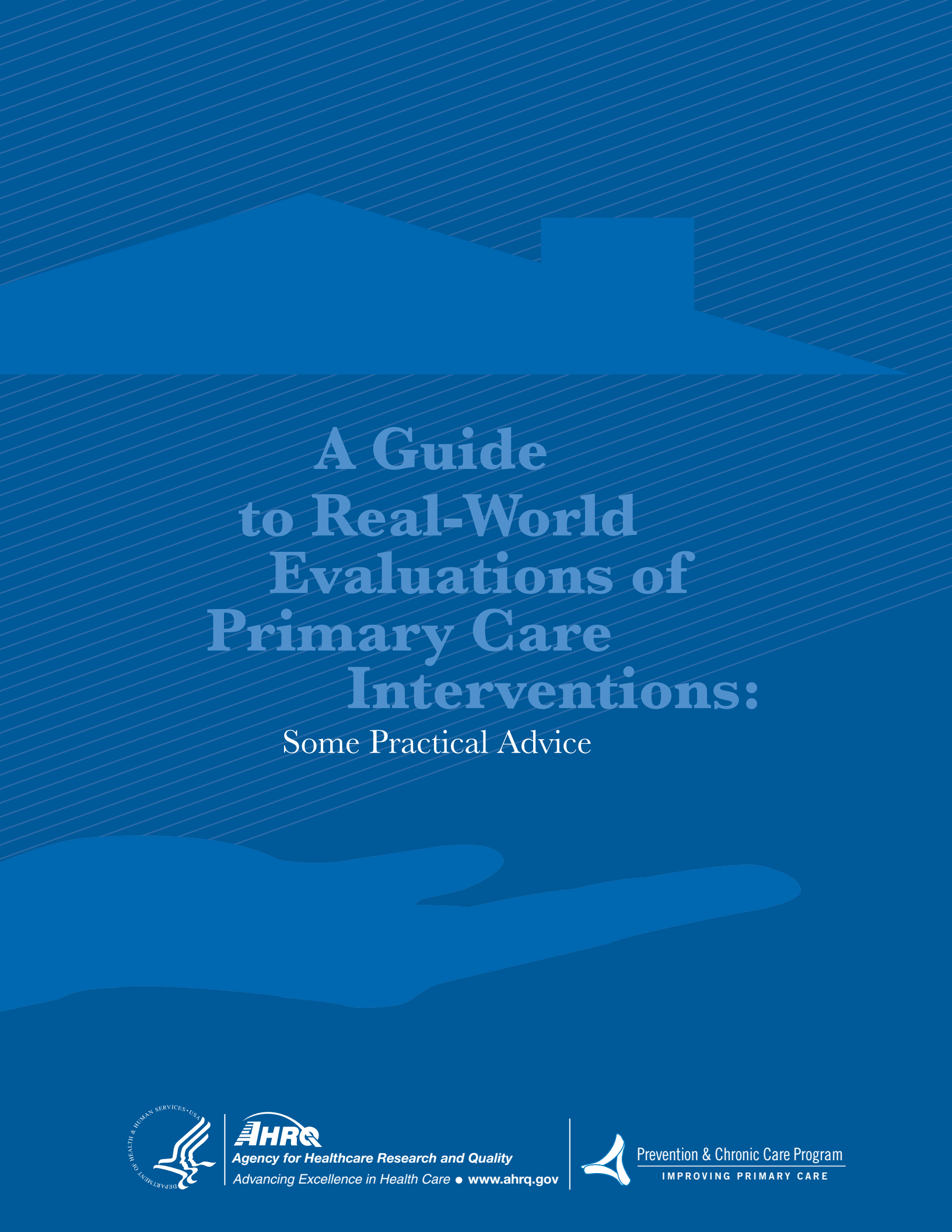


The background of the cover is a solid blue color with a subtle pattern of fine, light blue diagonal lines. In the upper half, there is a dark blue silhouette of a mountain range with a small, square-shaped building or structure on the right side. In the lower half, there is a dark blue silhouette of a body of water with a small, irregular landmass or island in the center.

A Guide to Real-World Evaluations of Primary Care Interventions: Some Practical Advice



Agency for Healthcare Research and Quality
Advancing Excellence in Health Care • www.ahrq.gov



Prevention & Chronic Care Program
IMPROVING PRIMARY CARE

None of the authors has any affiliations or financial involvement that conflicts with the material presented in this guide. The authors gratefully acknowledge the helpful comments on earlier drafts provided by Drs. Eric Gertner, Lehigh Valley Health Network; Michael Harrison, AHRQ; Malaika Stoll, Sutter Health; and Randall Brown and Jesse Crosson, Mathematica Policy Research. We also thank Cindy George and Jennifer Baskwell of Mathematica for editing and producing the document.

A Guide to Real-World Evaluations of Primary Care Interventions: Some Practical Advice

Prepared for:

Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

Prepared by:

Mathematica Policy Research, Princeton, NJ
Project Director: Deborah Peikes
Principal Investigators: Deborah Peikes and Erin Fries Taylor

Authors:

Deborah Peikes, Ph.D., M.P.A., Mathematica Policy Research
Erin Fries Taylor, Ph.D, M.P.P, Mathematica Policy Research
Janice Genevro, Ph.D., Agency for Healthcare Research and Quality
David Meyers, M.D., Agency for Healthcare Research and Quality

October 2014
AHRQ Publication No. 14-0069-EF

Quick Start to This Evaluation Guide

Goals. Effective primary care can improve health and cost outcomes, and patient, clinician and staff experience, and evaluations can help determine *how best* to improve primary care to achieve these goals. This Evaluation Guide provides practical advice for designing real-world evaluations of interventions such as the patient-centered medical home (PCMH) and other models to improve primary care delivery.

Target audience. This Guide is designed for evaluators affiliated with delivery systems, employers, practice-based research networks, local or regional insurers, and others who want to test a new intervention in a relatively small number of primary care practices, and who have limited resources to evaluate the intervention.

Summary. This Guide presents some practical steps for designing an evaluation of a primary care intervention in a small number of practices to assess the implementation of a new model of care and to provide information that can be used to guide possible refinements to improve implementation and outcomes. The Guide offers options to address some of the challenges that evaluators of small-scale projects face, as well as insights for evaluators of larger projects. Sections I through V of this Guide answer the questions posed below. A **resource collection** in Section VI includes many AHRQ-sponsored resources as well as other tools and resources to help with designing and conducting an evaluation. Several appendices include additional technical details related to estimating quantitative effects.

- I. **Do I need an evaluation?** Not every intervention needs to be evaluated. Interventions that are minor or inexpensive, have a solid evidence base, or are part of quality improvement efforts may not warrant an evaluation. But many interventions would benefit from study. To decide whether to conduct an evaluation, it's important to identify the specific decisions the evaluation is expected to inform and to consider the cost of carrying out the evaluation. An evaluation is useful for interventions that are substantial and expensive and lack a solid evidence base. It can answer key questions about whether and how an intervention affected the ways practices deliver care and how changes in care delivery in turn affected outcomes. Feedback on implementation of the model and early indicators of success can help refine the intervention. Evaluation findings can also help guide rollout to other practices. One key question to consider: Can the evaluation that you have the resources to conduct generate reliable and valid findings? Biased estimates of program impacts would mislead stakeholders and, we contend, could be worse than having no results at all. This Guide has information to help you determine whether an evaluation is needed and whether it is the right choice given your resources and circumstances.
- II. **What do I need for an evaluation?** Understanding the resources needed to launch an intervention and conduct an evaluation is essential. Some resources needed for evaluations include (1) leadership buy-in and support, (2) data, (3) evaluation skills, and (4) time for the evaluators and the practice clinicians and staff who will provide data to perform their roles. It's important to be clear-sighted about the cost of conducting a well-designed evaluation and to consider these costs in relation to the nature, scope, and cost of the intervention.

- III. [How do I plan an evaluation?](#) It's best to design the evaluation before the intervention begins, to ensure the evaluation provides the highest quality information possible. Start by determining your purpose and audience so you can identify the right research questions and design your evaluation accordingly. Next, take an inventory of resources available for the evaluation and align your expectations about what questions the evaluation can answer with these resources. Then describe the underlying logic, or theory of change, for the intervention. You should describe why you expect the intervention to improve the outcomes of interest and the steps that need to occur before outcomes would be expected to improve. This logic model will guide what you need to measure and when, though you should remain open to unexpected information as well as consequences that were unintended by program designers. The logic model will also help you tailor the scope and design of your evaluation to the available resources.
- IV. [How do I conduct an evaluation, and what questions will it answer?](#) The next step is to design a study of the intervention's implementation and—if you can include enough practices to potentially detect statistically significant changes in outcomes—a study of its impacts. Evaluations of interventions tested in a small number of practices typically can't produce reliable estimates of effects on cost and quality, despite stakeholders' interest in these outcomes. In such cases, you can use qualitative analysis methods to understand barriers and facilitators to implementing the model and use quantitative data to measure interim outcomes, such as changes in care processes and patient experience, that can help identify areas for refinement and the potential to improve outcomes.
- V. [How can I use the findings?](#) Findings from implementation evaluations can indicate whether it is feasible for practices to implement the intervention and ways to improve the intervention. Integrating the implementation and impact findings (if you can conduct an impact evaluation) can (1) provide a more sophisticated understanding about the effects of the model being tested; (2) identify types of patients, practices, and settings that may benefit the most; and (3) guide decisions about refinement and spread.
- VI. [What resources are available to help me?](#) The resource collection in this Guide contains resources and tools that you can use to develop a logic model, select implementation and outcome measures, design and conduct analyses, and synthesize implementation and impact findings.

I. Do I Need an Evaluation?

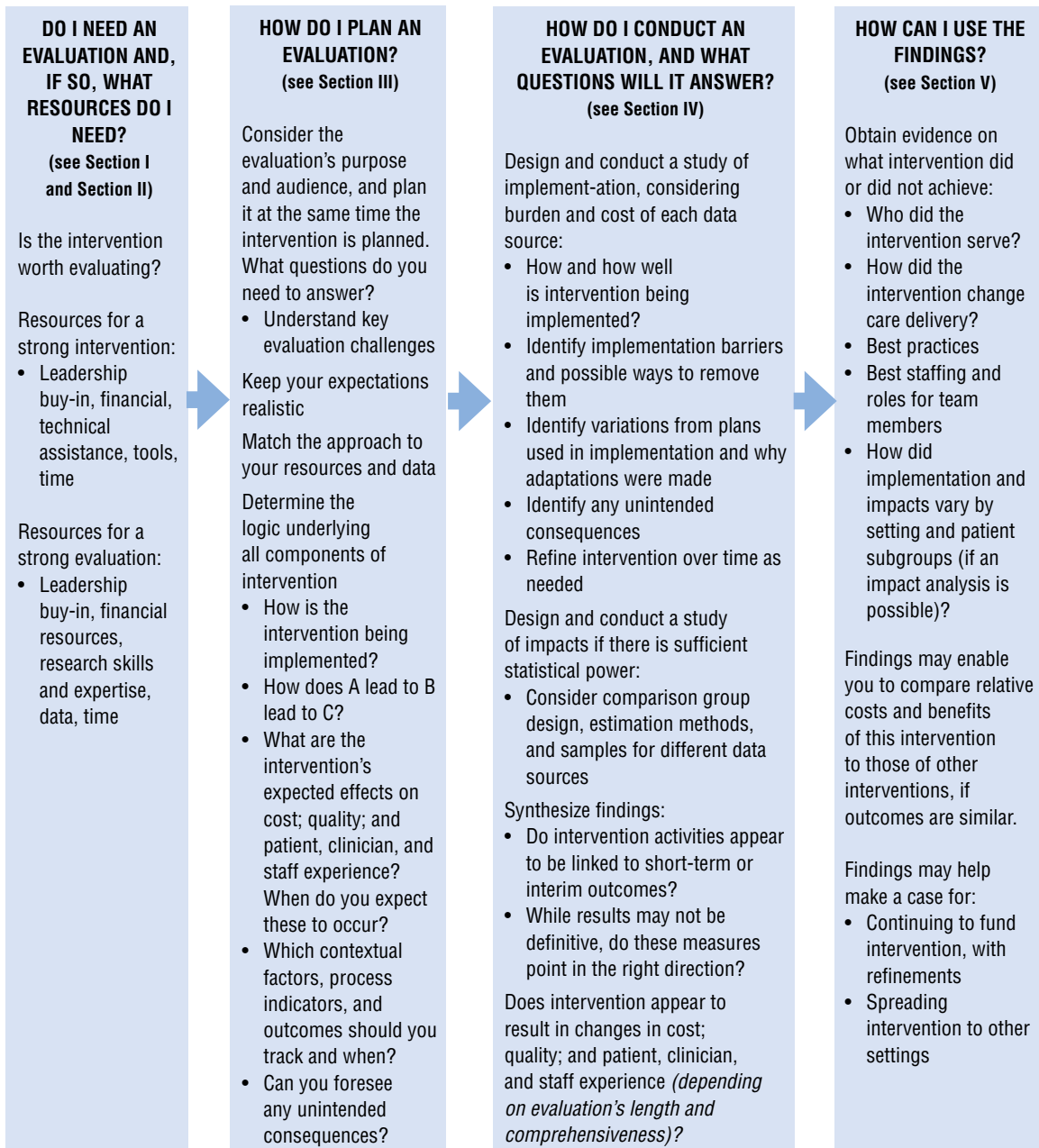
Your organization has decided to try to change the way primary care practices deliver care, in the hope of improving important outcomes. The first question to ask is whether you should evaluate the intervention.

Not all primary care interventions require an evaluation. When it is clear that a change needs to be made, the practice may simply move to adoption. For example, if patients are giving feedback about lack of evening hours, and business is being lost to urgent care centers, then a primary care practice might decide to add evening hours without evaluating the change. You may still want to track utilization and patient feedback about access, but a full evaluation of the intervention may not be warranted. In addition, some operational and quality improvement changes can be assessed through locally managed Plan-Do-Study-Act cycles. Examples of such changes include changing appointment lengths and enhancing educational materials for patients. Finally, when previously published studies have provided conclusive evidence in similar settings with similar populations, you do not need to re-test those interventions.

A more rigorous evaluation may be beneficial if it is costly to adopt the primary care intervention and if your organization is considering whether to spread the intervention extensively. An evaluation will help you learn as much as possible about how best to implement the intervention and how it might affect outcomes. You can examine whether it is possible for practices to make the changes you want, how to roll out this (or a refined intervention) more smoothly, and whether the changes made through the intervention sufficiently improve outcomes to justify the effort. You also may be able to ascertain how outcomes varied by practice, market, and patient characteristics. Outcomes of interest typically include health care cost and quality, and patient, clinician, and staff experience. Results from the implementation and impact analyses can help make a case for refining the intervention, continuing to fund it, and/or spreading it to more practices, if the effects of the intervention compare favorably to its costs.

Figure 1 summarizes the steps involved in planning and implementing an evaluation of a primary care intervention; the two boxes on the right-hand side show the evaluation's benefits.

Figure 1. Steps in Planning and Implementing an Evaluation of a Primary Care Intervention



COMMON CHALLENGES IN EVALUATING PRIMARY CARE INTERVENTIONS

Timeframes are too short or intervention too minor to observe changes in care delivery and outcomes.

Small numbers of practices make it hard to detect effects statistically due to clustering.

Data are limited, of poor quality, or have a significant time lag.

Results are not generalizable because practices participating in intervention are different from other practices (e.g., participants may be early adopters).

Outcomes may improve or decline for reasons other than participation in the intervention and the comparison group or evaluation design may not adequately account for this.

Differences exist between intervention practices and comparison practices even before the intervention begins.

Comparison practices get some form or level of intervention.

II. What Do I Need for an Evaluation?

A critical step is understanding and obtaining the resources needed for successfully planning and carrying out your evaluation. The resources for conducting an intervention and evaluation are shown in Table 1 and Figure 1. We suggest you take stock of these items during the early planning phase for your evaluation. Senior management and others in your organization may need to help identify and commit needed resources.

The *resources available for the intervention* are linked to your evaluation because they affect (1) the extent to which practices can transform care and (2) the size of expected effects. How many practices can be transformed? How much time do staff have available to implement the changes? What payments, technical assistance to guide transformation, and tools (such as shared decisionmaking aids or assistance in developing patient registries) will practices receive? Are additional resources available through new or existing partnerships? Is this intervention package substantial enough to expect changes in outcomes? Finally, how long is it likely to take practices to change their care delivery, and for these changes to improve outcomes?

Similarly, the *resources available for your evaluation* of the intervention help shape the potential rigor and depth of the evaluation. You will need data, research skills and expertise, and financial resources to conduct an evaluation. Depending on the skills and expertise available internally, an organization may identify internal staff to conduct the evaluation, or hire external evaluators to conduct the evaluation or collaborate and provide guidance on design and analysis. External evaluators often lend expertise and objectivity to the evaluation. Regardless of whether the evaluation is conducted by internal or external experts or a combination, ongoing support for the evaluation from internal staff—for example, to obtain claims data and to participate in interviews and surveys—is critical. The amount of time available for the evaluation will affect the outcomes you can measure, due to the time needed for data collection, as well as the time needed for outcomes to change.

Inventory the financial, research, and data resources you can devote to the evaluation, and adjust your evaluation accordingly.

Table 1. Inventory of Resources Needed for Testing a Primary Care Intervention

Resource Type	Examples
Resources for Intervention	
Leadership buy-in	Motivation and support for trying the intervention.
Financial resources	Funding available for the intervention (including the number of practices that can test it).
Technical assistance	Support available to help practices transform such as data feedback, practice facilitation/coaching, expert consultation, learning collaboratives, and information technology (IT) expertise.
Tools	Tools for practices such as registries, health IT, and shared decisionmaking tools.
Time	Allocated time of staff to implement the intervention; elapsed time for practices to transform and for outcomes to change.
Resources for Evaluation	
Leadership buy-in	Motivation and support for evaluating the intervention.
Financial resources	Funding available for the evaluation, including funds to hire external evaluation staff if needed.
Research skills, expertise, and commitment	Skills and expertise in designing evaluations, using data, conducting implementation and impact analyses, and drawing conclusions from findings. Motivation and buy-in of evaluation staff and other relevant stakeholders, such as clinicians and staff who will provide data. Expertise in designing the evaluation approach and analysis plan, creating files containing patient and claims data, and conducting analyses.
Data	Depending on the research questions, could include claims, electronic medical records, paper charts, patient intake forms, care plans, patient surveys, clinician and practice staff surveys, registries, care management tracking data, qualitative data from site visit observations and interviews, and other information (including the cost of implementing the intervention). Data should be of adequate quality.
Time	Time to obtain and analyze data and for outcomes to change.

III. How Do I Plan an Evaluation?

Develop the evaluation approach before the pilot begins.

Start planning your evaluation as early as you can. Careful and timely planning will go a long way toward producing useful results, for several reasons. First, you may want to collect pre-intervention data and understand the decisions that shaped the choice of the intervention and the selection of practices for participation. Second, you may want to capture early experiences with implementing the intervention to understand any challenges and refinements made. Finally, you may want to suggest minor adaptations to the intervention's implementation to enhance the evaluation's rigor. For example, if an organization wanted to implement a PCMH model in five practices at a time, the evaluator might suggest randomly picking the five practices from those that meet eligibility criteria. This would make it possible to compare any changes in care delivery and outcomes of the intervention practices to changes in a control group of eligible practices that will adopt the PCMH model later. If the project had selected the practices before consulting with the evaluator, the evaluation might have to rely on less rigorous nonexperimental methods.

Consider the purpose and audience for the evaluation. Identifying stakeholders who will be interested in the evaluation's results, the decisions that your evaluation is expected to inform, and the type of evidence required is crucial to determining what questions to ask and how to approach the evaluation. For example, insurers might focus on the intervention's effects on the total medical costs they paid and on patient satisfaction; employers might be concerned with absentee rates and workforce productivity; primary care providers might focus on practice revenue and profit, quality of care, and staff satisfaction; and labor unions might focus on patient functioning, satisfaction, and out-of-pocket costs. Potential adverse effects of the intervention and the reporting burden from the evaluation should be considered as well.

Who is the audience for the evaluation?
What questions do they want answered?

Consider, too, the form and rigor of the evidence the stakeholders need. Perspectives differ on how you should respond to requests for information from funders or other stakeholders when methodological issues mean you cannot be confident in the findings. We recommend deciding during the planning stage how to approach and reconcile tradeoffs between rigor and relevance. Sometimes the drawbacks of a possible evaluation—or certain evaluation components—are serious enough (for example, if small sample size and resulting statistical power issues will render cost data virtually meaningless) that resources should not be used to generate information that is likely to be misleading.

Questions to ask include: Do the stakeholders need numbers or narratives, or a combination? Do stakeholders want ongoing feedback to refine the model as it unfolds, an assessment of effectiveness at the end of the intervention, or both? Do the results need only to be suggestive of positive effects, or must they rigorously demonstrate robust impacts for stakeholders to act upon them? How large must the effects be to justify the cost of the intervention? Thinking through these issues will help you choose the outcomes to measure, data collection approach, and analytic methods.

Understand the challenges of evaluating primary care interventions. Some of the challenges to evaluating primary care interventions include (see also the bottom box of Figure 1):

- ▲ *Time and intensity needed to transform care.* It takes time for practices to transform, and for those transformations to alter outcomes.¹ Many studies suggest it will take a minimum of 2 or 3 years for motivated practices to really transform care.^{2, 3, 4, 5, 6} If the intervention is short or it is not substantial, it will be difficult to show changes in outcomes. In addition, a short or minor intervention may only generate small effects on outcomes, which are hard to detect.
- ▲ *Power to detect impacts when clustering exists.* Even with a long, intensive intervention, clustering of outcomes at the practice level^a may make it difficult for your evaluation to detect anything but very large effects without a large number of practices. For example, some studies spend a lot of time and resources collecting and analyzing data on the cost effects of the PCMH model. However, because of clustering, an intervention with too few practices might have to generate cost reductions of more than 70 percent for the evaluation to be confident that observed changes are statistically significant.⁷ As a result, if an evaluation finds that the estimated effects on costs are not statistically significant, it's not clear whether the intervention was ineffective or the evaluation had low statistical power (see Appendix A for a description of how to calculate statistical power in evaluations).
- ▲ *Data collection.* Obtaining accurate, complete, and timely data can be a challenge. If multiple payers participate in the intervention, they may not be able to share data; if no payers are involved, the evaluator may be unable to obtain data on service use and expenditures outside the primary care practice.
- ▲ *Generalizability.* If the practices that participate are not typical, the results may not be generalizable to other practices.
- ▲ *Measuring the counterfactual.* It is difficult to know what would have occurred in the intervention practices in the absence of the intervention (the “counterfactual”). Changing trends over time make it hard to ascribe changes to the intervention without identifying an appropriate comparison group, which can be challenging. In addition, multiple interventions may occur simultaneously or the comparison group may undertake changes similar to those found in the intervention, which can complicate the evaluation.

Adjust your expectations so they are realistic, and match the evaluation to your resources. The goal of your evaluation is to generate the highest quality information possible within the limits of your resources. Given the challenges of evaluating a primary care intervention, it is better to attempt to answer a narrow set of questions well than to study a broad set of questions but not provide definitive or valid answers to any of them. As described above, you need adequate resources for the intervention and evaluation to make an impact evaluation worthwhile.

^a Clustering arises because patients within a practice often receive care that is similar to that received by other patients served by the practice, but different from the care received by patients in other practices. This means that patients from a given practice cannot necessarily be considered statistically independent from one another, and the effective size of the patient sample is decreased.

With limited resources, it is often better to scale back the evaluation. For example, an evaluation that focuses on understanding and improving the implementation of an intervention can identify early steps along the pathway to lowering costs and improving health care. We recommend using targeted interviews to understand the experiences of patients, clinicians, staff, and other stakeholders, and measuring just a few intermediate process measures, such as changes in workflows and the use of health information technology. Uncovering any challenges encountered with these early steps can allow for refinement of the intervention before trying out a larger-scale effort.

Describe the logic model, or theory of change, showing why and how the intervention might improve outcomes of interest. In this step, the evaluators and implementers work together to describe each component of the intervention, the pathways through which they could affect outcomes of interest, and the types of effects expected in the coming months and years. Because primary care interventions take place in the context of the internal practice and the external health care environments, the logic model should identify factors that might affect outcomes—either directly or indirectly—by affecting implementation of the intervention.

Consider all factors, even if you may not be able to collect data on all of them, and you may not have enough practices to control for each factor in regression analyses to estimate impacts. Practice- or organization-specific factors include, for example, patient demographics and language, size of patient panels, practice ownership, and number and type of clinicians and staff. Other examples of practice-specific factors include practice leadership and teamwork.⁸ Factors describing the larger health care environment include practice patterns of other providers, such as specialists and hospitals, community resources, and payment approaches of payers. Intervention components should include the aspects of the intervention that vary across the practices in your study, such as the type and amount of services delivered to provide patient-centered, comprehensive, coordinated, accessible care, with a systematic focus on quality and safety. They may also include measures capturing variation across intervention

Find resources on logic models and tools to conduct implementation and impact studies in the Resource Collection.

practices in the offer and receipt of: technical assistance to help practices transform; additional payments to providers and practices; and regular feedback on selected patient outcomes, such as health care utilization, quality, and cost metrics.

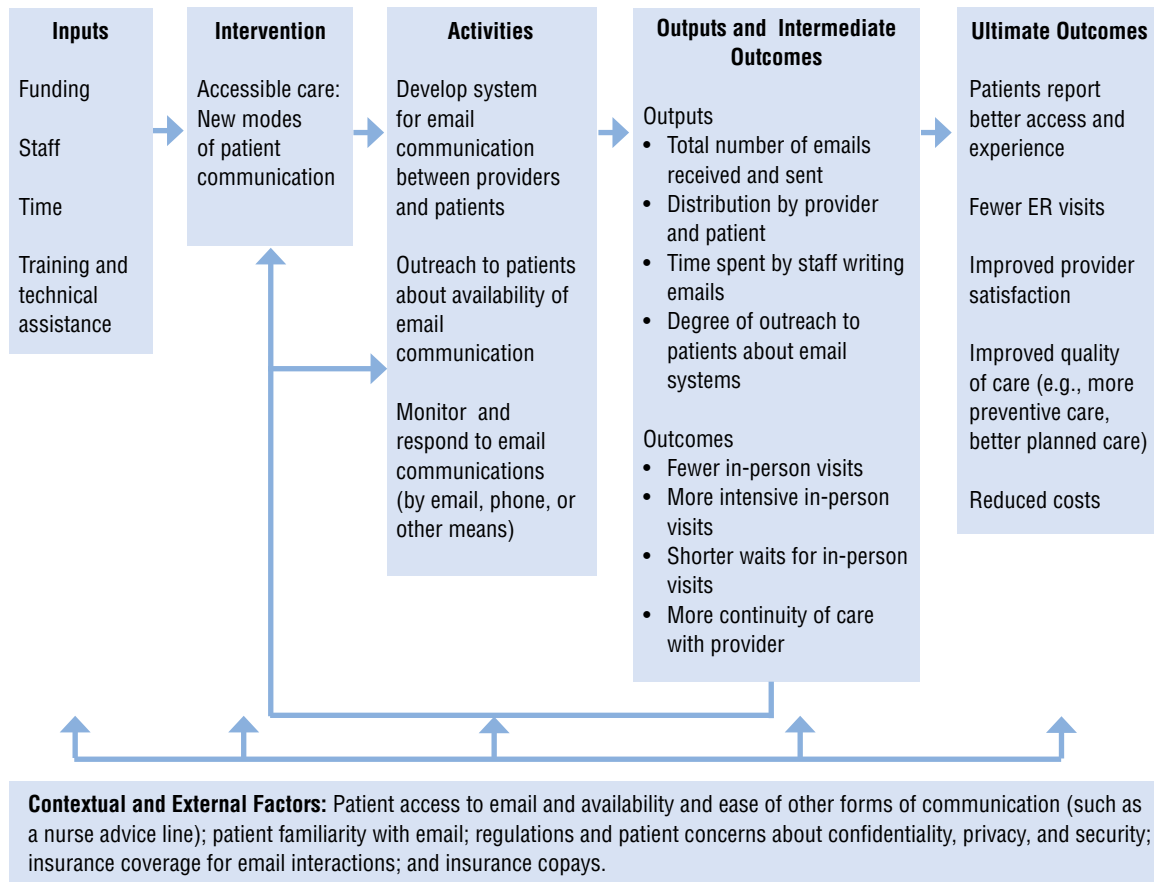
A logic model serves several purposes (see Petersen, Taylor, and Peikes⁹ for an illustration and references). It can help implementers recognize gaps in the logic of transformation early so they can take appropriate steps to modify the intervention to ensure success. As an evaluator, you can use the logic model approach to determine at the outset whether the intervention has a strong underlying logic and a reasonable chance of improving outcomes, and what effect sizes the intervention might need to produce to be likely to yield statistically significant results. In addition, the logic model will help you decide what to measure at different points in time to show whether the intervention was implemented as intended, improved outcomes, and created unintended outcomes, and identify any facilitators and barriers to implementation. However, while the logic model is important, you should remain open to unexpected information, too. Finally, the logic model's depiction of how the intervention is intended

The evaluator and implementers should work together to describe the theory of change. The logic model will guide what to measure, and when to do so.

to work can be useful in helping you interpret findings. For example, if the intervention targets more assistance to practices that are struggling, the findings may show a correlation between more assistance and worse outcomes. Understanding the specifics of the intervention approach will prevent you from mistakenly interpreting such a finding as indicating that technical assistance worsens outcomes.

As an example of this process, consider a description of the pathway linking implementation to outcomes for expanded access, one of the features the medical home model requires. Improved access is intended to improve continuity of care with the patient's provider and reduce use of the emergency room (ER) and other sites of care. If the intervention you are evaluating tests this approach, you could consider how the medical home practices will expand access. Will the practices use extended hours, email and telephone interactions, or have a nurse or physician on call after hours? How will the practices inform patients of the new options and any details about how to use them? Because nearly all interventions are adapted locally during implementation, and many are not implemented fully, the logic model should specify *process indicators* to document how the practices implemented the approach. For practices that use email interactions to increase access, some process indicators could include how many patients were notified about the option by mail or during a visit, the overall number of emails sent to and from different practice staff, the number and distribution by provider and per patient, and time spent by practice staff initiating and responding to emails (Figure 2). You could assess which process indicators are easiest to collect, depending on the available data systems and the feasibility of setting up new ones. To decide which measures to collect, consider those likely to reflect critical activities that must occur to reach intended outcomes, and balance this with an understanding of the resources needed to collect the data and the impact on patient care and provider workflow.

Figure 2. Logic Model of a PCMH Strategy Related to Email Communication



Source: Adapted from Petersen, Taylor, and Peikes. Logic Models: The Foundation to Implement, Study, and Refine Patient-Centered Medical Home Models, 2013, Figure 2.⁹

Expected changes in *intermediate outcomes* from enhanced email communications might include the following:

- ▲ Fewer but more intensive in-person visits as patients resolve straightforward issues via email
- ▲ Shorter waits for appointments for in-person visits
- ▲ More continuity of care with the same provider, as patients do not feel the need to obtain visits with other providers when they cannot schedule an in-person visit with their preferred provider

Ultimate outcomes might include:

- ▲ Patients reporting better access and experience
- ▲ Fewer ER visits as providers can more quickly intervene when patients experience problems
- ▲ Improved provider experience, as they can provide better quality care and more in-depth in-person visits
- ▲ Lower costs to payers from improved continuity, access, and quality

You should also track *unintended outcomes* that program designers did not intend but might occur. For example, using email could reduce practice revenue if insurers do not reimburse for the interactions and fewer patients come for in-person visits. Alternatively, increased use of email could lead to medical errors if the absence of an in-person visit leads the clinician to miss some key information, or could lead to staff burnout if it means staff spend more total time interacting with patients.

Some *contextual factors* that might influence the ability of email interactions to affect intended and unintended outcomes include: whether patients have access to and use email and insurers reimburse for the interactions; regulations and patient concerns about confidentiality, privacy, and security; and patient copays for using the ER.

What outcomes are you interested in tracking? The following resources are a good starting point for developing your own questions and outcomes to track:

- ▲ The Commonwealth Fund's PCMH Evaluators Collaborative provides a list of core outcome measures. www.commonwealthfund.org/Publications/Data-Briefs/2012/May/Measures-Medical-Home.aspx
- ▲ The Consumer Assessment of Healthcare Providers and Systems (CAHPS®) PCMH survey instrument developed by the Agency for Healthcare Research and Quality (AHRQ) provides patient experience measures. <https://cahps.ahrq.gov/surveys-guidance/cg/pcmh/index.html>.

IV. How Do I Conduct an Evaluation, and What Questions Will It Answer?

You've planned your evaluation, and now the work of design, data collection, and analysis begins.

There are two broad categories of evaluations: studies of implementation and studies of impact. Using both types of studies together provides a comprehensive set of findings. However, if the number of practices is small, it is difficult to detect a statistically significant “impact” effect on outcomes such as cost and utilization. In such cases, you could review larger, published studies to learn about the impact of the intervention, and focus on designing and conducting an implementation study to understand how best to implement the intervention in your setting.

See the Resource Collection for resources on designing and conducting an evaluation.

Design and conduct a study of implementation. Some evaluators may want to focus exclusively on measuring an intervention's effects on cost, quality, and experiences of patients, families, clinicians, and staff. However, you can learn a great deal from a study of how the intervention was implemented, the degree to which it was implemented according to plan in each practice, and the factors explaining both purposeful and unintended deviations. This includes collecting and analyzing information on: the practices participating in and patients served by the intervention; how the intervention changed the way that practices delivered care, how this varied in intended and unintended ways, and why; and any barriers and facilitators to successful implementation and achieving the outcomes of interest.

Whenever possible, data collection should be incorporated into existing workflows at the practice to minimize the burden on clinicians and other staff. You should consider the cost of collecting each data source in terms of burden on respondents and cost to the evaluation.

We recommend including both quantitative and qualitative data when studying how your intervention is being implemented. Although implementation studies tend to rely heavily on qualitative data, using some quantitative data sources (a mixed-methods approach) can amplify the usefulness of the findings.^{10, 11, 12}

If resources do not permit intensive data collection, a streamlined approach to studying implementation might rely on discussions with practice clinicians, staff, and patients and their families involved in or affected by the intervention; analysis of any data already being collected in tracking systems or medical charts; and review of available documents, as follows:

An implementation study can provide invaluable insights and often can be done inexpensively.

- ▲ **Interviews and informal discussions with patients and their families** provide their perceptions of and experiences with care.
- ▲ **Interviews and discussions with practice clinicians and staff** over the course of the intervention (including clinicians, care managers, nurses, medical assistants, front office staff, and other staff) using semi-structured discussion guides will provide information on how (and how consistently) they implemented the intervention, their general perceptions of it, how it changed their interactions and work with patients, whether they think it improved patient care and other outcomes effectively, whether it gained buy-in from practice leadership and staff, its financial viability, and its strengths and areas for improvement.

- ▲ **Data from a tracking system** are typically inexpensive to gather and analyze, if a system is already in place for operational reasons. The tracking system might document whether commonly used approaches in new primary care models such as care management, patient education, and transitional support are implemented as intended (for example, after patients are discharged from the hospital or diagnosed with a chronic condition). If a tracking system is in place, modifying it to greatly enhance its usefulness for research is often relatively easy.
- ▲ **Medical record reviews** can be inexpensive to conduct for a small sample. Such reviews can illustrate whether and how well certain aspects of the primary care intervention have been implemented. For example, you might review electronic charts to determine the proportion of patients for whom the clinician provided patient education or developed and discussed a care plan. You could also look more broadly at the effects of various components of the intervention on different patients with different characteristics. You could select cases to review randomly or focus on patients with specific characteristics, such as those who have chronic illness; no health problems; or a need for education about weight, smoking, or substance abuse. Unwanted variation in care for patients, as well as differences in care across providers, might also be of interest.
- ▲ **Review of documents**, including training manuals, protocols, feedback reports to practices, and care plans for patients, among others, can provide important details on the components of the intervention.

This information can be relatively inexpensive to collect and can provide insights about how to improve the intervention and why some outcome goals were achieved but others were not.

With more resources, your implementation study might also collect data from the following data sources:

- ▲ **Surveys with patients and their families** can be used to collect data from a large sample of patients. The surveys might ask about the care patients receive in their primary care practices (including accessibility, continuity, and comprehensiveness); the extent to which it is patient-centered and well coordinated across the medical neighborhood of other providers; and any areas for improvement.
- ▲ **Focus groups with patients and families** allow for active and engaged discussion of perspectives, issues, and ideas, with participants building on one another's thinking. These can be particularly useful for testing out hypotheses and developing possible new approaches to patient care challenges.
- ▲ **Site visits to practices** can enable you to directly observe team functioning, workflow, and interactions with patients to supplement interviews with practice staff.
- ▲ **Surveys of practice clinicians and staff** can provide data from a large number of clinicians and staff about how the intervention affects the experience of providing care.
- ▲ **Medical record reviews of a larger sample of patients** can provide a more comprehensive assessment of how the team provided care.

Your analysis should synthesize data from multiple sources to answer each research question. Comparing and contrasting information across sources strengthens the findings considerably and yields a more complete understanding of implementation. Organizing the information by the question you are trying to answer rather than by data source will be most useful for stakeholders.

Depending on the duration of the intervention, you may be able to use interim findings from an implementation study to improve and refine the intervention. Although such refinements allow for midcourse corrections and improvements, they complicate the study of the intervention's impact—given that the intervention itself is changing over time.

Most small pilots will not be able to detect effects on cost because they do not have enough practices to detect such effects. Devoting resources to an impact study with a small number of practices is not a good investment.

Design and conduct a study of impacts. Driving questions for most stakeholders are: What are the intervention's impacts on health care cost; quality; and patient, family, clinician, and staff experience? These are critical questions for a study of impacts. Unfortunately, most studies of practice-level interventions in a small number of practices would be wise to not invest resources in answering them, due to the statistical challenges inherent in evaluating the impacts of such interventions. If your organization can support a large-scale test of this kind, or has

sufficient statistical power with fewer practices because their practice patterns are very similar, this section of the Guide provides some pointers. We begin by explaining how to assess whether you are transforming enough practices to be able to conduct a study of impacts.

Assess whether the sample size is adequate to detect effects that are plausible to generate and substantial enough to encourage adoption. If you are considering conducting a study of impacts, you should first calculate whether the sample is large enough to detect effects that are moderate enough in size to be plausible, but large enough that stakeholders would consider adopting the intervention if the effects were demonstrated. Your assessment of statistical power must account for clustering of patient outcomes within practices. In most cases, evaluations of primary care interventions require surprisingly large numbers of practices, regardless of how many patients are served, to be confident that plausible and adequate-sized effects on cost and utilization measures will be shown to be statistically significant (described in more detail in Appendix A). Your evaluation would likely need to include more than 50 intervention practices (unless the practice patterns are very similar) to be confident that observed differences in outcomes are true effects of the intervention.⁷

For most studies, power estimates will show that it will not be possible to detect the effects of an intervention in a small number of practices unless the effects are much larger than could plausibly be generated. (Exceptions are practices with similar practice patterns.) In these cases, we advise against evaluating and estimating program effects (impacts). Doing so is likely to lead to erroneous conclusions that the intervention did not work (if the analysis accurately accounts for clustering of patient outcomes within practices), when the evaluation may not have had a sufficient number of practices to differentiate between real program effects and natural variation in outcomes. In those cases, we recommend that the evaluation focus on conducting an implementation study.

Consider these pointers for your impact study. If you have sufficient statistical power to include an impact study, here are things to consider.

Comparing the intervention group to a comparison group that is similar before the intervention is critical. The evaluation can select the comparison group using a randomized or nonexperimental design. If possible, try to use a randomized design.

1. Have a method for estimating the outcomes patients would have experienced in the absence of the intervention. Merely looking at changes in trends over time is unlikely to correctly identify the effects of the intervention because trends and external factors unrelated to the intervention affect outcomes. Skeptics will find such studies dubious because changes over time in health care costs may have affected all practices. For example, if total costs for patients treated by a PCMH declined by 5 percent, but health care costs of all practices in that geographic

region declined by 5 percent over the same period, the evaluation should conclude that the PCMH is unlikely to have had a meaningful effect on costs.

You should, therefore, consider what would have happened to the way intervention practices delivered care and to patients' outcomes if the practice had not adopted the intervention—that is, the “counterfactual.” Comparing changes in outcomes between the intervention practices and a group of comparable practices helps to isolate the effect of the intervention from the effects of other factors. If you can, select the comparison group of practices using a randomized or experimental design. A randomized design will give you more confidence in your results and should be used whenever possible. Appendix B contains a few additional details on different approaches to selecting a comparison group, but we caution that the appendix does not cover the many considerations that go into selecting a comparison group.

2. Make sure comparison practices are as similar as possible to intervention practices before the intervention begins. If the intervention and comparison groups are similar before the intervention begins, you can be more confident that the intervention caused any subsequent differences in outcomes between the two groups. If available, your evaluation should select comparison practices with similar patient panels, including age, gender, and race; insurance source; chronic conditions; and prior expenditures and use of hospitalizations, ER visits, and skilled nursing facility stays. Ideally, practice-level variables such as practice size; whether the practice is independent or part of a larger system; the number, types, and roles of nonphysician staff; and urban/rural location should also be similar. To improve confidence further, if data are available, you should examine how similar outcomes were in both groups for several years before the intervention began to ensure patients in the two groups had a similar trajectory of costs. Moreover, if there are preexisting differences in cost trends, you can control for them. You can examine the comparability of the intervention and comparison practices along as many of these dimensions as possible even if you cannot use all of them to select the comparison group.

The Patient Centered Outcomes Research Institute also provides useful recommendations for study methodology. www.pcori.org/assets/2013/11/PCORI-Methodology-Report.pdf

3. **Use solid analytical methods to estimate program impacts.** If you have selected a valid comparison group and included enough practices, appropriate analytical methods will generate accurate estimates of program effects. These include using a difference-in-difference approach (which compares changes in outcomes before and after the intervention began for the intervention group to changes in outcomes over the same time period for the comparison group), controlling for patient- and practice-level variables (“risk adjustment”), and adjusting standard errors for clustering and multiple comparisons (see Appendix C).

4. **Conserve resources by using different samples to measure different outcomes.** Calculating statistical power for each outcome can help you decide which sample to use to collect data on various outcomes. It is generally costly to collect survey data. For most survey-based outcomes, evaluations typically need data from 20 to 100 patients per practice to be confident that they can detect a meaningful effect. Collecting survey data from more patients might increase the precision of estimates of a practice’s average outcome for its own patients, but it will only slightly improve the precision of the estimated effect for the intervention as a whole (that is, increase your ability to detect a small effect). It typically adds relatively little additional cost to analyze data from claims or electronic health records (EHRs) on all of the practices’ patients rather than just a sample, so we recommend analyzing claims- and EHR-based outcomes using data for as many patients as you can. On the other hand, some interventions can be expected to generate bigger effects for high-risk patients, so knowing how you will define “high-risk” and separately analyzing outcomes for patients who meet those criteria may improve your ability to detect specific effects.¹³

Synthesize findings from the implementation and impact analyses. Most evaluations generate a lot of information. If your evaluation includes both implementation and impact analyses, using both types of findings together will provide a considerably more sophisticated understanding about the effects of the model being tested than either alone. Studying the connections between findings from both—arrayed according to their appearance in the logic model—can help illuminate how a primary care intervention is working, suggest refinements to it, and, if it is successful, consider how to spread it to other practices.

Ideally, you will be able to integrate your implementation and impact work so that they will inform one another on a regular and systematic basis. This type of integrated approach can provide insights about practice operations, and barriers and facilitators to success. It can also help generate hypotheses to test with statistical models of impact, as well as explanations for differences in impacts across geographic areas or types of practices or patients. This information, in turn, can be used to improve the interventions being implemented and inform practices about the effectiveness of changes they are making. If you collect implementation and impact results at the same time, you can use them to validate findings and strengthen the evidence for the evaluation’s conclusions. Moreover, information from implementation and impact analyses is useful for understanding how to refine and spread successful interventions.

V. How Can I Use the Findings?

This Evaluation Guide describes some steps for planning and conducting an evaluation of a primary care intervention such as a PCMH model. The best evaluation design and approach for a particular intervention will depend on your goals and the available data and resources, as well as the way practices are selected to participate and the type of intervention they are testing.

A well-designed and well-conducted evaluation can provide evidence on what the intervention did or did not achieve. Specifically, it can describe (1) who the intervention served; (2) how it changed health care delivery; and (3) the effects on patient quality, cost, and experience outcomes as well as on clinician and staff experience. The evaluation also can identify how both implementation and impact results vary by practice setting and by patient subgroups, which has important implications for targeting interventions. Finally, the evaluation can use information about variations in implementation approaches across practices to identify best practices in staffing, roles of team members, and specific approaches to delivering services.

Evaluation results can provide answers to stakeholders' questions and help in making key decisions. Results can be compared with those of alternative interventions being considered. In addition, if the results are favorable, they can support plans to continue funding the intervention and to spread the model, perhaps with refinements, on a larger, more sustainable scale.

VI. Resource Collection for Evaluations of Primary Care Models

This resource collection contains resources and tools that evaluators can use to develop logic models; select outcomes; and design, conduct, and synthesize implementation and impact findings.

Logic Models

Innovation Network, Inc. Logic Model Workbook. Washington, DC: Innovation Network, Inc.; n.d. www.innonet.org/client_docs/File/logic_model_workbook.pdf.

Petersen D, Taylor EF, Peikes D. Logic Models: The Foundation to Implement, Study, and Refine Patient-Centered Medical Home Models. AHRQ Publication No.13-0029-EF. Rockville, MD: Agency for Healthcare Research and Quality; March 2013.

W.K. Kellogg Foundation. Logic Model Development Guide. Battle Creek, MI: W.K. Kellogg Foundation; December 2001:35-48. www.wkkf.org/knowledge-center/resources/2006/02/WK-Kellogg-Foundation-Logic-Model-Development-Guide.aspx.

Outcomes

Agency for Healthcare Research and Quality. Consumer Assessment of Healthcare Providers and Systems (CAHPS). Rockville, MD: Agency for Healthcare Research and Quality; 2012.

Rosenthal M, Abrams M, Bitton A, the Patient-Centered Medical Home Evaluators' Collaborative. Recommended core measures for evaluating the patient-centered medical home: cost, utilization, and clinical quality. Commonwealth Fund; May 2012(12). www.commonwealthfund.org/~media/Files/Publications/Data%20Brief/2012/1601_Rosenthal_recommended_core_measures_PCMH_v2.pdf.

Implementation Analysis

Alexander JA, Hearld LR. The science of quality improvement implementation: developing capacity to make a difference. *Med Care*. 2011 December;49(Suppl):S6-20.

Bitton A, Schwartz GR, Stewart EE, et al. Off the hamster wheel? Qualitative evaluation of a payment-linked patient-centered medical home (PCMH) pilot. *Milbank Q*. 2012 September;90(3):484-515. doi: 10.1111/j.1468-0009.2012.00672.x.

Bloom H, ed. Learning More from Social Experiments. New York: Russell Sage Foundation; 2006.

Crabtree B, Workgroup Collaborators. Evaluation of Patient Centered Medical Home Practice Transformation Initiatives. Washington: The Commonwealth Fund; November 2010.

Damschroder L, Aron D, Keith R, et al. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. *Implement Sci.* 2009;4(1):50.

Damschroder L, Peikes D, Petersen D. Using Implementation Research to Guide Adaptation, Implementation, and Dissemination of Patient-Centered Medical Home Models. AHRQ Publication No.13-0027-EF. Rockville, MD: Agency for Healthcare Research and Quality; March 2013.

Goldman RE, Borkan J. Anthropological Approaches: Uncovering Unexpected Insights About the Implementation and Outcomes of Patient-Centered Medical Home Models. Rockville, MD: Agency for Healthcare Research and Quality; March 2013. AHRQ Publication No.13-0022-EF.

Nutting PA, Crabtree BF, Stewart EE, et al. Effects of facilitation on practice outcomes in the National Demonstration Project model of the patient-centered medical home. *Ann Fam Med.* 2010;8(Suppl 1):s33-s44.

Plsek PE, Greenhalgh T. Complexity science: the challenge of complexity in health care. *BMJ* 2001;323:625-8.

Potworowski G, Green, LA. Cognitive Task Analysis: Methods to Improve Patient-Centered Medical Home Models by Understanding and Leveraging Its Knowledge Work. AHRQ Publication No.13-0023-EF. Rockville, MD: Agency for Healthcare Research and Quality; March 2013.

Rossi PH, Lipsey MW, Freeman HE. *Evaluation: A Systematic Approach.* 7th ed. Thousand Oaks, CA: Sage Publications; 2004.

Stange K, Nutting PA, Miller WL, et al. Defining and measuring the patient-centered medical home. *J Gen Int Med.* June 2010;25(6):601-12. www.ncbi.nlm.nih.gov/pmc/articles/PMC2869425/pdf/11606_2010_Article_1291.pdf.

Wholey JS, Hatry HP, and Newcomer KE, eds. *Handbook of Practical Program Evaluation.* San Francisco: Jossey-Bass/John Wiley & Sons; 2010.

Impact Analysis

Overview

Hickam, D, Totten A, Berg A, et al., eds. The PCORI Methodology Report. November 2013. www.pcori.org/assets/2013/11/PCORI-Methodology-Report.pdf.

Jaén CR, Ferrer RL, Miller WL, et al. Patient outcomes at 26 months in the patient-centered medical home National Demonstration Project. *Ann Fam Med.* 2010;8(Suppl 1):s57-s67.

Meyers D, Peikes D, Dale S, et al. Improving Evaluations of the Medical Home. Rockville, MD: Agency for Healthcare Research and Quality; September 2011. AHRQ Publication No. 11-0091. <http://pcmh.ahrq.gov/page/patient-centered-medical-home-decisionmaker-brief-improving-evaluations-medical-home>.

Orr, L. Social Experiments: Evaluating Public Programs with Experimental Methods. Thousand Oaks, CA: Sage Publications; 1999.

Shadish, WR, Cook, TD, Campbell, DT. Experimental and Quasi-Experimental Designs for General Causal Inference. Boston: Houghton Mifflin Company; 2001.

Stepped Wedge (Randomized Rollout) Designs

Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. Contemp Clin Trials. 2007;28:182-91. <http://faculty.washington.edu/peterg/Vaccine2006/articles/HusseyHughes.2007.pdf>.

Calculating Power When Data Are Clustered

Peikes D, Dale S, Lundquist E, et al. Building the Evidence Base for the Medical Home: What Sample and Sample Size Do Studies Need? Rockville, MD: Agency for Healthcare Research and Quality; October 2011. AHRQ Publication No. 11-0100-EF. White paper prepared by Mathematica Policy Research under contract no. HHSA290200900019I TO2. <http://pcmh.ahrq.gov/page/building-evidence-base-medical-home-what-sample-and-sample-sizes-do-studies-need>.

Adjusting for Multiple Comparisons

Schochet, PZ. An Approach for Addressing the Multiple Testing Problem in Social Policy Impact Evaluations. Eval Rev. 2009 December;33(6).

Orthogonal Design

Collins LM, Murphy S, Strecher V. The multiphase optimization strategy (MOST) and the sequential multiple assignment randomized trial (SMART). New methods for more potent ehealth interventions. Am J Prev Med. 2007;32:112-18.

Zurovac J, Peikes D, Zutshi A, Brown R. Efficient Orthogonal Designs: Testing the Comparative Effectiveness of Alternative Ways of Implementing Patient-Centered Medical Home Models. AHRQ Publication No.13-0024-EF. Rockville, MD: Agency for Healthcare Research and Quality; March 2013.

Propensity Score Matching

Peikes D, Moreno L, Orzol S. Propensity score matching: A note of caution for evaluators of social programs. Am Statistician 2008;62:222-31. http://econpapers.repec.org/article/besamstat/v_3a62_3ay_3a2008_3am_3a2008_3a222-231.htm.

Mixed Methods

Creswell JW, Klassen AC, Plano Clark VL, et al., for the Office of Behavioral and Social Sciences Research. Best Practices for Mixed Methods Research in the Health Sciences. Bethesda, MD: National Institutes of Health; August 2011.

Miller WL, Crabtree BF, Harrison MI, et al. Integrating mixed methods in health services and delivery system research. *Health Serv Res.* 2013 December;48(6 Part II).

Wisdom J, Creswell JW. Mixed Methods: Integrating Quantitative and Qualitative Data Collection and Analysis While Studying Patient-Centered Medical Home Models. AHRQ Publication No.13-0028-EF. Rockville, MD: Agency for Healthcare Research and Quality; March 2013.

Appendix A: Calculating Statistical Power with Clustering to Assess Potential for Detecting Meaningful Effects

This appendix describes how to calculate statistical power to determine whether to conduct an impact study, and if so, which outcomes you can examine. To do so, first consider effect size: What effect size would convince stakeholders to adopt the intervention? The answer to this question will determine the effect size you should aim to detect with confidence. Discussions about the effect sizes that stakeholders consider meaningful should occur when you describe the logic model/theory of change. Stakeholders focused on return on investment may need to see a reduction in costs large enough to more than offset any extra payments made to the practice to adopt an intervention. Others may need to see improvements in patient experience ratings. Next, assess the feasibility of generating that effect in the time frame you have to work with. Finally, calculate statistical power to figure out how many practices and patients you need to be confident that the evaluation will detect an effect of this size. If you already know the maximum number of practices you will be transforming, you can skip to the final step and see how likely it is that the evaluation will find an effect of a given size to be statistically significant. If the number of practices is too small, the evaluation will be unable to reliably determine whether the intervention generated an effect.

Calculating the minimum effect that the evaluation is likely to detect using different design approaches is particularly important with evaluations of primary care interventions such as the PCMH, because testing them often requires large samples. In practice-level interventions, patient outcomes are clustered within practices and this clustering reduces the effective sample size. Clustering arises because patients within a practice often receive care similar to that received by other patients served by the same practice, but different from the care received by patients in other practices—given differences in ways that clinicians practice medicine and other factors. This means that patients from a given practice cannot necessarily be considered statistically independent from one another, which lowers the effective sample size of patients. As a result, the number of practices (not the number of patients) in the intervention largely determines the size of the minimum effect the evaluation is likely to detect with high confidence. The amount of clustering in data varies for different samples of patients and practices and for different outcomes. When calculating the minimum detectable effect (and when analyzing outcomes), you must account for the clustered nature of the data.⁷

One strategy—measuring effects among members of a high-risk subgroup—might help improve power to detect effects, depending on the extent of clustering in the data for a given set of intervention practices and outcomes. Although models like the PCMH target all patients in a practice, studying sicker patients can increase the power to detect effects on continuous claims-based outcomes, such as cost and service use, for several reasons. First, among healthy patients, we expect relatively few hospitalizations and limited service use regardless of the intervention's effectiveness, leaving little opportunity to reduce health care use and cost. Among sicker patients, we expect more opportunities for reductions in cost and service use. As a result, evaluations can use smaller samples of those patients. Additionally, because service use and cost vary more widely among all patients than among sicker patients, it is often harder to distinguish an effect of the intervention from regular variation in these outcomes among all patients. Therefore, we suggest calculating minimum detectable effects for different outcomes and samples, to help you decide which outcome measures to track for which patients.

Appendix B: Using a Comparison Group to Account for What Would Have Happened Without the Intervention

This appendix briefly describes the complex issue of selecting a comparison group using a randomized or nonexperimental design. The goal is to identify a group of comparison practices that are as similar as possible to the intervention practices. To better understand your design options, it is often most efficient to consult with an experienced evaluator. You can also obtain background from a good textbook (such as Orr¹⁴ or Shadish, Cook, and Campbell¹⁵).

When Possible, Use a Randomized Design

The most rigorous and credible way to develop a counterfactual is to randomize practices interested in participating in the intervention to an intervention or control group.^b The control group will then provide a good proxy of what would have happened to intervention practices had they not adopted the model. However, many stakeholders believe they cannot conduct a randomized trial for ethical or fairness reasons. In such cases, a key question is: Are there more practices interested in transforming than resources to transform them? If the answer is yes, two pragmatic ways to randomize practices are available—both of which provide a strong randomized design to study the effects of a primary care intervention.

The first approach to selecting practices to participate is to conduct a lottery among all practices that volunteer. A lottery is a randomized controlled trial in which practices selected by lottery receive the intervention, and practices that are not selected serve as a control group.

Another approach is to allow all practices that volunteer to participate, but stagger the rollout of implementation across them. This is called a staggered randomized or stepped wedge design. The late starters serve as a control group—before they begin the intervention—for the early starters.¹⁶ The advantages of this design are (1) all interested practices have the opportunity to participate, and (2) operational support can be provided to small groups of practices at a time, reducing resource demands on the system. The disadvantage is that the late starters can only serve as a pure control group until they begin the intervention. For example, if they begin 1 year later than the early starters, your evaluation will have only 1 year of data to use to compare outcomes between the intervention and control groups—which might be too short a period to realize many potential improvements associated with primary care transformation.¹⁷ However, you can also use a staggered randomized design to examine outcomes at different stages, such as comparing practices with 2 years of experience with the intervention and practices with only 1 year of experience.

If your evaluation uses a randomized design by lottery or by staggered rollout, it is critical to select practices at random. Picking a practice for the intervention group because it seemed to have the strongest physician commitment, or because it had better or worse patient outcomes, makes it difficult

^b A control group is a comparison group that is selected randomly from the set of potential practices eligible to implement the study.

to disentangle the effects of the intervention from those of the practice's existing performance or motivation. Similarly, in a staggered randomized design, be sure to randomize practices into rollout periods, avoiding the urge to start with practices that are more sophisticated or more eager to begin the intervention.

If stakeholders want to introduce the intervention in all practices, another option would be to analyze the effectiveness of different approaches to implementing the components of the intervention within the practices. At the outset of the study, each practice could be randomized to receive a combination of different approaches to implementing the intervention. For example, the practices could be randomly assigned to use either a social worker or nurse to coordinate care, and randomly assigned to follow up with patients within 2 days of a hospital discharge, either in person or by telephone. This approach, called orthogonal design, enables every practice to test at least some of the components (that is, no practice would be a pure control), while generating important operational lessons about the best ways to deliver the different components.^{18, 19}

When Randomized Designs Are Not Feasible, Use a Strong Comparison Group Design

Sometimes randomized designs are not feasible. In this case, it is critical to determine how the participating practices chose (or were chosen) to participate in the intervention and mimic those factors to the extent possible when selecting a nonexperimental comparison group. The factors driving participation include formal and informal selection criteria by the organization and decisions made by practices. For example, if the organization selects all practices in a particular city to test the intervention, the comparison group should contain practices in a city with a comparable market and patient mix. If only practices that had certain health IT in place were chosen, practices with similar health IT, as well as size, patient mix, and outcomes—before the intervention—should be selected for the comparison group. Ideally, the group of comparison practices should have the same characteristics as the intervention practices. Two popular options for selecting a comparison group are regression discontinuity (RD) designs, and propensity score matching (PSM) designs. However, both PSM and RD designs may not have sufficient power for interventions with a small number of practices.

Appendix C: Using Solid Analytical Methods to Estimate Program Impacts

This appendix briefly describes some analytical methods that can help you estimate program impacts: using a difference-in-difference approach, conducting multivariate analysis that controls for patient- and practice-level variables, and adjusting standard errors for clustering and multiple comparisons. These are complex topics, and these summaries are intended as an introduction to the general concepts underlying them. Again, it's often most efficient to consult with an experienced evaluator to explore the analytic methods that are most appropriate for your evaluation questions and design.

Estimate Effects Using a Difference-in-Difference Approach

We recommend that evaluations calculate difference-in-difference estimates of program impacts by subtracting the difference in a given outcome between the intervention and comparison groups before the intervention began from the difference in that same outcome during the intervention. This approach assumes that any differences between intervention and comparison practices in both levels and trends in outcomes before the intervention would have persisted after the intervention if the intervention had not occurred. Thus, for example, in the case of improved access through email described in Figure 2, the impact of the intervention is the change in access over time for patients in intervention practices after netting out the change in access over time experienced by patients in comparison practices.

Control for Differences in Patient and Practice Characteristics

In your analyses, you should use multivariate regressions to adjust estimates for differences in important patient- and practice-level variables (described previously) or control for practice fixed effects (that is, practice-level characteristics that do not change over time) because pre-existing differences in intervention and comparison practices can affect outcomes.

Adjust Standard Errors for Clustering and Multiple Comparisons

You must account for clustering when determining the statistical significance of the estimates of program effects. If clustering is ignored, a test of statistical significance might show a difference in outcomes between intervention and comparison practices to be statistically significant when it is not. In other words, ignoring the clustered nature of the data can lead to a false positive—finding an effect that does not exist.⁷ Similarly, if you test the effect of the intervention on numerous outcomes, you risk finding some effects just by chance. Therefore, you should assess whether you are finding more statistically significant findings than would be expected by chance for the number of tests you are conducting. There are also more formal ways to adjust standard errors for multiple comparisons.²⁰

Endnotes

- ¹ McNellis RJ, Genevro JL, Meyers DS. Lessons learned from the study of primary care transformation. *Ann Fam Med*. 2013 May-June;11(Suppl 1):S1-5. doi: 10.1370/afm.1548. PubMed PMID: 23690378; PubMed Central PMCID: PMC3707240.
- ² McMullen CK, Schneider J, Firemark A, et al. Cultivating engaged leadership through a learning collaborative: lessons from primary care renewal in Oregon safety net clinics. *Ann Fam Med*. 2013 May-June;11(Suppl 1):S34-40. doi: 10.1370/afm.1489. PubMed PMID: 23690384; PubMed Central PMCID: PMC3707245.
- ³ Solberg LI, Crain AL, Tillema J, et al. Medical home transformation: a gradual process and a continuum of attainment. *Ann Fam Med*. 2013 May-June;11(Suppl 1):S108-14. doi: 10.1370/afm.1478. PubMed PMID: 23690379; PubMed Central PMCID: PMC3707254.
- ⁴ Nutting PA, Crabtree BF, Miller WL, et al. Transforming physician practices to patient-centered medical homes: lessons from the national demonstration project. *Health Aff (Millwood)*. 2011 March;30(3):439-45. doi: 10.1377/hlthaff.2010.0159. PubMed PMID: 21383361; PubMed Central PMCID: PMC3140061.
- ⁵ Reid RJ, Johnson EA, Hsu C, et al. Spreading a medical home redesign: effects on emergency department use and hospital admissions. *Ann Fam Med*. 2013 May-June;11(Suppl 1):S19-26. doi: 10.1370/afm.1476. PubMed PMID: 23690382; PubMed Central PMCID: PMC3707243.
- ⁶ Reid RJ, Fishman PA, Yu O, et al. Patient-centered medical home demonstration: a prospective, quasi-experimental, before and after evaluation. *Am J Manag Care*. 2009 September;15(9):e71-87. PubMed PMID: 19728768.
- ⁷ Peikes D, Dale S, Lundquist E, et al. 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; October 2011. White paper prepared by Mathematica Policy Research under contract no. HHS290200900019I TO2. <http://pcmh.ahrq.gov/page/building-evidence-base-medical-home-what-sample-and-sample-size-do-studies-need>.
- ⁸ Nutting PA, Crabtree BF, Stewart EE, et al. Effects of facilitation on practice outcomes in the National Demonstration Project model of the patient-centered medical home. *Ann Fam Med*. 2010;8(Suppl 1):s33-s44.
- ⁹ Petersen D, Taylor EF, Peikes D. Logic Models: The Foundation to Implement, Study, and Refine Patient-Centered Medical Home Models. AHRQ Publication No.13-0029-EF. Rockville, MD: Agency for Healthcare Research and Quality; March 2013.
- ¹⁰ Wisdom J, Creswell JW. Mixed Methods: Integrating Quantitative and Qualitative Data Collection and Analysis While Studying Patient-Centered Medical Home Models. AHRQ Publication No.13-0028-EF. Rockville, MD: Agency for Healthcare Research and Quality; March 2013.

- ¹¹ Creswell JW, Klassen AC, Plano Clark VL, et al., for the Office of Behavioral and Social Sciences Research. Best Practices for Mixed Methods Research in the Health Sciences. Bethesda, MD: National Institutes of Health; August 2011.
- ¹² Miller WL, Crabtree BF, Harrison MI, et al. Integrating mixed methods in health services and delivery system research. *Health Serv Res.* 2013 December;48(6 Part II).
- ¹³ Meyers D, Peikes D, Dale S, et al. Improving Evaluations of the Medical Home. AHRQ Publication No. 11-0091. Rockville, MD: Agency for Healthcare Research and Quality; September 2011. <http://pcmh.ahrq.gov/page/patient-centered-medical-home-decisionmaker-brief-improving-evaluations-medical-home>.
- ¹⁴ Orr, L. Social Experiments: Evaluating Public Programs with Experimental Methods. Thousand Oaks, CA: Sage Publications; 1999.
- ¹⁵ Shadish, WR, Cook, TD, Campbell, DT. Experimental and Quasi-Experimental Designs for General Causal Inference. Boston: Houghton Mifflin Company; 2001.
- ¹⁶ Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemp Clin Trials.* 2007;28:182-91. <http://faculty.washington.edu/peterg/Vaccine2006/articles/HusseyHughes.2007.pdf>.
- ¹⁷ Jaén CR, Ferrer RL, Miller WL, et al. Patient outcomes at 26 months in the patient-centered medical home National Demonstration Project. *Ann Fam Med.* 2010;8(Suppl 1):s57-s67.
- ¹⁸ Zurovac J, Peikes D, Zutshi A, Brown R. Efficient Orthogonal Designs: Testing the Comparative Effectiveness of Alternative Ways of Implementing Patient-Centered Medical Home Models. AHRQ Publication No.13-0024-EF. Rockville, MD: Agency for Healthcare Research and Quality; March 2013.
- ¹⁹ Collins LM, Murphy S, Strecher V. The multiphase optimization strategy (MOST) and the sequential multiple assignment randomized trial (SMART). New methods for more potent ehealth interventions. *Am J Prev Med.* 2007;32:112-18.
- ²⁰ Schochet, PZ. An Approach for Addressing the Multiple Testing Problem in Social Policy Impact Evaluations. *Eval Rev.* 2009 December;33(6).

