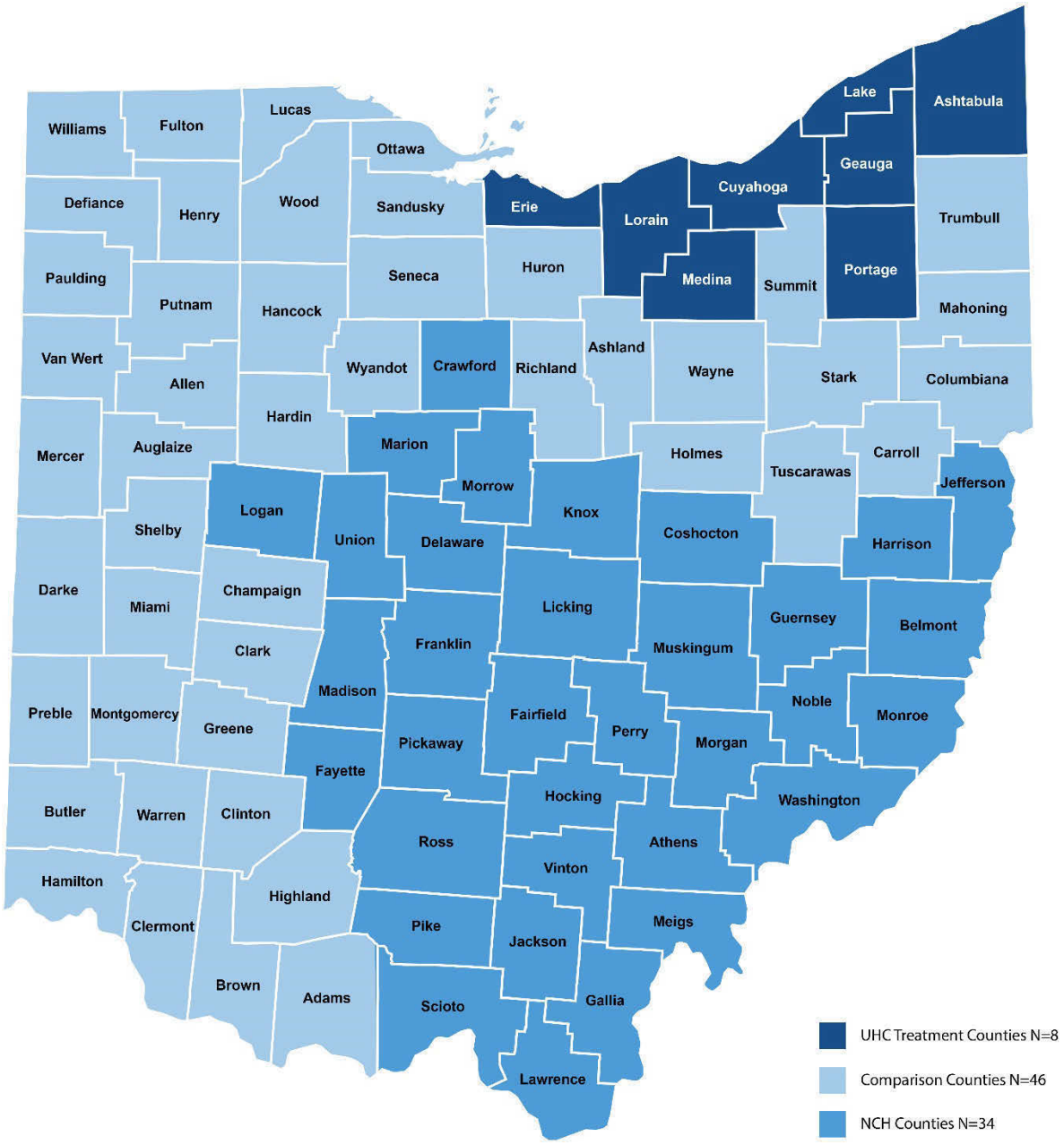
2. Assign beneficiaries to comparison practices.

When assigning Medicaid beneficiaries to the practices, we used the same attribution and practice assignment logic that we used for the treatment practices, as described previously.

3. Develop matching variables.

We developed matching variables defined at the start of the intervention for all treatment and potential comparison practices. These variables were defined relative to the start date of each cohort (so that, for example, a potential comparison practice might be included in the comparison pool four times: once for each cohort's start date). These matching variables included characteristics of the practice (for example, the number of pediatric primary care providers in the practice and whether a hospital or health system owned the practice); and characteristics of Medicaid beneficiaries assigned to the practices (for example, mean number of assigned beneficiaries and utilization in the baseline period).

Figure II.1. Ohio counties included for treatment and potential comparison practices



Note: Two UHC treatment practices are located in Summit County, which is not considered part of the UHC service area.

NCH = Nationwide Children's Hospital; UHC = University Hospitals of Cleveland.

4. **Restrict the potential comparison pool to limit the influence of outliers on matching.**

   Before matching, we narrowed the potential comparison pool of practices by excluding practices with characteristics not observed among the treatment group. We dropped potential comparison practices from the pool that were unlike treatment practices because they had very high or low numbers of attributed beneficiaries in the baseline period, low proportions of attributed beneficiaries enrolled in managed care, high proportions of attributed beneficiaries who were categorized as high-risk (defined as beneficiaries classified as "children with complex chronic disease" or "children with noncomplex chronic disease" based on the Pediatric Medical Complexity Algorithm [Simon et al. 2014]), or an ED visit rate more than twice the mean among the treatment practices. After we applied these restrictions, 198 non-FQHC and 83 FQHC potential comparison practices remained available for matching.

5. **Conduct matching.**

   We used distance-based optimal matching (Stuart 2010) to select 109 comparison practices from the comparison pool that were similar to the 37 treatment practices on the matching variables. We matched each treatment practice to one or more comparison practices, with the aim of minimizing the average difference between matched pairs on the matching variables. We did not allow the same comparison practice to match to more than one treatment practice. This matching approach, however, does not ensure that each comparison practice matches exactly to its treatment practice on all matching variables. We specified that comparison practices had to match exactly to the treatment practices on two characteristics: whether the practice was an FQHC and, for the FQHCs, whether the practice participated in the CMMI FQHC Advanced Primary Care Practice Demonstration, because one of the FQHC treatment sites participated in this demonstration.

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

III. Theory of action

Based on extensive review of UHC's program activities and goals, we developed a theory of action to depict the mechanisms through which program administrators expected the program to impact the study domains. UHC expected that its HCIA-funded program would improve the selected outcomes for Medicaid-enrolled children through two pathways: (1) practice facilitation to promote the delivery of preventive care services and (2) ED avoidance.

Pathway 1: Practice facilitation.

This component can affect all children receiving preventive care at participating practices. Practice facilitators provide continuous feedback on quality measure performance to promote quality improvement. Planned mechanisms of this pathway include the following:

1. **Program leaders obtain buy-in from practices for participating, including by offering financial incentives for performance on specified quality measures.**

- Clinicians and other practice staff are motivated by their interest in improving care; access to additional resources (for example, clinicians' decision tools and practice organizational tools) to support family education and practice organization; and program financial incentives.

2. **Practice facilitators provide feedback on practice performance.**
   - Facilitators visit practices weekly, perform a small number of chart audits to assess performance on quality measures, share measure results with providers and staff, and discuss opportunities for improvement. Clinicians and other practice staff are motivated to improve through comparisons with their own prior performance and how other practices performed.
3. **Clinicians and other practice staff identify and act on the feedback.**
   - Facilitators' feedback helps clinicians and other practice staff identify opportunities for improvement—both changes that practices can make on their own and changes for which they need additional assistance. As needed, facilitators provide practices with resources such as patient educational materials, clinician decision tools, and practice organizational tools.
4. **Practices' efforts increase the delivery of age- and risk-indicated preventive services to children served by the practice.**
   - Children are more likely to receive well-child visits, lead screening, and dental fluoride application.
5. **Increases in dental fluoride application lead to decreases in early childhood caries.**
   - Fewer children require dental caries treatment.
6. **The program assesses practices' performance and makes incentive payments.**
   - Facilitators perform monthly audits on larger numbers of charts than the number completed at the weekly audits to assess progress toward quality targets and eligibility for incentive payments. The program pays incentives to practices based on performance every six months. Incentive payments aimed to help offset the costs of practice staff time spent on quality improvement and encourage clinicians to make further improvement efforts.

Pathway 2: ED avoidance.

Planned mechanisms of this pathway include the following:

1. **Program services provide caregivers of children with alternatives to ED care.**
   - Enhanced telephone triage enables caregivers with urgent concerns to speak with a nurse, and a physician as needed, when primary care practices are not open.
   - The nurse and physician can advise worried caregivers about managing an illness at home and seeking care in the child's primary care office when an ED visit is not necessary. In addition, because the nurse could call in certain medications for children

with low-acuity conditions per standing orders of a physician, some caregivers can avoid having to visit another provider (such as the ED) for necessary but straightforward care.

- Telemedicine hubs provide an option for children to receive care when primary care practices are not open.
- In-person medical attendants and on-call pediatricians can provide care in a community setting for less urgent conditions using telemedicine tools, avoiding the need for the family to seek care in the ED.
- An after-hours urgent care clinic on the UHC main campus provides an additional option for children to receive care for low- to moderate-acuity conditions when primary care practices are not open.
- Nurse practitioners can provide care in an office setting after-hours, avoiding the need for the family to seek care in the ED.

2. **The program promotes awareness of the new services through multimodal outreach in the community.**
   - Caregivers learn about alternatives to ED use from billboards, bus advertisements, and other sources.
3. **Increased caregiver awareness of alternatives to ED care leads to increased use of those alternative services.**
4. **Increased use of the new ED alternative services leads to decreased reliance on the ED.**
   - This results in a decrease in ED visits, especially unnecessary or preventable visits, among the target population.
5. **The program identifies children with ongoing high ED use despite ED alternatives, and nurse case managers contact the children's families to identify and address barriers to receiving care in a setting other than the ED.**
   - This leads to a further decrease in ED visits.

REFERENCES

Simon, Tamara D., Mary Lawrence Cawthon, Susan Stanford, Jean Popalisky, Dorothy Lyons, Peter Woodcox, Margaret Hood, Alex Y. Chen, and Rita Mangione-Smith. "Pediatric Medical Complexity Algorithm: A New Method to Stratify Children by Medical Complexity." *Pediatrics,* vol. 133, no. 6, 2014, pp. e1647–e1654.

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