Milbank Memorial Fund

REPORT

Assessing the Effects of Primary Care Transformation:

Emerging Themes and Practical Strategies to Strengthen the Evidence

by Nancy McCall, ScD, and Kristin Geonnotti, PhD

Table of Contents

Acknowledgments

The authors are very grateful to the eight Multi-State Collaborative members who agreed to participate in this study to help build a solid foundation for strengthening future primary care evaluations.

They would also like to thank Deborah Peikes, PhD, MPA (Mathematica Policy Research), for her perspectives and insights in developing this report, and Lisa Dulsky Watkins, MD (Milbank Memorial Fund), for her guidance and collaboration throughout the project.

Message from the President

Dear Reader,

There is ample evidence internationally—much of it analyzed by Barbara Starfield on the pages of *The Milbank Quarterly*—that indicates that a high-quality, low-cost, and socially inclusive health care system must be built on a strong foundation of primary care services.

Here in the US, we seem resolute in ignoring this evidence. We spend a smaller proportion of our health care expense on primary care than any other developed economy—and we value primary care services less, relative to other medical specialties. We pay primary care practices to provide isolated visits, not to provide team-based, patient-centered care.

Critical to reversing these policies is learning what constitutes high-quality primary care and how to increase its adoption. The 18 projects involved in the Milbank Memorial Fund's Multi-State Collaborative are doing this hard work: working with 1,775 practices and about seven million patients to align standards, payment mechanisms, and measurement and technical assistance across payers in an effort to improve the practice of primary care.

Just as the physician practices in the Collaborative need measurement and feedback to improve, so do the projects themselves. Are the projects making a difference? How? What could be improved in their execution? These questions can only be answered with credible project evaluations.

Some of the projects in the Collaborative have independently funded evaluations. Others have engaged in their own evaluations. To help them learn how to do their evaluations better and improve the effectiveness of their projects, the Fund engaged Mathematica Policy Research to "evaluate the evaluations." Regardless of the stated outcomes of the evaluations, how credible are they? The results follow.

Learning to do evaluations better is not about arcane arguments over measurement methodology. Better evaluations will result in more robust results and more effective learning—for the projects themselves and for anyone concerned with a sustainable health care delivery system in the US.

The eight members of the Multi-State Collaborative who offered up their evaluations for assessment are to be commended for their leadership. The question is not whether we transform primary care in the US, but how. As multi-payer primary care work matures and spreads, we hope this report helps make that needed transformation more likely.

Christopher F. Koller
President, Milbank Memorial Fund

Introduction

The Milbank Memorial Fund's Multi-State Collaborative (MC) is a working group of 18 states and regions actively engaged in multi-payer primary care transformation through the implementation of patient-centered medical home (PCMH) programs. These innovative efforts include payment reform and enhanced multidisciplinary support services. Each of the MC programs were early adopters of multi-payer primary care—their leaders made investments of time and resources before knowing what the outcomes would be.

Assessing their own programs has been an important component of the MC members' work—and each MC member has been conducting an evaluation of the effectiveness of its PCMH program. As the programs developed, they grew in complexity—as did the methods needed to evaluate them. The challenge was to ensure that promising results were not missed—or impacts overstated.

In order to understand the extent to which MC PCMH programs are improving outcomes for a core set of key health care utilization and spending measures, the Fund asked Mathematica Policy Research to develop this report. It analyzes eight of the 18 MC member evaluations in order to assess the strength of the evidence being reported and to provide a foundation for learning how to strengthen future advanced primary care evaluations.

The Multi-State Collaborative

In 2009, with support from the Milbank Memorial Fund (MMF), five states—Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont—began to collaboratively share their experiences in transforming their states' primary care delivery systems. In 2010, the group formally became the "Multi-State Collaborative." The MC continued to share its outcomes and data and to advocate with the Centers for Medicare & Medicaid Services (CMS) to improve collaboration between the states and federal government in support of multi-payer primary care initiatives. As of 2015, the MC included 17 states and 18 programs (there are two programs in New York State). Most of the programs participate in the Multi-Payer Advanced Primary Care Practice (MAPCP) demonstration or the Comprehensive Primary Care (CPC) initiative sponsored by CMS. MMF views this growth as evidence of the benefit states receive through collaboratively learning from, and sharing information with, their peers.[1]

Participating MC Patient-Centered Medical Home (PCMH) Programs

As described in detail here, eight of the 18 MC members participated in this study to assess the strength of the evidence generated by participating members' evaluations.[2]

These eight programs have robust, multifaceted PCMH programs (Table 1). They have all implemented innovations to enhance capacity and provide additional services in primary care practices, combined with payment reform mechanisms. Without conducting a formal implementation analysis, we briefly describe the participating programs' attributes here:

Enhanced Services and Capacity

- Enhanced self-management
- Mental health
- Social and economic needs
- Substance use disorders
- Practice transformation support
- Care coordinators

- Almost all participating MC members' programs provide practice transformation support—including practice facilitators or coaches—and care coordination services.
- All participating MC members' programs provide support for enhanced self-management.
- Participating MC members' programs provide varying levels of support for mental health, social and economic needs, and substance use disorders.

Payment Reform Mechanisms

- Routine fee-for-service (FFS)
- Per-member-per-month (PMPM)
- Shared savings

Participating MC members' programs generally include payment incentives aligned with promoting quality and building capacity toward measurable outcomes that address population health.[1]

- Almost all participating MC programs preserved traditional fee-for-service (FFS) arrangements among a majority of payers. This reflects the current insurance market, as well as the pilot nature of the participating MC programs, particularly in the earlier years.
- All participating MC programs utilize per-member-per-month (PMPM) arrangements with practices or provider organizations, often to support care coordination or care management activities or tied to meeting PCMH standards.
- One-half of participating MC programs include shared savings arrangements, with some of these added as the programs matured.

PMPM payments often fund care management or care coordination services. In some cases, a staff member is directly assigned to oversee these efforts, such as a nurse care manager. In addition to funding care management/coordination services, some programs use the PMPM fees to cover other efforts, such as practice transformation, preventive care, administrative support, or performance incentives. While one participating MC program specifically noted that all payers share equally in their PMPM payments, most participating MC programs experience variation in payment amounts by payer.

Table 1. Attributes of Participating MC Programs and Payment Models

Intervention Attributes	A	B	C	D	E	F	G	H
Enhanced Services and Capacity								
Enhanced self-management	Y	Y	Y	Y	Y	Y	Y	Y
Mental health	V	N	Y	Y	V	Y	?	Y
Social and economic needs	N	N	Y	Y	V	?	?	Y
Substance use disorders	N	N	Y	Y	V	Y	?	?
Practice transformation support	Y	Y	Y	Y	V	?	Y	Y
Care coordinators	Y	Y	Y	V	Y	Y	Y	Y
PCMH Payments								
Routine FFS	Y	Y	Y	Y	Y	Y	Y	N
PMPM to PCMH	Y	Y	Y	Y	Y	Y	Y	Y
Shared savings to PCMH	Y	Y	N	N	N	Y	Y	N

Y = participating MC program consistently utilizes service/capacity or payment
N = participating MC program does not consistently utilize service/capacity or payment
V = variation across practices; ? = not enough information to determine.

Motivation and Study Aims

This paper is written for use by evaluators, funders of PCMH evaluations, implementers of PCMH programs, and policymakers with the aim of understanding the strength of evidence of the effectiveness of the PCMH model of primary care generated from this set of evaluations and using the information to inform strategies for strengthening future initiatives or evaluations.

In this section, we briefly describe the motivation and study aims for this project. In the following section, we note the factors used to assess the strength of the eight participating MC evaluations. We then present our assessment of the strength of the evaluations, the findings for six outcomes, and our assessment of the strength of the evidence given the evaluation methods used. Next, we suggest practical strategies for strengthening evidence in evaluation of primary care interventions. In the final sections, we offer perspectives on the future evaluations of payment reform demonstrations and provide evaluation resources.

The PCMH concept has evolved over the past 40 years, from an initial focus on improving care for children with special health care needs to a broader primary care transformative health system. It is designed to provide physician-directed care that is "accessible, continuous, comprehensive and coordinated and delivered in the context of family and community," as defined by four primary care physician specialty societies.[3,4] Since 2007, the PCMH concept has been endorsed by a range of purchaser, labor, and consumer organizations.[5]

Evidence from peer-reviewed studies of early and evolving PCMH models shows some promising results in decreasing the cost of care and use of acute care services, improving processes of and access to care, as well as improving patient satisfaction. However, a review of early studies reported mostly inconclusive results because of shortcomings in evaluation methods.[6,7]

As of January 2014, over 90 private health insurers, dozens of large employers, Medicare, the Veterans Administration, TriCARE, the Federal Employee Health Benefits Program, and 25 state Medicaid programs were making significant financial investments in the PCMH concept through upfront payments to practices, performance bonuses, and additional care management supports.[5] Further, almost 7,000 primary care practices have each made significant investments to receive recognition as a PCMH from the National Committee of Quality Assurance (NCQA).[8] The US Congress recently passed legislation enabling practices that are PCMHs to qualify as Medicare alternative payment models, making them eligible to receive bonus payments in the future.[9] Current evidence suggests that more mature PCMH models show stronger improvements, and that multi-payer models may hold greater promise for improving cost and utilization outcomes.[5] It is, however, still too early to state the definitive impacts of the PCMH, particularly with a full understanding of the strength of evidence being generated.

Since 2009, the MC members have aimed to provide this type of sound evidence by assessing their own programs in primary care transformation.[10] The MC PCMH programs have grown over time in complexity—in terms of number and types of participating practices, patients, and payers; supports offered to practices; and payment mechanisms. As programs become more complex, so do the methods needed to evaluate their effectiveness and provide useful information to key stakeholders to allow for continued evolution of the PCMH model. The evaluation challenge is to ensure that promising results are not being missed or do not overstate the actual impact of each program on health outcomes.

Identifying MC Member Participants and Collecting Information

The MMF aims to understand the extent to which MC PCMH programs are improving outcomes for a core set of key health care utilization and spending measures.[11] A critical first step is to systematically assess the strength of the evidence being reported by participating programs. All 18 current MC members were invited to participate in the current study and eight agreed to participate. We did not analyze any evaluation methods or results for non-participating MC programs. The eight MC participants provided us with documents describing their most recent evaluations as well as their results. In some cases, methods and results varied by participating payer, and we analyzed them separately. We did not receive nor analyze any raw data, nor did we conduct a formal meta-analysis of the results.

The eight participating MC members' evaluations included in this study use a variety of methods with some strong components. Five of the evaluations report favorable reductions in acute care utilization and expenditures. Six of the eight evaluations report findings that are not statistically significant for one or more key outcomes. Our assessment of the strength of evidence indicates that there is some reason to question both the favorable results and the results indicating no effects because several important gaps exist in the methods.

This study is designed to serve as an important foundation for collaborative learning going forward, specifically on how to strengthen evaluations in order to increase confidence that the evidence is credible. To achieve this foundation, we have the following four aims:

- Catalog the evaluation methods and results of the participating MC member's most recent evaluation.
- Assess the strength of each evaluation effort.
- Identify common strengths and gaps in methods.
- Identify practical strategies for strengthening future primary care evaluations.

Approach and Study Methods

A. Evaluation Domains

To assess the strength of methods used and reported in the eight evaluations, we identified four domains that contribute to the overall strength of an evaluation. We assessed each MC member's evaluation approach for adequacy within the four domains. Our primary aim was to assess the degree to which participating MC members use methods that maximize *internal validity* of study findings by minimizing two types of errors:

- False positives, whereby no "real" changes in outcomes actually occurred but the study findings show favorable findings (e.g., increase in quality of care or a lower rate of growth in expenditures).
- False negatives, whereby there are "real" changes in outcomes but the study findings show no changes.

We define "real" changes as those in which we have a high degree of statistical confidence that the change occurred. Two primary drivers of false positive and false negative findings are (1) lack of appropriate statistical adjustment for clustering of outcomes within practices and (2) small sample sizes.[6]

Second, we aim to assess the degree to which participating MC evaluation designs maximize *external validity*. In the case of non-experimental designs, where there is self-selection of practices that transform to PCMHs, we define external validity as the generalizability of study findings to a similar set of practices and patients.[12]

Evaluation domain 1: Comparison group is sound

There is widespread consensus in the research community that a sound comparison group is critical to producing unbiased causal effects.[13] Generally, choosing practices from the same geographic areas is best unless there is widespread diffusion of medical homes within the area. This leaves too few practices or systematically different practices available for the comparison group, which creates a selection bias and is a threat to external validity. In this instance, it is necessary to identify study areas with similar patient sociodemographic characteristics, geographic characteristics, and health care system characteristics. Using geographic areas outside of the same state adds an additional level of evaluation complexity, in particular, when analyzing effects on Medicaid beneficiaries due to the fact that benefits and payment levels vary considerably across states. The following four factors were considered when assessing soundness of the selected comparison group:

Four Domains Used to Evaluate Strength of Evidence of a Study

1. Comparison group is sound
2. Evaluation design is rigorous
3. Study is well-powered to detect effects
4. Statistical methods are sound

Secular changes that affect outcomes are captured. It is difficult to capture secular trends in factors correlated with study outcomes without a comparison group that is subject to the same external influences as the intervention group. For example, closure of an emergency department (ED) in a rural area will reduce the future likelihood of patients receiving care in an ED because of increased travel burden. If the intervention practices are located in the rural area and comparison practices are in an urban area with no change in ED availability, the study may erroneously conclude that ED usage declined during the study period within the intervention group because of the presence of PCMHs.

Intervention and comparison groups resemble one another. Ensuring that comparison groups resemble intervention groups in terms of unobserved characteristics, such as motivation to change care delivery, as well as observed characteristics is the fundamental design challenge. By definition, comparability of unobserved characteristics cannot be ensured. Matching should be conducted at the level at which the intervention occurs; that is, comparing practices to practices for practice-level interventions, patients to patients for patient-level interventions, and so on. Most PCMH initiatives are considered practice-level interventions, so matching most often occurs at the practice level. There are a variety of methods for matching, but most evaluations use propensity score matching. Different approaches to matching appear to work better for different situations, making it important to select the approach that best fits the data. Checking and reporting equivalency of intervention and comparison groups' practice characteristics (such as practice size and ownership status) and their patients' sociodemographic, health status, and pre-intervention values of study outcomes are necessary regardless of the approach taken for matching. Ensuring that there is overlap between the intervention and comparison groups in the range of covariate values (e.g., the age range is similar across the intervention and comparison group) is also important. As with matching, there are a variety of methods for ensuring equivalency in observed characteristics.

Outcomes are measured using comparable data sources and at similar times. Use of disparate data sources between the intervention and comparison groups (such as clinical records for the intervention group and claims data for the comparison group) or collecting data at different times can lead to spurious conclusions about program effects.

Attribution method of the intervention group is replicated in the comparison group. Many PCMH programs use administrative data to attribute patients to practices based upon decision rules such as the patient received the majority of their primary care in the year prior to the study year at that practice. Because these assignment algorithms are based upon prior use of health care services, it is necessary to apply the same attribution algorithm when assigning comparison group patients to similar practices.

Evaluation domain 2: Evaluation design is rigorous

Although there is no absolutely right or wrong way to conduct an evaluation of a non-experimental design, there are a number of key features that strengthen evaluations of such designs:

Intervention and comparison groups followed over time. Transformation to a PCMH reflects a dynamic process of practice or system change aiming to achieve better quality of care, improve population health, and decrease costs. This type of non-experimental intervention is most often evaluated by tracking the difference in outcomes over time between the intervention and comparison group. Although no perfect method exists for drawing inferences from non-experimental data, a well-known tool in program evaluation, difference-in-differences, can be employed. It compares changes between two time periods in the outcome for an intervention group with changes for a comparison group, and is typically the most reliable way to draw inferences from observational data. It allows evaluators to remove effects of confounding influences, providing a less biased estimate of the effect of the PCMH intervention. In contrast, a pre- to post-study with no comparison group is considered a weak design, in which significant reductions in utilization or expenditures, particularly for chronically ill participants, may reflect a regression to the mean rather than the *true* impact of the PCMH program. There are a variety of specification choices for difference-in-differences models.

Intention-to-treat design. In an intention-to-treat design, patients attributed to the intervention and comparison groups are followed throughout the study period regardless of whether they drop out or are disenrolled. This minimizes non-random attrition of patients which can affect estimates of program effects.

Enough follow-up data to enable at least a minimum length of exposure to intervention. It is a common practice to set a minimum length of possible exposure to the intervention to ensure an adequate potential "dosage" effect. Including patients who could only have received a few days or months of exposure will dilute estimates of program effects.

Outcomes adjusted for partial period eligibility. A common estimation problem in PCMH evaluations is that patients will lose health insurance coverage or die during the study period. Researchers typically "annualize" expenditures—increasing the patient's total expenditure estimate to the full year or estimating a per-member-per-month expenditure. Because there is less statistical certainty associated with the "annualized" estimate when the patient is not observed for a full year relative to observed expenditures for patients observed for a full year, correcting the standard errors through weighting is necessary.

Evaluation domain 3: Study is well-powered to detect effects

Evidence from early evaluations of PCMHs suggests that one might expect to find statistically significant savings in the range of 5% for a general population and upwards of 15% for a chronically ill population.14 We evaluated the degree to which studies reported the statistical power they had to find the sizes of the effects in this range statistically significant:

Minimum detectable effect (MDE) reported. Why is it important to report MDEs? Inconclusive findings of PCMH initiatives may be the result of ineffective implementation of the PCMH model, the PCMH model may itself be ineffective, or implausibly large performance requirements are needed to result in statistical significance. Without knowing what the MDEs are, findings of no changes cannot be interpreted as no "real" change. An MDE is the minimum true effect of a PCMH intervention that can be detected with a given level of statistical significance (p-value 0.05 or 0.10) for a given level of knowledge that the effect is not due to chance (typically 80%). Calculating MDEs will also provide guidance on the number of patients and practices that are required to detect substantively important clinical or financial changes.

Design effect—longitudinal cohort versus cross-sections. Regression models can be estimated using a longitudinal cohort design—following the same intervention and comparison patients, multiple cross-sections, or different intervention and comparison patients over time. A cohort design—following the same study participants over time—has several advantages over a cross-sectional design. A cohort design will capture the correlation between patients' outcomes before and after the intervention began. Because baseline levels of utilization and expenditures are generally strong predictors of future utilization and expenditures, cohort designs allow for smaller program effects to be found statistically significant. It requires approximately one-quarter fewer patients than a cross-sectional design for the same level of statistical power. In other words, using a cross-sectional study design substantially increases sample size requirements for the same level of statistical power to detect a given effect size.

Evaluation domain 4: Statistical methods are sound

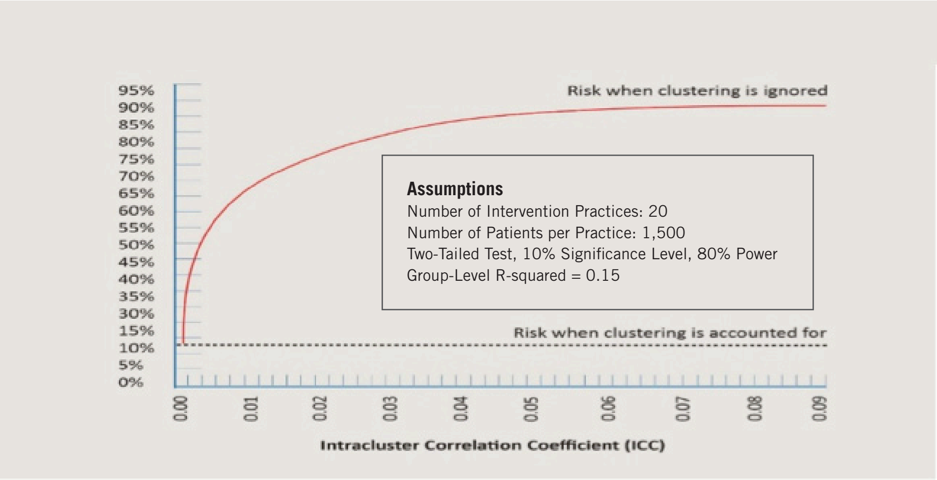
There are several methodological considerations when conducting statistical analyses to determine program effectiveness:

Covariates that influence outcomes are included in regression models. Propensity score matching of comparison practices to intervention practices using geographic-, practice-, and patient-level characteristics related to study outcomes is intended to create a comparison group that has similar observable characteristics as the intervention group. In addition, it is common practice to include these same characteristics as explanatory variables in regression models estimating program effects. The intent is to further control for any remaining imbalances, thereby removing bias from the estimate of program effects because of differences in the characteristics of the intervention and comparison groups before the start of the intervention.

Results are adjusted for clustering at the practice level. PCMH programs are generally considered practice-level interventions. Patients attributed to a PCMH can be expected to receive care (and associated outcomes) that is more similar to care received by other patients in the same practice than care received in other practices. The degree of similarity in outcomes can be measured using a statistic called the intracluster correlation coefficient (ICC), which ranges from zero to one. The practical implication of this clustering of outcomes within practices is to reduce the effective sample size for an evaluation.[14] As the ICC increases, the effective sample size decreases, which means that we need to observe a very large change in an outcome, such as expenditures, before it will become statistically significant.[15]

Not accounting for clustering at the practice level (if the ICC ≠ 0) also increases the likelihood of false-positive findings, whereby no "real" changes in outcomes actually occurred, but the study findings show an increase in quality of care or a lower rate of growth in expenditures. Thus, standard errors need to be adjusted downward to reflect the lack of independence of patient outcomes within practices. How important is this adjustment? Figure 1 displays the risk of false-positive findings of estimated savings when we account for clustering at a desired level of significance (for example, 5% or 10%), for various levels of the ICC, ranging from 0% to 9% (0.09), a relevant range for health care costs and use of services.[14] As the figure shows, if the ICC equals 1% (0.01), then failure to account for clustering increases the false-positive risk to almost 70%! ICCs can vary substantially across programs and study populations and should be calculated during the design phase of a program.

Figure 1. Risk of False Positive Findings of Savings Based on Clustering



Assumptions
Number of Intervention Practices: 20
Number of Patients per Practice: 1,500
Two-Tailed Test, 10% Significance Level, 80% Power
Group-Level R-squared = 0.15

Peikes D, Dale S, Lundquist E, Genevro J, Meyers D. *Building the evidence base for the medical home: what sample and sample size do studies need?* White Paper (Prepared by Mathematica Policy Research under Contract No. HHSA290200900019I TO2). AHRQ Publication No. 11-0100-EF. Rockville, MD: Agency for Healthcare Research and Quality. October 2011.

Minimum length of run-out of outcome data needs to be the same for treatments and comparisons. When using administrative data to conduct evaluations, it is important to ensure that there is a minimum amount of time that has transpired from the date of service to allow for the claims adjudication process to be fairly complete. It is common to allow a three- to six-month claims "run-out" period before using data for evaluations. Using different periods of run-out for the intervention and comparison groups will likely affect utilization and expenditure estimates.

Sensitivity analyses/robustness tests are conducted. Results can be sensitive to a variety of influences. It is important to "kick the tires" to ensure that results are not heavily dependent on particular observations, people or data points, or model specifications. Common sensitivity tests for robustness of findings include comparing the main results to results with (1) truncation of expenditure data (or other outcomes with a skewed distribution) at the 98th or 99th percentile, log transformation of data or use of a two-part utilization model; (2) removal of individuals as outliers identified using influence statistics; and (3) estimating different model specifications, such as changing the number of years in the pre-intervention period.

Significance is adjusted for multiple comparisons. Most PCMH evaluations test a large number of outcomes, potentially creating the problem of multiple test bias—increasing the likelihood of a positive study finding when there is no "real" change due to chance. For example, if an evaluation tests 15 outcome measures over a one-year period using a quarterly regression model, 60 tests would be conducted leading to the statistical expectation that six of the estimates would be statistically significant (if tests are conducted at the

0.10 significance level using a two-tailed test), even if the program had no "real" effect. Several common approaches that can be used to guard against multiple test bias include identifying a few outcomes as the primary outcomes for policymaking purposes with others as secondary outcomes to help inform the primary outcomes, or using a more conservative significance level (e.g., 0.01 rather than 0.05) or a Bonferroni correction factor.

B. Review of Evaluation Methods

We assessed the strength of evidence from the evaluation methods used by MC members participating in this study. We used the template found in Appendix A to abstract information from secondary data sources provided by the participants. After completing the initial abstraction, we conducted follow-up telephone calls with each participating MC member and/or their local evaluator to ensure accuracy of the information abstracted. During these calls, we clarified questions and obtained additional information that was not available from the reviewed documents. After finalizing templates for each participating MC member's evaluation, we created a set of standardized data categories to increase the comparability of the data collected across the evaluations. We also identified evaluations for which there was not sufficient information available from the participating MC member, its evaluator, or the participating payers to fully assess the methods, or for which study results were not available to include in this report.

We focused this report on key utilization and expenditures outcomes measures adapted from the set of measures recommended by the Commonwealth Fund's Evaluators' Collaborative and generally considered to be actionable goals by PCMH initiatives.[11] Because we observed a wide variation in participating MC programs' reporting on quality of care measures, we ultimately excluded those from this report.

Results

A. Strength of Evaluation Methods

We assessed the participating MC members' evaluations on four domains to determine the strength of the evaluation methods. Here, we provide a high-level assessment of the strength of evidence generated from the evaluation methods used among the participating MC members:

Evaluation domain 1: Comparison group is sound

Seven of the eight evaluations include a comparison group in the same or similar geographic areas to capture secular changes in similar markets that could affect outcomes (Table 2). The evaluation without a comparison group uses an approach in which the intervention group's baseline utilization is trended forward, which may or may not adequately account for potentially confounding factors during the study period. Among the MC evaluations that use a comparison group, all but one apply attribution criteria similar to those that apply to the intervention group, thereby increasing similarity between the two groups.

There was less uniformity in how the participating MC evaluations approach ensuring that the intervention and comparison groups had similar observable characteristics that are likely to be correlated with outcomes. Six evaluations used either propensity score matching or inclusion of covariates in the outcomes models. However, only three evaluations provided confirmation on whether the two groups were indeed comparable. All but one evaluation with a comparison group was able to observe comparison group patients in the pre- and post-periods. All evaluations using a comparison group measured relevant outcomes using comparable data sources and at similar times. Both of these factors contribute to the ability of the evaluations to generate strong evidence.

Table 2. Comparison Group Is Sound

Domain	A	B	C	D	E*	F	G**	H
Secular changes that affect outcomes captured	✓	✓	✓	✓	✓	✓	X	✓
Intervention group eligibility criteria and attribution method replicated	✓	✓	✓	✓	✓	✓	X	n/a
Comparison group observed in pre- and post-periods	✓	✓	✓	✓	✓	✓	X	X
Findings from testing for equivalency between intervention and comparison group reported	✓	✓	X	X	n/a	✓	X	X
Outcomes measured using comparable data sources and at similar times	✓	✓	✓	✓	✓	✓	X	✓

* Responses indicate plans, though analyses not carried out yet.

** No comparison group used in assessment.

✓ = Participating MC member evaluation met criteria; X = participating MC member evaluation did not meet criteria; ? = not enough information about evaluation provided to determine whether criteria met; n/a = not applicable.

Evaluation domain 2: Evaluation design is rigorous

The participating MC members used a variety of evaluation approaches, all with elements of a rigorous evaluation design. All but two evaluations compared an intervention and comparison group over time, a strong non-experimental design evaluation approach (Table 3). All used an intention-to-treat design, which minimizes non-random attrition that could affect estimates of program effects. We confirmed that five of the eight evaluations require patients to have had a minimum length of potential exposure to the PCMH intervention (or attribution to participating practices), which minimizes dilution of findings. In addition, five of the eight evaluations adjusted their outcomes for partial eligibility of patients, which appropriately gives more statistical weight to patients with longer exposure to medical homes.

Table 3. Evaluation Design Is Rigorous

Domain	A	B	C	D	E*	F	G**	H
Pre- and post-data for intervention and comparison group used	✓	✓	✓	✓	✓	✓	X	X
Intention-to-treat design used	✓	✓	✓	✓	✓	✓	✓	✓
Participants had minimum length of potential exposure	✓	✓	X	X	✓	✓	✓	?
Outcomes were adjusted for partial eligibility	✓	✓	✓	✓	X	X	✓	X

* Responses indicate plans, though analyses not carried out yet.

** No comparison group used in assessment.

✓ = Participating MC member evaluation met criteria; X = participating MC member evaluation did not meet criteria; ? = not enough information about evaluation provided to determine whether criteria met; n/a = not applicable.

Evaluation domain 3: Study is well-powered to detect effects

None of the participating MC members presenting results reported on the statistical power they had to detect plausibly sized effects or the minimum detectable effect (MDE) they could statistically detect, which may lead to erroneous null findings (Table 4). Even with large sample sizes, we cannot assume that evaluations are adequately powered to detect changes in utilization and expenditures in the 5% to 15% range. Seven of eight evaluations use a longitudinal cohort design; one uses a cross-section cohort design. Relative to a randomized control trial, a cross-sectional design requires upward of four times the sample to detect the same size of effect. Without reporting statistical power or MDEs, it is not possible to know with certainty whether studies are underpowered to detect changes in utilization or expenditures, which could explain insignificant findings.

Table 4. Study Is Well-Powered to Detect Effects

Domain	A	B	C	D	E*	F	G**	H
Study is well-powered to detect plausible effects	X	X	X	X	X	X	X	X

* Responses indicate plans, though analyses not carried out yet.

** No comparison group used in assessment.

X = participating MC member evaluation did not meet criteria.

Evaluation domain 4: Statistical methods are sound

There is a lack of uniformity in the statistical methods used to estimate program effects among participating MC members' evaluations (Table 5). Three of eight evaluations did not include characteristics that likely influence outcomes in the construction of risk-adjusted rates or regression models. In this situation, there may be missed opportunities to improve covariate balance between the intervention and comparison groups. Five of the evaluations required a minimum length of claims run-out for claims/encounter data used in analyses. Four of the eight evaluations adjusted standard errors for patient clustering within practices, but only one evaluation adjusted for multiple comparisons. Both approaches may lead to erroneous positive findings. Two of the evaluations did not use statistical hypothesis testing to determine whether some or all results were statistically significant. Lastly, only three of eight evaluations conducted sensitivity analyses to confirm the robustness of their findings to specification error. This may reflect limits in resources or time available for additional statistical analyses. Confidence in findings would be stronger with additional analyses.

Table 5. Statistical Methods Are Sound

Domain	A	B	C	D	E*	F	G**	H
Regression model(s) estimated included baseline outcome covariates or risk adjusters (for repeated cross-sections)	✓	✓	✓	✓	✓	X	X	X
Minimum length of run-out of outcomes data used	?	✓	✓	✓	✓	✓	✓	?
Results adjusted for clustering (practice level)	✓	✓	X	✓	✓	X	X	X
Sensitivity analyses/robustness testing conducted	✓	✓	✓	X	X	X	X	X
Significance adjusted for multiple comparison	X	X	X	X	✓	X	X	X

* Responses indicate plans, though analyses not carried out yet.

** No comparison group used in assessment.

✓ = Participating MC member program met criteria; X = participating MC member program did not meet criteria; ? = not enough information about evaluation provided to determine whether criteria met; n/a = not applicable.

B. Changes in Utilization and Expenditures

We report results from seven of the eight MC participating members' published and unpublished findings (Table 6) for four acute-care utilization measures and total expenditures. One participating MC member was able to provide us with detailed information on their self-assessment design and methods but was unable to share results, as analyses have not been completed. What we present here spans different types of payers and patients, interventions, follow-up periods, and estimation methods; and the measures themselves are not necessarily consistently defined. Upward and downward arrows reflect statistically significant results in the noted direction (e.g., a downward arrow for all cause admissions means the rate of growth was favorably lower among patients being treated at PCMHs than patients treated at non-PCMHs). Results that are reported as not being statistically significant (denoted in the table as NS) may reflect either lack of detection at the given level of significance used by the evaluation or lack of statistical significance testing by the evaluator. Outcomes not reported by a MC member are denoted in the table as NR.

Five of the seven evaluations reported one or more favorable impacts of medical homes on outcomes—decreases in acute care utilization or expenditures over time among patients attributed to PCMHs. We observe a reduction in the rate of all-cause hospitalizations in three evaluations (C, G, H), the rate of ambulatory care-sensitive condition (ACSC) hospitalizations in one evaluation (B), the rate of all-cause emergency department (ED) visits in two evaluations (B, G), and the rate of ACSC ED visits in one evaluation (A). Among these five evaluations, one also reported an unfavorable finding—an increase in the rate of ACSC ED visits (C). Total expenditures declined in three of the seven participating MC programs that tracked expenditures (C, G, H).

While all of the participating MC evaluations possessed some strong evaluation features, our assessment of the strength of evidence indicates that there is also some reason to question both the favorable and unfavorable results. We linked these findings to each study's methodology to assess the degree to which statistically significant findings or no statistically significant findings likely reflect "real" changes in outcomes. Two of seven evaluations reported no reduction in any acute care utilization measure (D and F), and three of six evaluations reported no reduction in total expenditures (B, D, F). Yet none of these evaluations reported what size reduction would have been necessary for them to have determined that the change was statistically significant. Reporting minimum detectable effects would allow us to know if each negative finding was due to no real changes occurring at the point of study or the size of the change was smaller than what one could expect to find given the statistical properties of the data. The reason for not showing a change has very different implications to practices, payers, and policymakers.

Of the five studies that reported a reduction in hospitalizations, ED visits, or total expenditures, two adjusted their standard errors for clustering (A, B). One evaluation did not perform any statistical testing (G) and a second did not perform statistical testing of expenditures (H). Thus, *the decrease in hospitalizations and total expenditures may be overstated*. Further, there was limited reporting of findings from sensitivity analyses conducted in all seven studies to test the robustness of their findings. For example, sensitivity tests could indicate that favorable findings remain when removing outliers or modifying the statistical model, thus lending more credibility to the main findings. Further, limited subgroup analyses were conducted or reported. Reporting results by subgroups might also show for which patients the PCMH model of care provides the greatest benefit.

Table 6. Changes in Rates of Acute Care Utilization and Total Medical Expenditures from Participating MC Evaluations of Their PCMH Programs

	A	B	C	D	E*	F	G**	H
Utilization								
All cause admissions	NS	NS	↓	NS	N/A	NS	↓	↓
ACSC admissions	NS	↓	NR	NS	N/A	NR	NR	NS
ED visits–total	NS	↓	NS	NS	N/A	NS	↓	NS
ED visits–ACSC	↓	NS	↑	NS	N/A	NR	NR	NR
Expenditures								
Total expenditures	NR	NS	↓	NS	N/A	NS	↓	↓**

* No results from evaluation reported yet.

** No statistical testing conducted.

Upward arrow = statistically significant increase in utilization or expenditures; Downward arrow = statistically significant decrease in utilization or expenditures; NS = not significant; NR = result not reported; NT = result not tested for significance; N/A = not applicable (no results reported yet).

Practical Strategies for Strengthening Evidence in Evaluations of PCMHs

Why are rigorous evaluations of PCMHs so important? First, PCMHs themselves continue to evolve, but they do so now in a health care environment that is itself rapidly adopting new health care service and delivery models designed to have the same effects on patient outcomes. This substantially increases the likelihood of “spillover effects” from these outcomes being measured in current PCMH evaluations, and on outcomes for both intervention and comparison patients.

A variety of evaluation approaches were used across the participating MC members, all with the intent of conducting a rigorous assessment of their PCMH programs. All possess some strong approaches. We find that each evaluation could be improved going forward by either enhancing methods or reporting of findings. The primary reason for improving is to reduce the risk that “real” changes in outcomes are missed or that there are no “real” changes but the study results suggest otherwise.

In this section, we provide practical strategies to consider when designing or conducting future evaluations of an advanced primary care program. Strengthening evaluation methods in a couple of key areas will have two benefits: (1) increase the level of confidence of policymakers and PCMH staff that positive program effects reflect real change and are worthy of continued support; and (2) ensure that a finding of no program effects reflects no meaningful change, rather than a statistically underpowered study design.

Address Statistical Power and Methods

- Calculate minimal detectable effects (MDE). Without reporting MDEs, evaluators cannot be sure whether a lack of favorable findings is due to ineffective implementation of a PCMH initiative, a PCMH initiative not being an effective model, or implausibly large performance requirements to be detected by the evaluation.

- Account for clustering when there are multiple practices in the study. Calculating intra-class coefficients can help determine whether it is important to account for clustering (at the practice level). If it is, consider adjusting your results accordingly. Failure to account for clustering when evaluating performance among a set of PCMHs will increase the likelihood of finding a statistically significant change when, in fact, no change has occurred.

- Conduct sensitivity analyses. It is preferable to conduct robustness tests, such as estimating effects with trimmed versus untrimmed outliers or varied functional forms of outcome variables, to see if the findings are similar to the main results. Sensitivity analyses often increase the confidence that the main findings are robust to different specifications of the methods.

- Transform outcome variables. Consider transforming outcomes, such as using dichotomous (likelihood of hospitalizations) rather than continuous (rate of hospitalizations) variables, to reduce the variance of estimates. This increases the likelihood of finding statistically significant effects.

- Try Bayesian methods. Unlike traditional methods, Bayesian methods allow evaluators to make intuitive probabilistic statements about the size of program effects (e.g., there is an 80% probability that PCMHs reduced the growth in medical expenditures by at least 10%). Bayesian methods also can “borrow information” across time or subgroups of patients to increase statistical power to detect true effects.

Capture Secular Changes that Affect Outcomes

- Use similar geographic areas from which to draw the comparison group. It is ideal if the evaluation uses a comparison group within the same geographic area and state. This increases the likelihood that changes in the health care market, independent of the PCMH initiative, will equally affect the intervention and comparison groups’ outcomes. When concern about selection bias is high because of likely systematic differences between participating and non-participating practices, select other geographic areas that are most similar in terms of population sociodemographic characteristics, geographic characteristics, and medical supply and demand factors.

- Select practices for the comparison group that are similar to the intervention group. When constructing a comparison group, practices should be selected with similar characteristics that affect outcomes unrelated to becoming a PCMH.

- Select patients for the comparison group that are similar to the intervention group. It is best to use patients for the comparison groups with similar sociodemographic and health status characteristics as well as the same insurance benefits and payments.

- Test baseline trends. Evaluators should test that baseline trends in outcomes are comparable for the intervention and comparison groups if using a longitudinal cohort evaluation design.

- Clearly identify threats to validity of the comparison group from “spillover effects” of other health system transformation activities. Evaluators should seek to understand if comparison group patients and practices are participating in other health system or payment reform initiatives that may reduce acute care utilization and expenditures and increase quality of care (e.g., accountable care organizations or ACOs). If PCMHs are participating in ACOs, as an example, evaluators should test for the incremental effect of PCMH status relative to comparison practices in ACOs and test for the joint effect of PCMH and ACO status relative to comparison practices not participating in ACOs.

Subgroup Analyses

- Analyze high-cost subgroups. Previous PCMH studies have shown cost savings of up to 15% in chronically ill populations, in contrast to 5% for a general population. Restricting analyses to a homogenous and sicker population will also improve the likelihood of finding a statistically significant effect because the coefficient of variation is smaller than observed in a general population. Analyzing program effects across a heterogeneous group of patients often dilutes the effect that one can observe among a sicker or more homogeneous group of patients.

- Analyze early versus late adopters of the PCMH model. Consider conducting analyses by timing of adoption of the PCMH intervention.

Future Directions in Evaluations of PCMHs

The principles and standards of primary care transformation and the PCMH are evolving to better serve patients’ physical, behavioral, and social and economic needs. This occurs in an increasingly complex environment with other, sometimes overlapping health care system reforms. Evaluation designs and methods also need to evolve to *identify* (1) the components of PCMHs, and (2) relationships of PCMHs with other health system and payment reform initiatives and *link how these affect outcomes*. Given the dynamic situation, future evaluations should consider examining both the overall effect that different PCMH approaches have on reducing costs and improving outcomes and if improvements in particular features of the medical home or interactions with other initiatives lead to improved outcomes.

Another fundamental challenge will be to identify the counterfactual in an era of a rapidly changing health care landscape. No longer will we be evaluating a single payment reform intervention, such as a bundled payment for heart bypass surgery, relative to traditional fee-for-service payment. We will instead be evaluating PCMHs relative to primary care practices that are also likely to be participating in health system and payment reform interventions. Thus, the net benefit of PCMH transformation is likely to be smaller, making it far more difficult to identify *real* changes from "noise." Adding to the complexity will be the increasing multi-payer nature of reform activities. While multi-payer collaboration can increase financial and other supports to practices, payer-specific requirements and different targeted populations and performance metrics may reduce the likelihood of an overall finding of effectiveness because of variation across payers.

Our qualitative and quantitative evaluation toolkits must expand to allow for identification of program elements associated with success or failure. Qualitative research tools must enable us to comprehensively capture implementation barriers and facilitators, clinically meaningful organizational changes, and complex intersections of reform activities. Linking these program features and factors to improvements in key program outcomes will provide important feedback on how best to define future interventions and challenges that need to be addressed going forward.

We must also move beyond classical regression methods and testing that gives a "thumbs up" or a "thumbs down." Application of methods should provide actionable information to practices and sponsors of the initiatives. Bayesian approaches to estimating program effects offer important advantages over more traditional evaluation methods. These are particularly compelling in program evaluations, where sponsoring organizations may wish to make a statement, such as "There is an 80% chance that the intervention reduced the cost of care by at least 10%." The Bayesian approach yields program estimates that are more precise, allowing program and subgroup effects to be identified that might otherwise go unrecognized due to insufficient statistical power.

Together, movement toward more creative designs upfront, stronger qualitative and quantitative methods in the evaluation, and linking practice changes to improved outcomes will more quickly and accurately inform policymakers responsible for crafting the next generation of primary care reforms.

Resources

Building the Evidence Base for the Medical Home: What Sample and Sample Size Do Studies Need?
Evaluations of the medical home should account for clustering of patients within practices. This paper describes why and how to do this and what samples of patients and practices are needed for studies to achieve adequate statistical power.
https://pcmh.ahrq.gov/sites/default/files/attachments/Building%20Evidence%20Base%20PCMH%20White%20Paper.pdf

The Medical Home: What Do We Know, What Do We Need to Know?: A Review of the Current State of the Evidence on the Effects of the Patient-Centered Medical Home Model
Amid burgeoning efforts to create medical homes across the US, this paper describes the evidence we have so far on the effects of precursors to the medical home model on key outcomes, and how to improve studies in the future.
https://pcmh.ahrq.gov/sites/default/files/attachments/the-medical-home-what-do-we-know.pdf

AHRQ PCMH Resource Center on Evidence and Evaluation
Policy decisions concerning the PCMH must rest on sound evidence about whether this model of care helps achieve the Triple Aim of improved patient outcomes, improved patient experience, and improved value. In this section, information and resources for PCMH researchers, evaluators, and decision makers are explored. Resources include:

- PCMH Research Methods Series is designed to "expand the toolbox" of methods used to evaluate and refine PCMH models and other health care interventions.
- Guide to Real-World Evaluations of Primary Care Interventions offers practical steps for designing an evaluation.
- White papers, briefs, and archived webinars.

https://pcmh.ahrq.gov/page/evidence-and-evaluation

PowerUp!
Dong N, Maynard R. "PowerUp"!: A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi-experimental design studies. *Journal of Research on Educational Effectiveness.* 2013; 6(1): 24-67.
http://repository.upenn.edu/cgi/viewcontent.cgi?article=1265&context=gse_pubs

The TREND Statement and Checklist
http://www.cdc.gov/trendstatement/

What Works Clearinghouse Procedures and Standards Handbook. Version 3.0.
Institute of Education Sciences. Their website offers over 700 publications and more than 10,500 reviewed studies in the online searchable database (available at http://ies.ed.gov/ncee/wwc/). *The What Works Clearinghouse Procedures and Standards Handbook* is available at http://ies.ed.gov/ncee/wwc/DocumentSum.aspx?sid=19.

Appendix

Multi-State Collaborative Data Collection Template

Lead MC state initiative: **Participating member (Respondent):** **Contact for obtaining follow-up information:**
TARGET POPULATION, SAMPLE SIZE, and INTERVENTION
Patient population (defined by insurance status; yes, no for each)
Medicaid
Commercial health plan
Medicare FFS
Self-insured employer
Sample size
Number of practices included in assessment
Attrition of practices: Number of practices that dropped out of assessment by end date
Number of patients included in assessment
Attrition of patients: Number of patients that dropped out of assessment by end date
Intervention attributes (yes, no for each)
Presence of a formal logic model (yes, no)
Additional care support services offered (yes, no, some)
Support for enhanced self-management/behavior change
Support for mental health
Support for social and economic needs
Support for substance use disorders
Practice transformation support for practices (practice facilitators/technical assistance)
Care coordinators hired by PCMHs
PCMH payments (yes, no for each)
Routine FFS
Enhanced FFS
PMPM to PCMH
Shared savings to PCMH
Other performance payment to PCMH
Timing of anticipated effects is stated in assessment documentation (yes, no)
If yes, describe:
EVALUATION DESIGN AND DATA SOURCES
Date range of assessment
Start date
End date
Does sample included in assessment differ from population served in PCMH intervention? (yes, no)
If yes, describe:

Intent to treat design (yes, no)
Identification of intervention group (yes, no)
Data collected for intervention group in pre-period
Data collected for intervention group in post-period
Attribution methods
Attribution method(s) used to identify intervention group (yes, no)
If attribution methods were used, were they: (note all that apply)
Claims-based, retrospective
Claims-based, prospective
Patient self-identification
Practice identification
If attribution methods were used, describe attribution rule
If attribution methods were not used, describe how intervention group was identified
Identification of comparison group (yes, no)
Data collected for comparison group in pre-period
Data collected for comparison group in post-period
Propensity score matching was used to select comparison group
If yes, matching was performed at which level:
Practice
Provider
Patient
Market
If yes, variables used to match[1]
Propensity score weights were used to balance comparison group to intervention group
If yes, method (e.g., inverse probability weights)
Covariate balance reported
If yes, method (e.g., standardized differences) and characteristics assessed
Credibility of comparison group selection methods (yes, no)
Can replicate eligibility criteria (if applied)
Comparison group subject to same external influences
Can be observed in pre- and post-periods
Outcomes data available from same source as for intervention group
Data sources
Medicare – FFS claims
Medicaid – FFS claims
Commercial FFS and/or encounter data
Self-insured FFS and/or encounter data
Electronic medical record/chart review/clinical registries
Survey – patient
If yes, interval(s) at which patient survey was conducted

Survey – provider

- If yes, interval(s) at which provider survey was conducted

Survey – other

- If yes, describe:
- If yes, interval(s) at which other survey was conducted

Time period unit of analysis

Quarterly

Annually

Other – describe:

CORE OUTCOME MEASURES

For each measure used in the assessment, describe its use and reported result

Clinical quality

To be determined

Health care utilization

All-cause inpatient admissions

Ambulatory care-sensitive (ACS) admissions (describe ACSCs included)

Emergency department visits (total or those that did not result in hospital admission; ACS or all cause)

Hospital all-cause 30-day (unplanned) readmissions (note if using 60- or 90- day measure)

Medical Expenditures

Total expenditures/average monthly expenditures (describe components included)

STATISTICAL METHODS

Statistical methods used to report quantitative results (yes, no) [If yes, report here:]

Unadjusted descriptive statistics reported

Adjusted descriptive statistics reported

Standard errors reported

Confidence intervals reported

Effect size reported

Model-based impact estimates reported (for example, regression/ANOVA/etc.)

Level of significance reported (0.05, 0.10, etc.)

Length of exposure and claims run-out

Minimum length of time patient is exposed to initiative before included in the analysis

Minimum length of run-out for claims/encounter data

Statistical power to detect effects

Minimum detectable effect (or any power calculation) is reported for core outcomes (yes, no)

- If yes, MDE for each core outcome measure

Results adjusted for clustering (yes, no)

- If yes, at what level:
- Practice
- Patient

Impact estimation method

Post-period only for treatment and comparison groups

Pre- and post-periods for treatment group only

Pre- and post-periods for treatment and comparison groups

Risk adjustment approach (yes, no)
Standardization of rates
If yes, method:
Regression model(s)
If yes, covariates included in regression model(s):
Weighting used in analyses (yes, no)
Propensity score matching weights used, if appropriate
Partial-period eligibility weights used
Other weights included
If yes, describe:
Utilization and expenditure adjustments
Annualization of utilization and expenditures
Price standardization
If yes, method for standardization:
Adjustment for multiple comparisons made (yes, no)
If yes, approach:
Sensitivity analyses were performed (yes, no)
If yes, tests used and reported:

[1] We will let respondents specify matching variables. For example, *practice characteristics* used for matching may include: percent of providers in a practice who meet meaningful use criteria for using electronic health records; number of primary care clinicians; percentage of clinicians at practice with primary care specialty; and whether NCQA-recognized medical home. *Beneficiary characteristics* used for matching may include: number of attributed patients' mean Medicare/Medicaid risk score; mean number of hospitalizations per person; and demographic mix of attributed patients (age, race, and gender categories). *Market characteristics* used for matching may include: whether in a medically underserved area; median income of the county; and whether in an urban area.

Source: Multi-State Collaborative member and partner assessment information

Notes

[1] Watkins L. *Aligning Payers and Practices to Transform Primary Care: A Report from the Multi-State Collaborative*. New York, NY: Milbank Memorial Fund; 2014. (Accessed March 8, 2016: http://www.milbank.org/uploads/documents/papers/Milbank%20-%20Aligning%20Payers%20and%20Practices.pdf.)

[2] The participating Collaborative programs are not identified to protect currently embargoed results or confidential agreements. They are instead denoted using letters A-H.

[3] American Academy of Family Physicians, American Academy of Pediatrics, American College of Physicians, and American Osteopathic Association. *Joint Principles of the Patient-Centered Medical Home.* 2007. (Accessed March 7, 2016: https://www.acponline.org/acp_policy/policies/joint_principles_pcmh_2007.pdf.)

[4] Barr MS. The need to test the patient-centered medical home. *JAMA.* 2008;300(7):834-835.

[5] Nielsen M, Gibson L, Buelt L, Grundy P, Grumbach K. *The Patient-Centered Medical Home's Impact on Cost and Quality: Review of Evidence, 2013-2014*. Washington, DC: Patient-Centered Primary Care Collaborative; 2015. (Accessed March 5, 2016: https://www.pcpcc.org/resource/patient-centered-medical-homes-impact-cost-and-quality#sthash.a8kLOb70.dpuf.)

[6] Peikes D, Zutshi A, Genevro JL, Parchman ML, Meyers DS. Early evaluations of the medical home: Building on a promising start. *Am J Manag Care*. 2012;18(2):105-116.

[7] Zutshi A, Peikes D, Smith K, et al. *The Medical Home: What Do We Know? What Do We Need to Know? A Review of the Current State of Evidence on the Effects of the Patient-Centered Medical Home Model*. Rockville, MD: Agency for Healthcare Research and Quality; 2014. (Accessed March 7, 2016: https://pcmh.ahrq.gov/sites/default/files/attachments/the-medical-home-what-do-we-know.pdf.)

[8] *The Future of Patient-Centered Medical Homes*. Washington, DC: NCQA; 2014. (Accessed January 14, 2016: http://www.ncqa.org/Portals/0/Public%20Policy/2014%20Comment%20Letters/The_Future_of_PCMH.pdf.)

[9] U.S. Congress. Public Law 114-10. August 16, 2015. "The Medicare Access and CHIP Reauthorization Act of 2015." 2015. (Accessed March 7, 2016: https://www.congress.gov/114/plaws/publ10/PLAW-114publ10.pdf.)

[10] Phillips RL, Jr., Han M, Petterson SM, Makaroff L, Liaw WR. Cost, utilization, and quality of care: An evaluation of Illinois' Medicaid primary care case management program. *Annals of Family Medicine*. 2014;12(5):408-417.

[11] Rosenthal M, Abrams M, Bitton A, Collaborative T. *Recommended Core Measures for Evaluating the Patient-Centered Medical Home: Cost, Utilization, and Clinical Quality*. New York, NY: The Commonwealth Fund; 2012. (Accessed March 6, 2016: http://www.commonwealthfund.org/~/media/Files/Publications/Data%20Brief/2012/1601_Rosenthal_recommended_core_measures_PCMH_v2.pdf.)

[12] Cook TD, Campbell DT. *Quasi-experimentation: Design and analysis for field settings.* Rand McNally; 1979.

[13] Institute of Education Sciences. *What Works Clearinghouse Procedures and Standards Handbook.* Version 3.0. (See listing in Resources.)

[14] Peikes D, Dale S, Lundquist E, Genevro J, Meyers D. *Building the Evidence Base for the Medical Home: What Sample and Sample Size Do Studies Need?* AHRQ Publication No. 11-0100-EF. Rockville, MD: Agency for Healthcare Research and Quality; 2011. (Accessed March 5, 2016: https://pcmh.ahrq.gov/sites/default/files/attachments/Building%20Evidence%20Base%20PC MH%20White%20Paper.pdf.) With an ICC of 0, the number of unique observations is the number of patients.

[15] With an ICC of 1, the number of unique observations reduces to the number of provider organizations included in the study. As an example, with 20 provider organizations (10 in the intervention group and 10 in a comparison group), the minimum detectable effect (MDE) for reducing total cost of care rises from 5% to 66% when there is a moderate degree of clustering. This is an implausibly large reduction in costs required before it is likely to be considered statistically significant (Peikes *et al*. 2011).

The Authors

Nancy McCall is a senior fellow at Mathematica Policy Research, which provides a full range of research and data collection services, including program evaluation and policy research, survey design, and data collection.

McCall has 30 years of experience conducting health services research assessing the effects of health system transformation on quality of care, health care utilization, costs, and health outcomes. Her research and technical assistance efforts have focused on a wide range of subjects, including the evaluation of health care transformation demonstrations within the Medicare and Medicaid programs; federal and state health care delivery reforms; and access to and quality of care for consumers within federal, state, and commercial insurance programs. She has extensive experience leading successful evaluations of health care delivery and payment reform demonstrations that require the use of rigorous quantitative, qualitative, and mixed methods and analysis of a broad set of outcomes using an array of quantitative data, including Medicare, Medicaid, commercial, and all-payer claims databases.

Before joining Mathematica, McCall was a chief scientist at RTI International. She was the principal investigator for the Evaluation of the Multi-Payer Advanced Primary Care Practice Demonstration, the Medicare Medical Home Demonstration Evaluation, the High Cost Medicare Beneficiary Demonstration Evaluation, and the Medicare Health Support Evaluation.

McCall was a cardiac research nurse at Brigham and Women's Hospital and a primary nurse at the Sidney Farber Cancer Institute, both in Boston, Massachusetts. She also served as an assistant research professor in the Clinical Economics Research Unit at Georgetown University Medical Center and as an economist at the Health Care Financing Administration. She holds a BS in nursing from Purdue University, an MS in health policy, and a ScD in health economics from the Harvard School of Public Health. McCall's work has been published in the *New England Journal of Medicine*, *Medical Care*, *Medical Care Research and Review*, and *Health Services Research*.

Kristin Geonnotti is an associate director and senior researcher at Mathematica Policy Research. She has expertise in primary care redesign and transformation, the patient-centered medical home, and quality of care. She works on numerous mixed methods evaluations of health policies and programs, focused primarily on evaluating delivery system reform initiatives.

Geonnotti's work on large federal evaluations for the Centers for Medicare & Medicaid Services (CMS) includes leading both Medicare claims analyses and implementation evaluations. She currently works on three major evaluations for CMS, which focus on a range of health care delivery and payment reforms that aim to achieve better quality of care; improve population health; and lower costs for Medicare, Medicaid, and Children's Health

Insurance Program beneficiaries. She also examined strategies for improving the delivery of primary care through medical homes for the Agency for Healthcare Research and Quality. Her research contributions included a paper describing strategies for practice facilitators (coaches) to engage primary care practices in ongoing quality improvement and a guide on developing practice facilitation programs. In addition, Geonnotti has provided technical assistance to states and payers on best practices for collecting and analyzing performance measures related to medical homes implementation. She joined Mathematica in 2010 and holds a PhD in health policy and management from the University of North Carolina at Chapel Hill.

Milbank Memorial Fund
645 Madison Avenue
New York, NY 10022
www.milbank.org

The Milbank Memorial Fund is an endowed operating foundation that engages in nonpartisan analysis, study, research, and communication on significant issues in health policy. In the Fund's own publications, in reports, films, or books it publishes with other organizations, and in articles it commissions for publication by other organizations, the Fund endeavors to maintain the highest standards for accuracy and fairness. Statements by individual authors, however, do not necessarily reflect opinions or factual determinations of the Fund.

ISBN 978-1887748-80-3

About the Milbank Memorial Fund

The Milbank Memorial Fund is an endowed operating foundation that works to improve the health of populations by connecting leaders and decision makers with the best available evidence and experience. Founded in 1905, the Fund engages in nonpartisan analysis, collaboration, and communication on significant issues in health policy. It does this work by publishing high-quality, evidence-based reports, books, and *The Milbank Quarterly*, a peer-reviewed journal of population health and health policy; convening state health policy decision makers on issues they identify as important to population health; and building communities of health policymakers to enhance their effectiveness. www.milbank.org.