



# Building the Evidence Base for the Medical Home: What Sample and Sample Size Do Studies Need?




WHITE PAPER




Agency for Healthcare Research and Quality

Advancing Excellence in Health Care • [www.ahrq.gov](http://www.ahrq.gov)

Prevention/Care Management



# Building the Evidence Base for the Medical Home: What Sample and Sample Size Do Studies Need?



WHITE PAPER

**Prepared for:**

U.S. Department of Health and Human Services  
Agency for Healthcare Research and Quality  
540 Gaither Road  
Rockville, MD 20850  
[www.ahrq.gov](http://www.ahrq.gov)

**Contract No. HHS290200900019I TO2**

**Prepared by:**

Mathematica Policy Research  
Princeton, NJ

Deborah Peikes, Mathematica Policy Research  
Stacy Dale, Mathematica Policy Research  
Eric Lundquist, Mathematica Policy Research

Janice Genevro, Agency for Healthcare Research and Quality  
David Meyers, Agency for Healthcare Research and Quality

AHRQ Publication No. 11-0100-EF  
October 2011



This document is in the public domain and may be used and reprinted without permission except those copyrighted materials that are clearly noted in the document. Further reproduction of those copyrighted materials is prohibited without the specific permission of copyright holders.

**Suggested Citation:**

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. HHS290200900019I TO2). AHRQ Publication No. 11-0100-EF. Rockville, MD: Agency for Healthcare Research and Quality. October 2011.

None of the investigators has any affiliation or financial involvement that conflicts with the material presented in this report.



## Acknowledgments

We would like to thank a number of people for helping with this paper. Asaf Bitton (Brigham and Women's Hospital/Harvard Medical School) reviewed the paper, raised several important questions, and suggested avenues to explore. Members of the Federal Patient-Centered Medical Home Collaborative and the Commonwealth Fund's Patient-Centered Medical Home Evaluators' Collaborative provided helpful comments and questions in response to presentations of this work. Michael Parchman at the Agency for Healthcare Research and Quality; and Randy Brown, Hanley Chiang, John Deke, and Barbara Carlson at Mathematica Policy Research, provided helpful comments and guidance during the development of this paper. Walt Brower edited the paper, and Jennifer Baskwell produced it.

## Abstract

**Background.** Well-designed and -implemented studies are critical to determining whether the patient-centered medical home (PCMH) model of primary care is effective in raising quality, reducing costs, and improving the experience of care and, if not, how the model should be altered.

**Purpose.** This paper provides information that can be used to improve the quality of the evidence about whether the medical home works. Practical suggestions are offered for generating credible evidence on the effects of the medical home model in the face of several challenges, which include the following:

- The PCMH is a practice-level intervention.
- Evaluations of the medical home typically include a small number of practices.
- Health care costs and service use vary substantially.

The specific goals of the paper are:

- To raise awareness about the need to account for clustering inherent in practice-based interventions.
- To provide information about some of the key inputs used to determine what effect sizes a given study can expect to detect.
- To identify the approximate number of patients and practices required to detect policy-relevant yet achievable effects.
- To show how varying the outcomes and types of patients included in analyses can increase the ability of a study to detect true effects on outcomes of interest.

**Suggestions for Generating Credible Evidence.** Designing evaluations of the PCMH that will produce credible evidence requires a combination of approaches.

- So that estimates of the effectiveness of interventions that alter the entire practice such as the PCMH are not inflated, statistical corrections must be made for the degree to which patients in a practice tend to be more similar to each other than to patients in other practices (clustering). The amount of clustering varies for different practices, markets, and outcomes, so researchers should ensure that their calculations reflect the data they will be analyzing.
- To ensure adequate statistical power to detect effects, studies should include as many practices as possible. Many ongoing studies of the PCMH are small (including 10 to 20 intervention practices) and will likely not have enough statistical power to detect effects on costs and hospitalizations for all patients a practice serves.
- Including more patients per practice contributes little to power. Therefore, for any given number of patients, it is better to have many practices with few patients per practice than few practices with many patients in each. For example, a study with 100 practices and 20 patients per practice has much greater power than a study of 20 practices with 100 patients each.
- To increase the likelihood of detecting the effects of PCMH interventions and to use evaluation resources efficiently, researchers can (and should) measure different outcomes in different subsamples of patients. Costs and service use should be measured among those most likely to incur high health care costs, such as chronically ill patients. To illustrate why this is helpful, using some assumptions

from published literature, if an evaluation has 20 intervention and 20 control practices and measures effects on all patients, the intervention would have to generate a cost reduction of 45 percent on all patients for a study to be likely to detect it. In contrast, if effects are measured among only the chronically ill, a study could detect effects of a 21 percent reduction in costs.

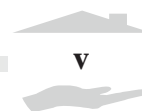
- Quality-of-care and satisfaction outcomes can be assessed in all patients, even in smaller studies. Evaluating these outcomes is easier because there is less variation in them than in cost and service use. In addition, a practice-level intervention should be expected to improve quality and experience for all patients.
- Although including more comparison practices for a given number of intervention practices makes it slightly easier for an evaluation to detect effects of a given size without substantially increasing costs, finding suitable comparison practices is often difficult. The modest improvement will likely not be sufficient to detect plausible effects on cost and service use for all patients, but may be helpful for studies that measure these outcomes for the chronically ill.
- There may be opportunities to undertake meta-analyses that combine the results of small studies or actually combine the data from several small studies to increase the power to detect effects. This will require collaboration among researchers and the utilization of common outcome metrics.

**Conclusions.** Decisionmakers contemplating whether to invest in implementation of the PCMH must consider two seemingly paradoxical consequences of less-than-optimal research designs.

First, decisions should not be based on analyses that show significant results but are not adjusted for clustering, as such findings are likely to be “false positives.” Analyses that do not account for clustering can incorrectly indicate that the interventions resulted in statistically significant (replicable) changes in outcomes when in reality they did not.

Second, the lack of significant findings from studies with too few practices does not necessarily indicate that the PCMH model does not work. These underpowered studies are unable to demonstrate statistical significance even when real effects are present. Findings from these studies may thus be “false negatives.” Decisionmakers should avoid drawing conclusions about effectiveness, especially conclusions about lack of effectiveness, from studies with few practices.

It is imperative that evaluators address these issues to design studies that can produce credible evidence regarding the PCMH and its effects on quality, cost, and experience of care.





# Contents

Chapter 1. Background.....	1
Chapter 2. The Challenges of Providing Evidence on the Medical Home.....	3
Chapter 3. What Size Effects Are We Looking For? .....	7
Chapter 4. What Size Effects Are Plausible for a Study to Detect?.....	9
Chapter 5. Inputs for Calculating the MDE from the Literature.....	11
Chapter 6. How Many Patients and Practices, and Which Patients, Should Be Included in the Study Sample? .....	13
How Many Patients Should Be Included per Practice? .....	13
How Many Intervention Practices Should Be Included? .....	14
Which Patients Should Be Included for Each Outcome?.....	17
Can it Help to “Transform” the Outcome Variable to Reduce Variation? .....	20
Can Adding More Comparison Practices Improve MDEs? .....	21
Can a Study Improve Power by Accounting for Clustering at the Clinician (or Team) Level, Rather Than the Practice Level? .....	21
Chapter 7. Summary and Conclusions .....	23
Implications .....	24
References .....	26
Appendix A AHRQ Definition of the Medical Home.....	28
Appendix B Calculating Minimum Detectable Effects and Effective Sample Sizes.....	30
Appendix C Explanation of Figure 1 on False Positive Rates When Clustering Is Ignored.....	35
Appendix D Sample Effect Sizes Found in Targeted Literature Review .....	36
Appendix E Inputs from the Literature for Calculating MDEs .....	37
Appendix F Sample Code to Calculate the ICC .....	43
Tables	
Table 1. Average coefficient of variation (CV) for continuous variables from studies in our literature review....	11
Table 2. Average ICC reported in studies in our literature review.....	12
Table 3. Minimum detectable effects for all patients and chronically ill patients, by number of intervention practices.....	18
Table D.1. Examples of selected effects found in a targeted literature review .....	36
Table E.1. Coefficients of variation (CVs) reported in the literature.....	39
Table E.2. Definitions of chronically ill used in selected studies .....	41
Table E.3. Intraclass correlation coefficients (ICCs) reported in the literature.....	42



Figures

Figure 1. False Positive Rates When Ignoring the Effects of Clustering (Assuming the Intervention Has No Effect) .....5

Figure 2. MDEs (%) for Cost or Hospitalization Outcomes Among All Patients— Varying the Number of Patients per Practice.....14

Figure 3. MDEs (%) for Cost or Hospitalization Outcomes Among All Patients— Varying the Number of Practices: Small Studies.....14

Figure 4. MDEs (%) for Cost or Hospitalization Outcomes Among All Patients— Varying the Number of Practices: Large Studies.....15

Figure 5. MDEs (%) for Cost or Hospitalization Outcomes Among All Patients— Varying the Number of Practices and the ICC.....16

Figure 6. MDEs (%) for Cost or Hospitalization Outcomes Among All Patients— Varying the Number of Practices and the CV .....16

Figure 7. MDEs (%) for Cost or Hospitalization Outcomes Among Chronically Ill Patients—Varying the Number of Practices .....17

Figure 8. MDEs (%) for Quality-of-Care or Satisfaction Outcomes Among All Patients— Varying the Number of Practices .....19

Figure 9. MDEs (%) for Quality-of-Care or Satisfaction Outcomes Among All Patients— Varying the Number of Patients per Practice.....20

Figure 10. MDEs (%) for Cost or Hospitalization Outcomes Among All Patients— Varying the Number of Comparison Practices .....22

Figure 11. MDEs (%) for Cost or Hospitalization Outcomes Among Chronically Ill Patients—Varying the Number of Comparison Practices .....22

# Chapter 1. Background

Insurers, policymakers, providers, and patients are waiting eagerly to see whether recent efforts to transform primary care delivery will pay off in improved quality, lower costs, and better patient and provider experience. The model being tested, most commonly called the patient-centered medical home (PCMH),<sup>1</sup> aims to strengthen the primary care foundation of the health care system by reorganizing the way primary care practices provide care (American Academy of Family Physicians et al., 2007; Rittenhouse, Shortell, and Fisher, 2009). According to the Agency for Healthcare Research and Quality (AHRQ), the medical home is supported by health information technology, workforce development, and payment reform and rests on five pillars: a patient-centered orientation, comprehensive team-based care, coordinated care, superb access to care, and a systems-based approach to quality and safety.<sup>2</sup> (Appendix A describes the PCMH model in more detail.)

Well-designed and -implemented studies are critical to determining whether the current PCMH model is effective in a variety of domains and, if it is not, how it should be altered. Payers are relying on evaluations to determine whether the PCMH improves outcomes enough to justify costs before attempting to replicate it more broadly among the providers who deliver care to their members. Results from evaluations will also guide practices that are contemplating whether to adopt the model, and patients when selecting their primary care practice.

This paper aims to improve the quality of the evidence about whether the PCMH works. We have four specific goals: (1) to raise awareness about the need to account for clustering inherent in PCMH effectiveness research, (2) to provide information about some of the key inputs into determining what effect sizes a given study can expect to detect, (3) to identify the approximate number of patients and practices required to detect policy-relevant yet achievable effects, and (4) to show how varying the outcomes and types of patients included in analyses can improve a study's ability to detect true effects. The paper focuses on quantitative evaluations designed to measure effectiveness, rather than the smaller pilots that are useful to test implementation of the intervention.

In Chapter 2, we describe the challenges to evaluating the medical home and the importance of accounting for clustering. In Chapter 3, we describe what types of effects we can expect well-designed interventions to generate by presenting examples of findings in the literature. Chapter 4 turns to what effect sizes an evaluation can likely detect. To be of value, evaluations must be powered to be able to detect effects that are both plausible and meaningful to policymakers. We discuss the minimum detectable effect (MDE) and the factors that can increase or decrease it. Chapter 5 presents estimates for these factors based on a review of 18 published and unpublished studies. These estimates may be helpful when researchers do not have access to the data needed to calculate their own MDEs. Chapter 6 presents likely MDEs for key outcomes based on reasonable assumptions about the number of patients per practice, the number of practices, and what types of patients are included in the sample when measuring different outcomes. Chapter 7 concludes by discussing the implications for study design and interpretation of findings.

---

<sup>1</sup> This model is also referred to as the health care home, advanced primary care, or patient-aligned care.

<sup>2</sup> Adapted from AHRQ's medical home definition (AHRQ, 2010).



## Chapter 2. The Challenges of Providing Evidence on the Medical Home

This paper makes practical suggestions for how to generate credible evidence of the effects of the medical home model in the face of several challenges. The first is the *small numbers of practices* being used in most of the existing and planned studies (Bitton, Martin, and Landon, 2010). From an operational perspective, including more practices is expensive because of the financial and organizational costs of facilitating individual practice transformation and making payments to each practice. However, from the perspective of evaluating the intervention, including only a small number makes it difficult to determine whether the intervention worked. Second, because *health care costs and service use vary substantially*, it is difficult to detect whether a difference in these patient outcomes between practices that become a medical home and similar practices that do not can be attributed to the medical home model or simply to random fluctuations (“noise”) in the data. Third, although the medical home is a practice-level intervention (that is, it alters the way the entire practice operates), it *can likely affect only the costs of a small fraction of the practices’ patients* in the 1- to 3-year period most studies examine (that is, patients most likely to incur high health care costs, such as the chronically ill). High-risk patients present more opportunities to improve health care and patient self-care. Conversely, there are fewer opportunities to improve health care for relatively healthy patients. A medical home model cannot substantially reduce hospitalizations for patients who are only infrequently hospitalized. For example, in a given year, under 30 percent of Medicare patients with preexisting chronic illnesses are hospitalized, and typically well under 10 percent of all patients in a practice (including both healthy and chronically ill patients) are hospitalized.

Perhaps the largest challenge stems from the fact that the medical home is generally a practice-level intervention, which means that statistical adjustments are needed because *patient outcomes are correlated or “clustered” within a practice*. Clustering arises when interventions are delivered to groups of people (in the medical home context, the clusters are practices or clinics), and outcomes are correlated (not independent) for individuals within the same practice or clinic. This lack of independence reduces the amount of unique information and, therefore, the effective sample size. Clustering is likely to occur when there is variation across practices (or clinics or clinicians) in the way patients are treated. For example, because of variation in practice patterns, the rate of foot exams for patients with diabetes may be higher in some practices than in others.<sup>3</sup> The degree of clustering can vary for different types of patients or practices, in different markets, or for different types of outcomes. An extreme case of clustering would arise if each practice in a study sample had a different approach to providing care and delivered that same type of care consistently for all patients within the practice. When there is a high degree of clustering due to providers having a strong influence on patient outcomes, study estimates will be more likely to fluctuate based on which specific providers are included in the intervention and comparison groups. By chance, better-performing (or worse-performing) providers might be assigned to the intervention group; to avoid erroneously concluding that an observed difference in outcomes between the intervention and comparison groups is an effect of the intervention, the sample needs to include more practices.

---

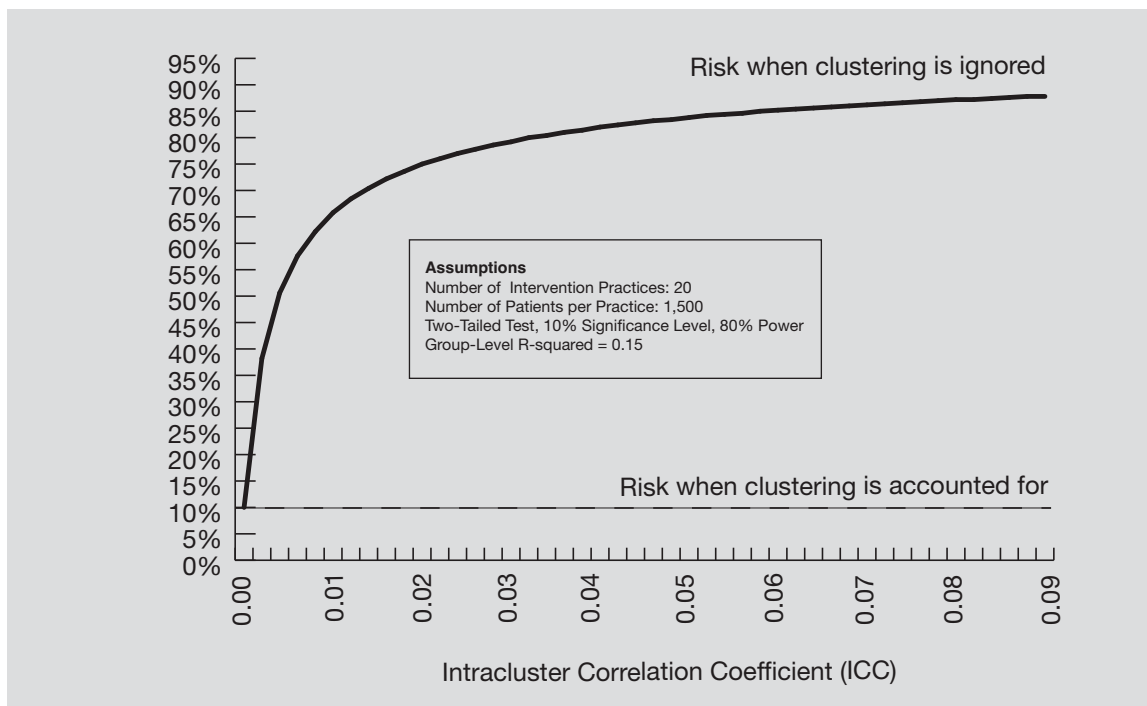
<sup>3</sup> Clustering can also occur if the types of patients in a practice vary or if patient characteristics cannot be adequately controlled in regressions.

The need to account for clustered designs has been well documented in statistical textbooks for many years and is standard practice in the education field and in the work of practice-based research networks (Murray, 1998; Bloom, 2005). It is also included in the guidelines used by leading medical publications (such as the *Journal of the American Medical Association* and the *British Medical Journal*) to ensure that studies of practice-level interventions generate externally valid results (Schochet, 2008; Bloom, 2005; Campbell, Elbourne, and Altman, 2004). However, many health researchers are not aware of the need to adjust for clustering, because clinical interventions (such as disease-specific treatments) are delivered at the patient level, and patients are considered independent units of analysis. In contrast, the medical home intervention affects the whole primary care practice, and individual patients are affected only through their relationship with the practice.

Failing to account for clustering leads to problems at both the design and the analysis phases of a study. Ignoring clustering when designing a study will lead researchers to believe the study has better ability to detect effects than it really does. This occurs because clustering reduces the effective sample size of a study. Study designs should take clustering into account when calculating the sample size needed to detect plausible effects. Doing so will typically increase the sample needed. The required sample size will be driven largely by the amount of clustering (the similarity of outcomes within and across practices) and the number of practices (Campbell, Mollison, and Grimshaw, 2001). If there is a high degree of clustering, the sample size used when calculating the MDEs will be driven largely by the number of practices. For example, if a study includes a total of 20,000 patients, but the patients are clustered within 20 practices, then the effective sample size is only 1,820 (assuming patient outcomes are moderately clustered within practices). (See Appendix B, Chapter 3, for details on this calculation as well as on methods of calculating effective sample sizes more generally.)

During the analysis phase, ignoring clustering in studies with fewer than 100 practices may lead researchers to conclude inaccurately that effects on costs or service use outcomes are statistically significant when they are not. If clustering is not accounted for, the chance of a false positive (that is, concluding that an intervention works when it does not) can be very large, often greater than 60 percent (see Figure 1, which assumes there are 20 intervention practices and 1,500 patients per practice, and the related discussion in Appendix C). Most studies are designed to accept a 5 or 10 percent chance of a false positive; most decisionmakers would view rates of 60 percent or more as too high.

**Figure 1. False Positive Rates When Ignoring the Effects of Clustering  
(Assuming the Intervention Has No Effect)**



Another important advantage of accounting for clustering is that it allows the research to gauge the potential generalizability of the study findings: it enables assessment of the likelihood that the findings are not limited to the set of practices that implemented the intervention. Intuitively, outcomes that were not adjusted for clustering would have the same MDE if based on a study with 2,000 patients in 1 practice (and a comparable control group) as they would from the same number of patients spread across 10 practices (with 10 control practices). However, a study done with 10 intervention and 10 control practices clearly provides more confidence than a single-site study that the findings are not due to the peculiarities of the intervention group practices.



## Chapter 3. What Size Effects Are We Looking For?

When designing a quantitative study of effectiveness, researchers often first consider the size of the effects the intervention could plausibly achieve, and whether such an expected effect size would be policy relevant to stakeholders. The size needed will likely vary based on a wide variety of factors, such as the cost of providing the intervention and the expected return relative to other possible interventions. In this paper, we focus on what sizes are plausible. Because the concept of a medical home is relatively new, to estimate plausible effect sizes, we examined evaluations of the medical home and selected health care interventions intended to improve quality of care and reduce health care costs, such as disease management. This list is not exhaustive, but we found that most of these evaluations were done among Medicare beneficiaries with chronic conditions (Appendix Table D.1). Although some results were unfavorable (such as increased cost), favorable results of 6 to 12 percent reductions for costs (without the intervention costs) and 6 to 44 percent reductions for hospitalizations were found (Gilfillan, Tomcavage, Rosenthal, et al., 2010; Counsell, Callahan, Clark, et al., 2007; Leff, Reider, Frick, et al., 2009; Boyd, Reider, Frey, et al., 2010; McCall, Cromwell, and Urato, 2010; Peikes, Peterson, Schore, et al., 2011). The one medical home study that treated all patients in a clinic reported that it reduced costs by 2 percent overall and hospitalizations by 6 percent overall (Reid, Coleman, and Johnson, 2010; Reid, Fishman, and Yu, 2009).<sup>4</sup>

Based on the studies in this targeted review, we assume that a successful program could hope to reduce costs or hospitalizations, on average, by 15 percent for chronically ill patients and 5 percent for all patients. These optimistic yet plausible effects serve as a rough guide; the nature and intensity of the intervention as well as the patient mix will shape the actual effect sizes a particular intervention can be expected to generate. Effects of 20 percent for patient experience and quality-of-care measures (for example, increasing the proportion who rate their practice as very good or excellent, or have had a flu shot, from 70 to 84 percent) regardless of whether the sample includes all patients or just the chronically ill appear to be reasonable for programs to achieve (Counsell, Callahan, Clark, et al., 2007; Boulton, Reider, Frey, et al., 2008).

---

<sup>4</sup> The lower effects for all patients make sense intuitively. Most of the effects for outcomes like cost, hospital bed days, and hospitalizations will be concentrated in, say, 25 percent of the patients in a practice because they are the ones at risk for those outcomes. Including other patients who benefited less or not at all will undoubtedly dilute the effect. Thus estimates for all patients should be lower or smaller than estimates for high-needs patients when considering cost and utilization.





## Chapter 4. What Size Effects Are Plausible for a Study to Detect?

After assessing what size effects might be feasible to generate as the result of the planned intervention, study designers should make sure they will have a large enough sample to ensure that their evaluation can detect impacts of that size. Also, when interpreting studies that show no effects, it is important to understand whether the study's sample was large enough to detect plausible effects. If a study's sample size is too small, statistical tests will be unlikely to find that an intervention worked—even if it was effective.

An MDE is defined as the minimum true intervention effect that can be detected by a particular statistical test with a given level of statistical significance at a given power level. In general terms, it is the smallest real program impact for which the study is quite likely to observe a statistically significant intervention-control group difference in the sample. For any study, a smaller MDE is more desirable.

Although many factors can affect the MDE of a study, we focus here only on four (Appendix B provides the equation and further explanation). In the design of practice-level evaluations, such as evaluations of the medical home, each of the following factors will lead to lower MDEs:

- Less variation in the outcome measure.
- Less clustering of patient outcomes within practices.
- More intervention and comparison practices in the sample, with equal numbers in each group.
- More patients per practice in the sample.<sup>5</sup>

---

<sup>5</sup> However, as we show later, increasing the number of study practices will improve the MDEs far more than adding patients per practice, even if the same total number of patients is included. In other words, if a study pays per patient and can afford to include 20,000 patients, it is better to have 20 practices of 1,000 patients each than 5 practices of 4,000 patients each.



## Chapter 5. Inputs for Calculating the MDE from the Literature

Because many evaluators do not have access to data in advance to calculate MDEs, study designers need to make some assumptions about the variability of their outcome measure (referred to as the “coefficient of variation,” or CV, defined as the standard deviation divided by the mean) and the degree of clustering (referred to as the “intracluster correlation coefficient,” or “intraclass correlation coefficient” (ICC), or the rate of homogeneity, which is the percentage of total variation accounted for between, as opposed to within, clusters) for their particular target practices and patients. To help inform these assumptions, we compiled CVs and ICCs from more than 18 published and unpublished studies (see Tables 1 and 2; Appendix E provides further details, and Appendix F provides code to calculate ICCs).

Although only a limited number of studies were available, and the CVs and ICCs varied considerably from study to study, we found some general patterns:

- Variation (the CV) for health care cost and service use is very large when the study population includes all patients (that is, both healthy patients and the chronically ill).
- The CV for health care cost and service use is considerably smaller when the study population is restricted to the chronically ill than when all patients are included.
- The CV for some measures, like hospital days, is so large that it is unlikely that most studies will be able to reliably detect effects for these outcomes.
- Clustering (the ICC) for health care cost and service use is relatively low, but still important to account for.
- The ICC for satisfaction, access to care, and quality process measures is relatively high.
- There was little information in the literature showing ICCs for different types of patients, providers, or markets.

**Table 1. Average coefficient of variation (CV) for continuous variables from studies in our literature review**

Population	Costs	Hospitalizations	Emergency Room Visits	Hospital Days
All, Privately Insured	3.73 (n=3)	4.7 (n=2)	2.83 (n=2)	NA
Chronically Ill, Privately Insured	2.50 (n=2)	3.18 (n=2)	2.21 (n=2)	5.80 (n=1)
All, Medicare	2.46 (n=3)	3.00 (n=1)	2.00 (n=1)	5.78 (n=1)
Chronically Ill, Medicare	1.64 (n=3)	1.99 (n=4)	2.75 (n=2)	3.07 (n=3)

Source: Estimates compiled by authors. Appendix E provides more details.

Note: n denotes the number of studies that were included in the average for each cell.

NA = not available.

**Table 2. Average ICC reported in studies in our literature review**

<b>Costs</b>	<b>Hospitalizations</b>	<b>Emergency Room Visits</b>	<b>Overall Satisfaction</b>	<b>Access to Care</b>	<b>Process Measures</b>
0.026 (n=2)	0.023 (n=3)	0.022 (n=4)	0.019 (n=3)	0.098 (n=3)	0.104 (n=3)

Source: Estimates compiled by authors. Appendix E provides more details.

Note: n denotes the number of studies that were included in the average for each cell.

Having described the importance of adjusting for clustering when deciding the appropriate sample size, the main factors influencing MDEs, and some estimates from the literature for the ICCs and CVs needed to calculate MDEs, we turn to calculating MDEs for a clustered design.

## Chapter 6. How Many Patients and Practices, and Which Patients, Should Be Included in the Study Sample?

People who commission or design evaluations have the opportunity to decide how many practices and patients per practice to include, as well as which subgroups of patients to analyze for specific types of outcomes, taking into account cost and feasibility. We provide more information for these decisions by graphing MDEs using estimated values of the ICC and CV (based on the literature). We also indicate on the graph the impacts a successful intervention might generate. Our goal is to demonstrate how the MDEs change by varying (1) the number of practices, (2) the number of patients per practice, and (3) the outcomes being examined.<sup>6</sup> Given the variation in estimates and the few published reports of these numbers, researchers will ideally calculate their own MDEs from real data.

### How Many Patients Should Be Included per Practice?

**Finding: including more patients improves MDEs only slightly.**

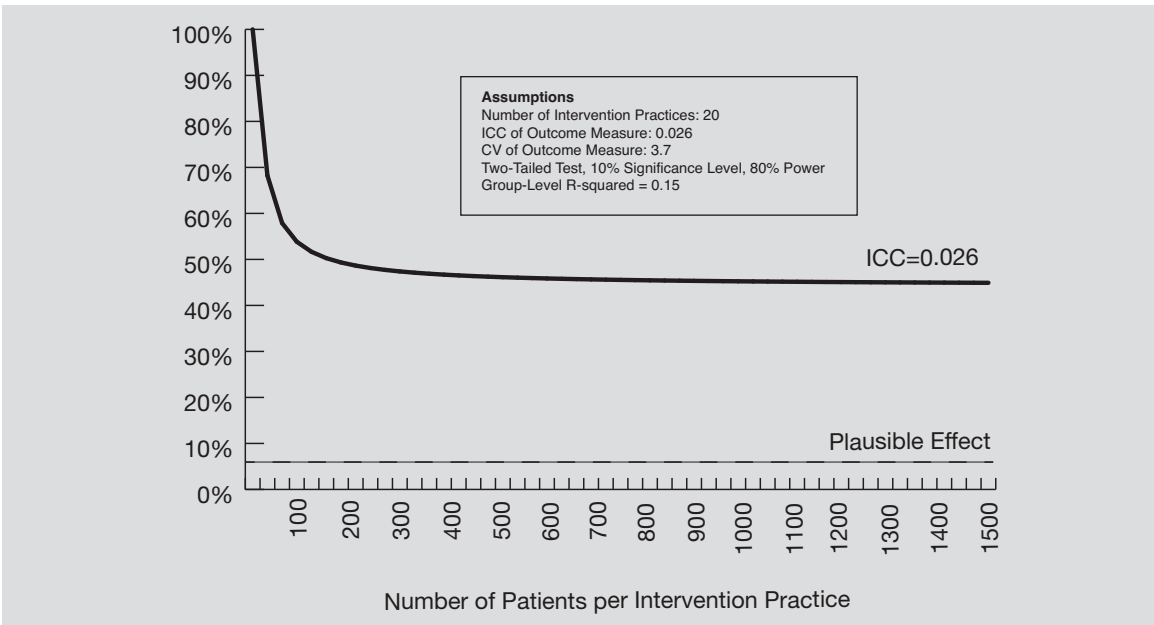
Figure 2 shows how MDEs for costs vary by the number of patients per practice, assuming there are 20 intervention and 20 comparison practices in a study. Here, we also assume that the CV is equal to 3.7 (the average CV for costs among studies in our literature review that used private payer data and included all patients in their sample), and that the ICC is equal to 0.026 (the average ICC for health care costs among studies in our literature review). As the number of patients per practice increases, the MDEs fall, but the gain from including more than several hundred patients (while holding the number of practices constant) is minimal (Figure 2). Thus, if including more patients increases costs (which is likely in studies involving surveys or chart abstraction), then it may be cost-effective to include fewer patients in the sample; in practice-level interventions, increasing the number of patients improves the MDE only slightly. Note also that even with 1,500 patients per practice, the MDEs in this graph are very large (about 45 percent, well above the effect an intervention might plausibly have on all patients); to further decrease the MDEs, studies must include more practices.<sup>7</sup>

---

<sup>6</sup> The graphs assume a 10 percent significance level (that is, a 10 percent chance of a false positive) and 80 percent power (chance of a true positive) and also assume that regression adjustment would result in the proportion of practice-level variance explained by regression adjustment (the group R-squared) of 0.15. As noted in Appendix B, a higher R-squared will improve the MDE. In fact, in education research, the R-squared is often higher than 0.5 when pre-intervention values of the outcome measure are included as control variables (Schochet 2008). However, in health care research, the R-squared is typically substantially lower; it is unlikely that practice-level health care studies will be able to improve the R-squared enough to substantially reduce the MDE. Also, depending on the level of clustering, individual-level control variables may be able to improve the MDE somewhat; however, greater gains in precision can generally be achieved in practice-level interventions by including practice-level control variables that reduce the group R-squared.

<sup>7</sup> Some readers may wonder why the figure does not examine more than 1,500 patients per practice. While large practices will have more than 1,500 patients, as the figure illustrates, these additional patients will barely alter the MDE unless clustering is very low. For example, with 30,000 patients per practice, using a typical ICC, the MDE is only slightly smaller (44.3 versus 45 percent) and is still larger than the effect size that most interventions could plausibly achieve.

**Figure 2. MDEs (%) for Cost or Hospitalization Outcomes Among All Patients—Varying the Number of Patients per Practice**

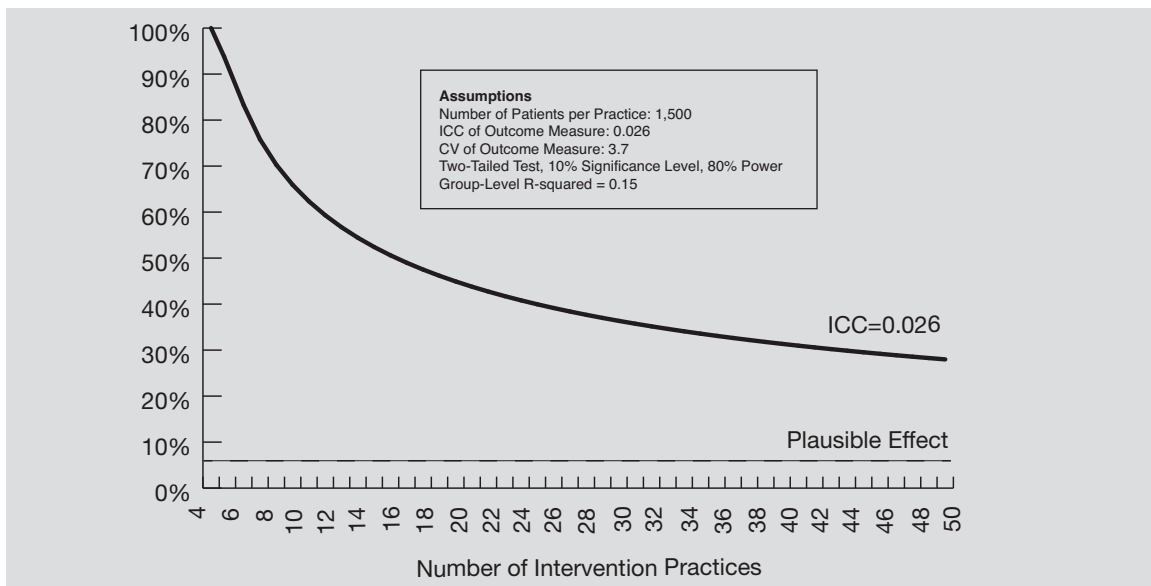


**How Many Intervention Practices Should Be Included?**

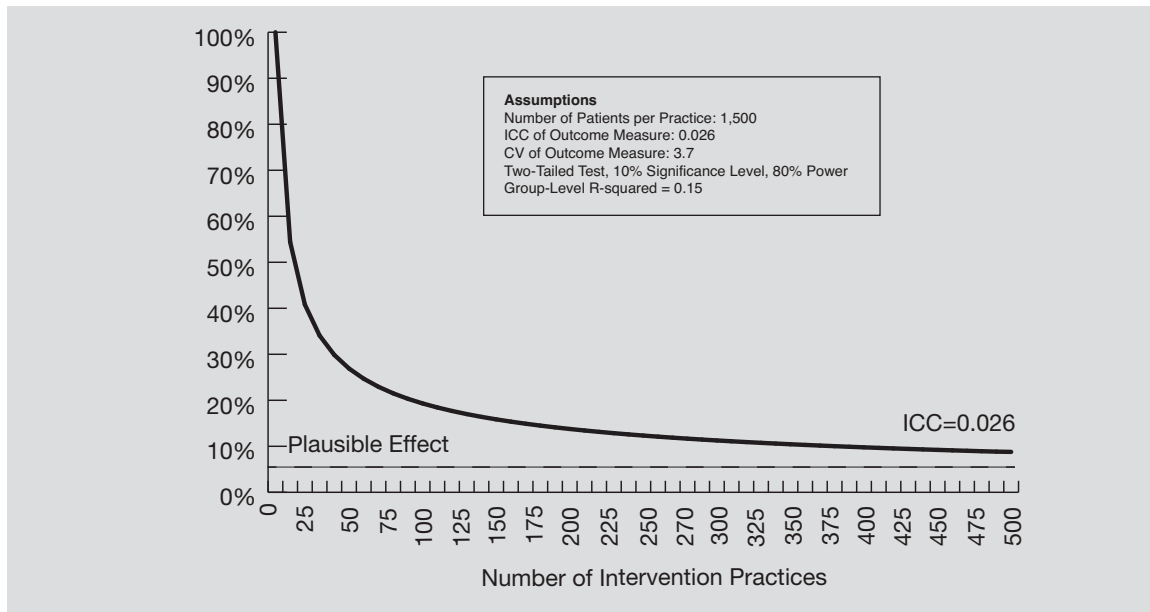
**Finding: it is critical to include as many practices as possible.**

Figures 3 and 4 vary the number of intervention practices (rather than the number of patients) and illustrate the importance of including as many practices as is feasible in practice-level interventions.

**Figure 3. MDEs (%) for Cost or Hospitalization Outcomes Among All Patients—Varying the Number of Practices: Small Studies**



**Figure 4. MDEs (%) for Cost or Hospitalization Outcomes Among All Patients—Varying the Number of Practices: Large Studies**



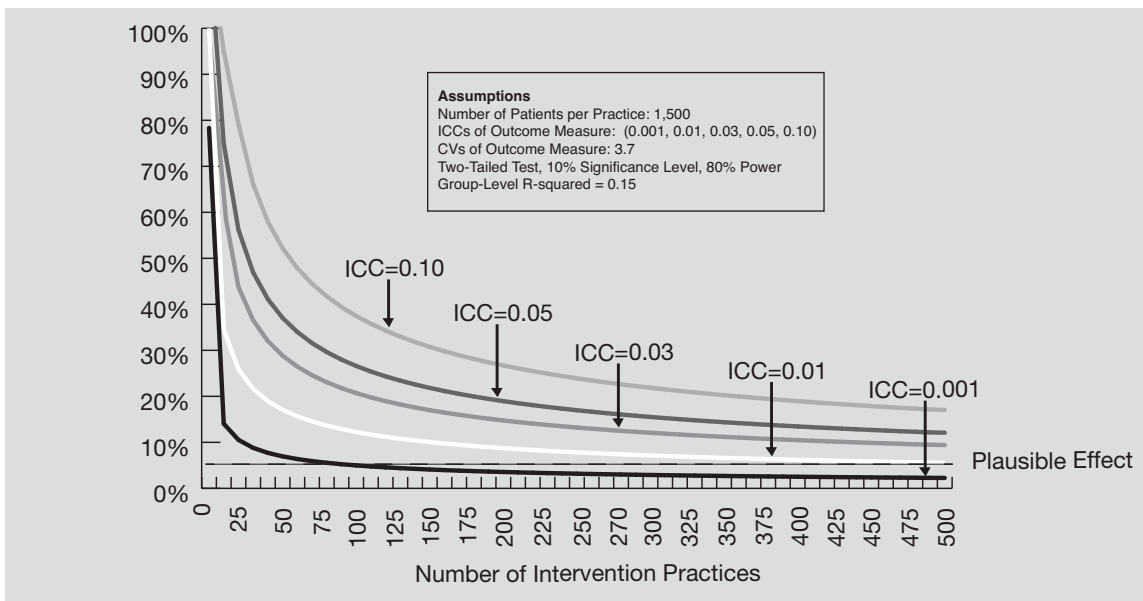
These figures have identical assumptions, but the scale of Figure 3 is blown up to show the MDEs for studies with fewer than 50 intervention and 50 control practices, as is common in medical home studies (Bitton, Martin, and Landon, 2010). In studies that have few practices, MDEs are implausibly large for adequate analysis of cost outcomes. For example, if a study has only 10 intervention and 10 control practices, MDEs are 66 percent when measured among all patients. MDEs fall as the number of practices increases, but even a study with 100 intervention practices (Figure 4) would have MDEs of about 20 percent. Because it is difficult to greatly reduce the costs of low-risk patients, it is virtually impossible for an intervention to have effects this large over all patients.

Although ICCs for health care costs and hospitalizations in the literature we reviewed ranged from 0.013 to 0.033, ICCs for a particular study could be much different, as it is difficult to generalize from the few ICCs that were reported in the literature. For example, proprietary data indicate that some ICCs are as low as 0.001. Therefore, Figure 5 shows MDEs under several different assumptions about the ICCs. Even with a very low ICC (0.001), a study would likely need 100 intervention and 100 control practices to be able to detect effects of 5 percent—what we believe would be an impressive impact for all patients.

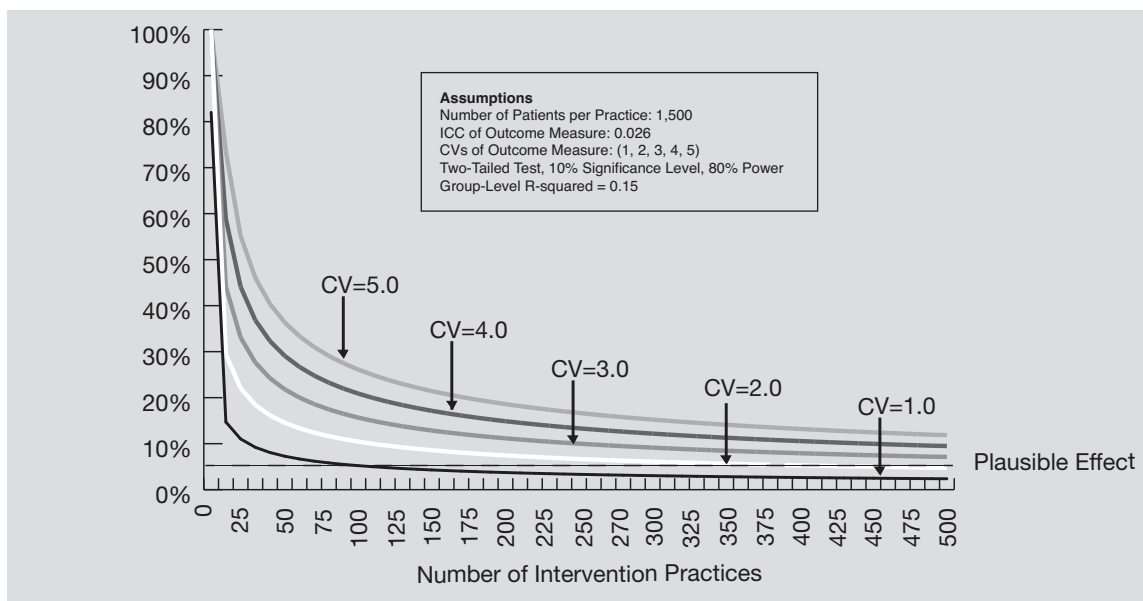
Similarly, across all study populations, there was also a wide range in CVs. Figure 6 illustrates how MDEs would change under different assumptions about the CV. Again, even a study with a CV of 1.0 (lower than the smallest CV reported in the literature for costs) would likely need 100 intervention and 100 control practices to be able to detect effects of 5 percent.



**Figure 5. MDEs (%) for Cost or Hospitalization Outcomes Among All Patients—Varying the Number of Practices and the ICC**



**Figure 6. MDEs (%) for Cost or Hospitalization Outcomes Among All Patients—Varying the Number of Practices and the CV**



Finally, although the graphs show MDEs for health care costs, the graphs would be fairly similar for hospitalizations and emergency room (ER) use, because the ICCs were similar in the studies we examined; however, compared to the MDEs for costs, the MDEs are likely to be slightly higher for hospitalizations (the result of a somewhat higher average CV) and slightly lower for ER visits (the result

of a slightly lower CV). Because hospital bed days have very high CVs, the impacts for bed days would have to be implausibly large for a study to be likely to detect them.

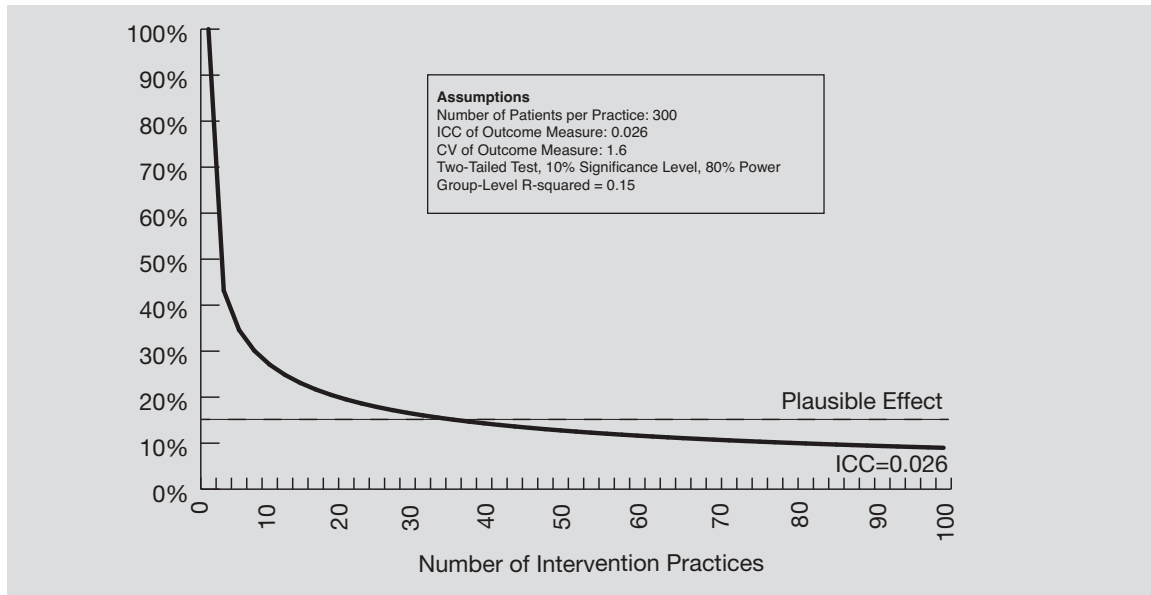
Given the large number of practices needed to detect effects for all patients, we turn to how researchers can design practice-level interventions so that the studies have sufficient power to detect impacts that are plausible.

## Which Patients Should Be Included for Each Outcome?

**Finding: treat all patients, but measure effects on costs and service use among the chronically ill.**

Most supporters of the medical home argue that it should be a model of care for all patients; however, research realities may necessitate measuring outcomes over subgroups of patients from within the medical home to increase the likelihood of finding a true effect. Our analyses indicate that limiting the population analyzed to the chronically ill (who are more likely to incur high health care costs) is an effective strategy to obtain lower MDEs, because the variation in outcomes for this subgroup is smaller (leading to a lower CV). Assuming the CV is 1.6 (based on the average CV for the chronically ill Medicare population), the MDE for the chronically ill is about 30 percent in small studies (with only 10 intervention practices) and about 13 percent in larger studies (with 50 intervention practices; see Figure 7 or Table 3 for MDEs).<sup>8</sup>

**Figure 7. MDEs (%) for Cost or Hospitalization Outcomes Among Chronically Ill Patients—Varying the Number of Practices**



<sup>8</sup> We used an assumption that 300 patients would have a chronic illness in each practice. Larger practices will have more patients but, as noted above, the additional patients will have a negligible effect on MDEs.

Not only are the MDEs for the chronically ill smaller than those for all patients, they are plausible to achieve (at least for studies with many practices), as there are more opportunities for a practice to alter the costs of care provided to those with chronic diseases. Therefore, researchers should consider assessing costs and hospitalizations for the chronically ill rather than for their full sample. Although the definition of *high risk* will depend on the study context, some examples of definitions are shown in Table E.2.

**Table 3. Minimum detectable effects for all patients and chronically ill patients, by number of intervention practices**

Number of Intervention Practices	Minimum Detectable Effect	
	All Patients	Chronically Ill Patients
500	9%	4%
200	14%	6%
100	20%	9%
50	28%	13%
20	45%	20%
10	66%	30%

Note: See Figures 4 and 7 for assumptions used to estimate these MDEs. Actual MDEs will vary for different practices, patients, and markets.

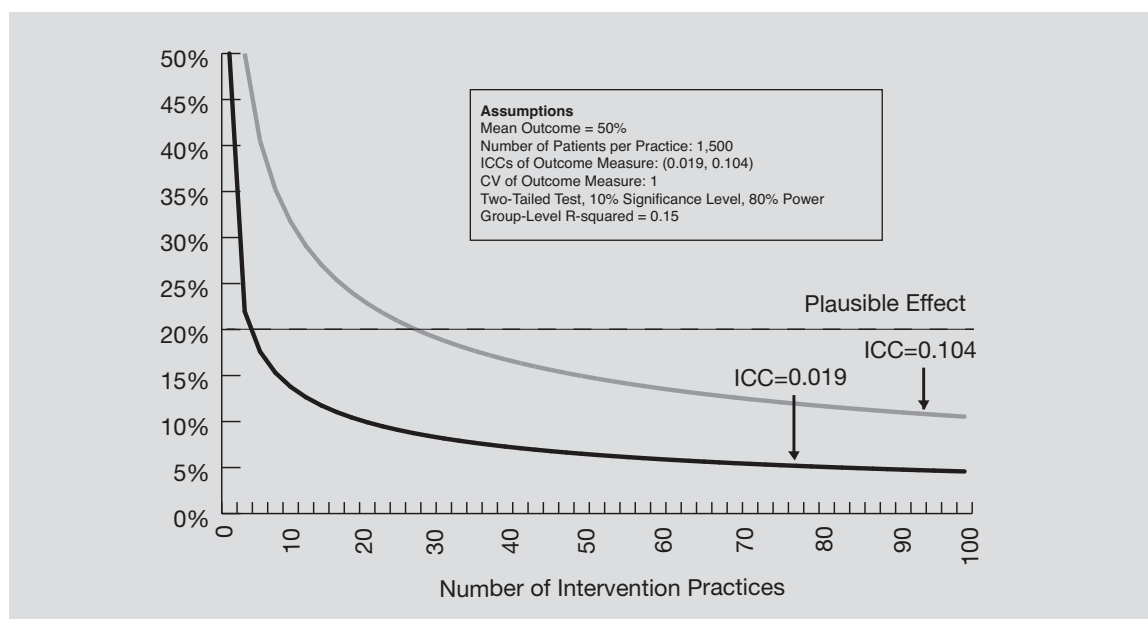
If researchers measure effects on cost and service use among a subset of patients, they can still estimate the costs of providing the intervention to all patients. The costs of delivering the medical home across all patients could be compared to the estimated effect for the chronically ill and the simulated effect for the remaining patients. To simulate the effects on health care cost among the non-chronically ill patients, researchers could generate three scenarios—one where there is no effect, an optimistic scenario (say, a 5 percent reduction), and a pessimistic scenario (say, a 5 percent increase).

While researchers may need to measure effects on continuous variables (such as cost and service use) among the chronically ill, they may be able to analyze—for all patients—binary or categorical variables, such as measures of quality of care and patient experience, which are expressed through a limited number of values. An example of a binary variable would be whether or not a patient reports being highly satisfied with care; an example of a categorical variable would be a five-point scale ranging from “very satisfied” to “very dissatisfied” with a specific aspect of experience. As discussed in the next section, continuous service use and cost measures could also be redefined to be categorical or binary variables; for example, for hospitalizations, a continuous measure would be defined as the number of hospitalizations, a binary variable as whether or not a patient was hospitalized.<sup>9</sup>

<sup>9</sup> While the ICCs for satisfaction and process outcomes tend to be larger and have a wider range than those for health utilization and cost outcomes, the variance for binary outcomes tends to be much smaller than the variance for continuous measures; the lower variance leads to a lower MDE (in spite of the higher ICCs).

It may be easier for a study to detect effects of a given proportionate size on binary and categorical variables. First, binary and categorical variables generally have lower CVs than continuous measures, and the MDEs are also generally lower.<sup>10</sup> Also, it may be easier for interventions to generate larger effects on many types of binary variables (such as quality-of-care process measures) than on continuous health care cost and service use measures. For example, it is somewhat easier for a practice to affect certain types of quality-of-care process measures, because the practice controls whether it orders certain tests and often needs only to successfully encourage patients to follow their instructions once annually (for example, for adults to get their flu shot); in contrast, not all hospitalizations are preventable, and to prevent those that are, the provider often needs to change the patients' long-term behavior, such as daily adherence to a diet, exercise, or prescription drug regimen. As shown in Figure 8, MDEs for a study with 10 intervention practices would be 15 percent to 35 percent, depending on the ICC. This size MDE is plausible; for example, an MDE of 20 percent is equivalent to changing the proportion of patients that received an influenza shot by 10 percentage points—from 50 percent to 60 percent [ $10/50 = 20\%$ ]. Many studies have reported effects of this size on quality-of-care process measures (for example, Counsell, Callahan, and Clark, 2007; Boulton, Reider, Frey, et al., 2008).

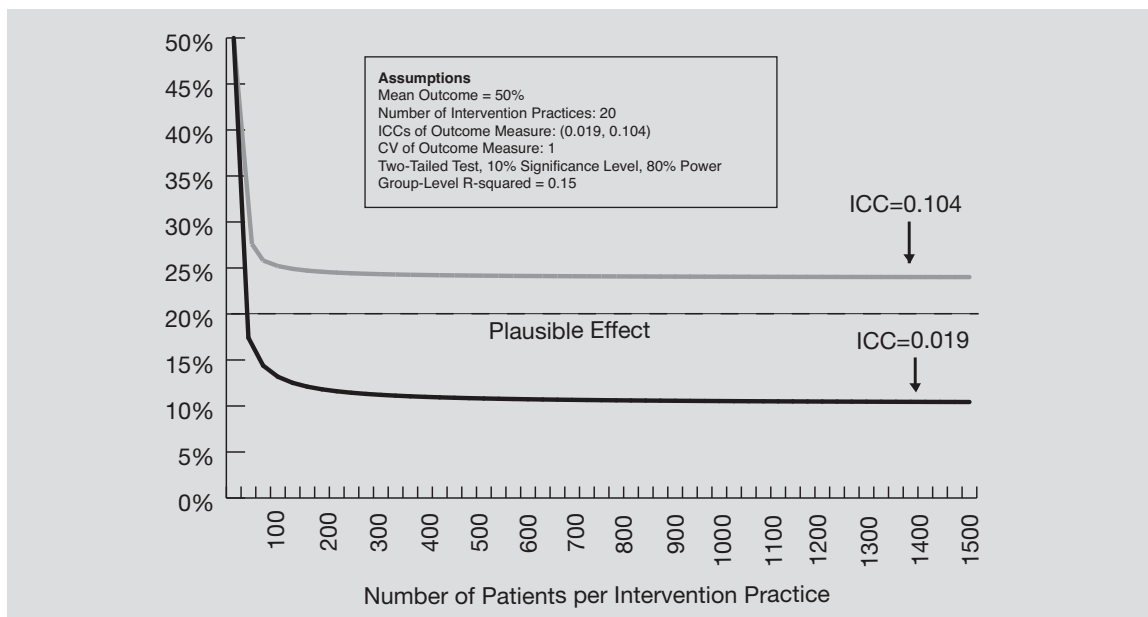
**Figure 8. MDEs (%) for Quality-of-Care or Satisfaction Outcomes Among All Patients—Varying the Number of Practices**



<sup>10</sup> Note that while the CVs for many of the quality-of-care and satisfaction measures are relatively low, the CVs for very rare outcomes (like mortality in all patients in a primary care practice) are relatively high.

Because MDEs are lower for binary variables, these types of measures can be assessed for all patients in a practice. However, the cost of data collection for each sample member could be high if the outcome measures are drawn from a survey or from chart review. Therefore, it might be wise to include only a fraction of the patients at each practice in the study sample, as adding more than 20 to 100 patients per practice reduces the MDE only slightly (see Figure 9).

**Figure 9. MDEs (%) for Quality-of-Care or Satisfaction Outcomes Among All Patients—Varying the Number of Patients per Practice**



### Can it Help to “Transform” the Outcome Variable to Reduce Variation?

**Finding:** while transforming outcomes might reduce the MDE somewhat, this strategy has several drawbacks.

A final strategy to consider is to reduce the influence of outliers for cost and service use outcomes. Doing so lowers the CV and therefore the MDE but raises challenges in estimation and interpretation. There are many ways to transform outcome measures. One is “capping” (also called “topcoding”) data. An example of capping a variable would be assigning, say, the value of the 95th percentile to all individuals who had costs greater than the 95th percentile.<sup>11</sup> Based on our past analyses using Medicare data, we suspect that capping the data would not lead to enough of an improvement in MDEs for all patients to make the MDEs for cost impacts low enough to detect effects of plausible size. Other strategies include taking the log of the outcome, creating a categorical variable (such as four quartiles of “number of hospitalizations”), or creating a binary variable (such as “any hospitalization”).

<sup>11</sup> For example, we found for one data set that capping costs to the 95th percentile reduced the CV and MDE by 30 percent. To translate this into more concrete terms, if the original MDE for costs was 50 percent, capping costs might lower the MDE to 35 percent [50% \* (1 - 30%) = 35%].

However, evaluators should weigh the tradeoffs that come with potentially lowering variance by reducing the influence of outliers. First, doing so could possibly reduce the apparent impacts of the intervention. For example, if the medical home is expected to reduce costs (or hospitalizations) for the highest-risk cases, then it may not make sense to arbitrarily limit the maximum costs (or number of hospitalizations) that a person might incur. Second, there are estimation challenges. Whenever researchers plan to transform variables, they will have to recalculate the MDEs and may need to adopt a more sophisticated estimation method (for example, using censored regressions to analyze topcoded data). Third, it may be difficult to interpret the effects on transformed variables. For example, if the researcher estimates the effects of the medical home on the distribution of costs (say, transformed into quartiles), it would be difficult for decisionmakers to interpret how changes in the distribution would translate into actual savings.

### **Can Adding More Comparison Practices Improve MDEs?**

Some readers may wonder whether adding more comparison practices lowers the MDE. Such a strategy appears attractive; although the cost of transforming and paying an additional medical home to include in the intervention group is often quite substantial, the main cost of adding comparison practices is typically much lower.

There are three important issues to assess when considering whether to add comparison practices. First, to avoid bias, comparison practices should be included only if they are well matched to intervention practices (that is, they must be similar in terms of the number and type of providers, use of health information technology, patient demographics, and other important baseline outcome measures).

Second, if the study has a fixed number of practices, allocating more to the comparison group will not help, because this increases the MDE. The MDE is, by definition, always lowest (best) when the study includes an equal number of intervention and control practices. For example, a study with 40 practices will have a lower MDE if they are split equally (20 intervention and 20 comparison) than if they are split unequally (for example, 10 intervention and 30 comparison).

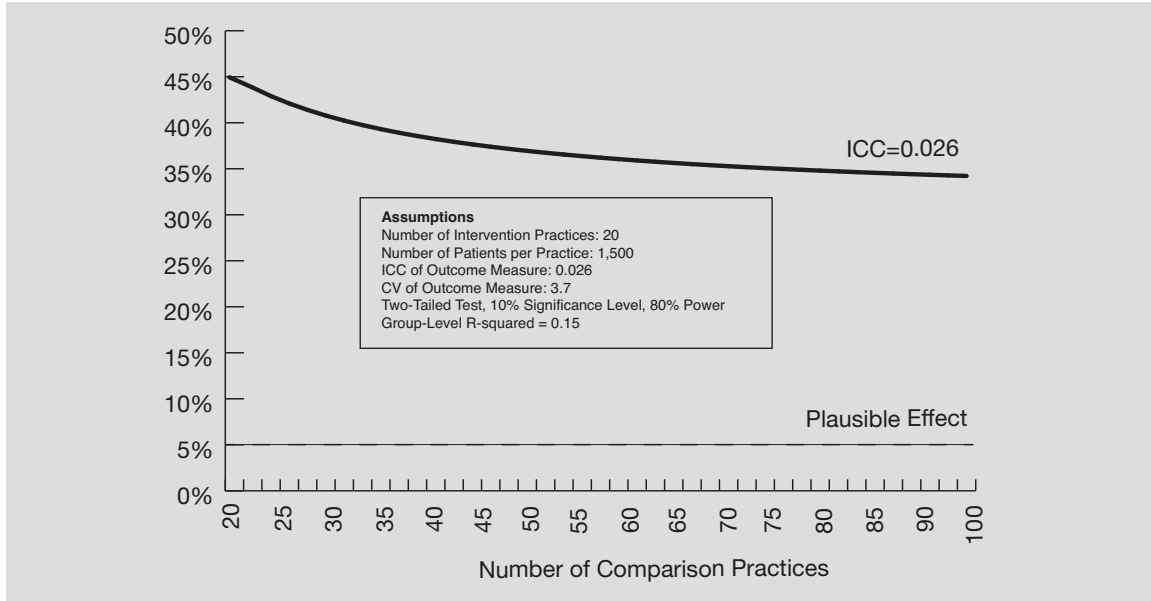
Third, for studies measuring cost and service use for all patients, the small improvement in the MDE is unlikely to be sufficient to measure plausible effects (see Figure 10). However, for studies that measure these outcomes only for the chronically ill, the MDE may fall enough to be plausible for smaller studies. For example, with 20 intervention and 20 comparison practices, the MDE for cost measured for the chronically ill is about 21 percent; this MDE falls to 18 percent if 40 rather than 20 comparison practices are used (see Figure 11). While the MDE drops even further by adding more comparison practices, finding a substantial number of potential comparison practices with similar characteristics and in markets similar to those of intervention practices is likely to be difficult.

### **Can a Study Improve Power by Accounting for Clustering at the Clinician (or Team) Level, Rather Than the Practice Level?**

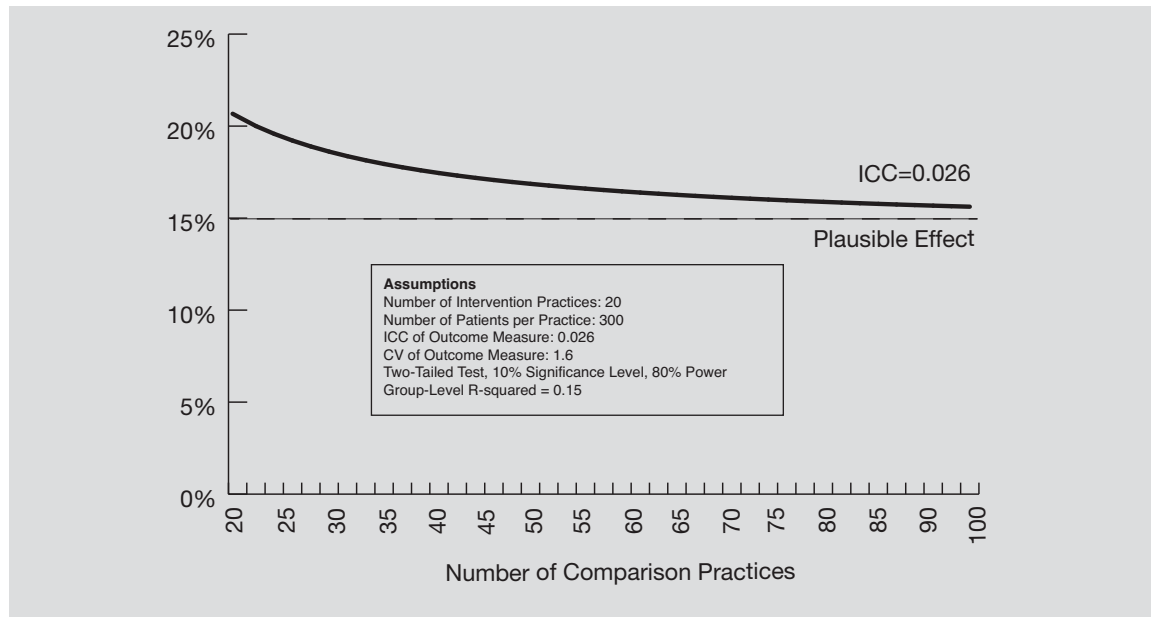
Addressing clustering only at the clinician level is not appropriate for tests of the full medical home intervention because it is a practice-level intervention. For example, expanded after-hours care is typically provided for the full practice, not for particular providers' patients. At a minimum, researchers need to adjust for clustering at the level of the intervention, so it is not appropriate to ignore the practice-level

clustering when evaluating medical home interventions. Furthermore, if outcomes of patients are clustered at the clinician level, researchers ideally would account for clustering at both the clinician and the practice level; adding this additional level of clustering would likely increase (rather than decrease) the MDE.

**Figure 10. MDEs (%) for Cost or Hospitalization Outcomes Among All Patients—Varying the Number of Comparison Practices** (Number of Intervention Practices Held Constant at 20)



**Figure 11. MDEs (%) for Cost or Hospitalization Outcomes Among Chronically Ill Patients—Varying the Number of Comparison Practices** (Number of Intervention Practices Held Constant at 20)



## Chapter 7. Summary and Conclusions

**Studies of the effectiveness of practice-level interventions will be unlikely to detect cost or service use reductions for all patients in a practice** because the cost-saving effects of such interventions are likely to be heavily concentrated in the subset of patients with chronic health problems, and the high variation in these outcomes over the full population of patients makes it hard to distinguish effects of programs from noise. Even with 50 intervention practices, cost reductions for all patients would likely have to be implausibly large (about 28 percent for our basic assumptions) to be detectable. Studies would probably need hundreds of intervention practices to be able to detect plausibly sized effects on cost and utilization for all patients in a practice.

**For a study to detect plausibly sized effects on costs and hospitalizations, analyses should be conducted on only the chronically ill/high-risk patients.** The MDEs for the chronically ill are lower because such patients have less variation in costs and hospitalizations. With 35 intervention and 35 control practices, studies will likely be able to detect cost reductions of 15 percent among the chronically ill. With 50 intervention and 50 control practices, the figure improves to as low as 13 percent. Perhaps more important, for this group, there are opportunities to reduce service use, so it is plausible for interventions to generate effects of this size.

**It is also possible to detect effects for quality-of-care and satisfaction outcomes across all patients** because these measures have less variation. In fact, a study with only 20 intervention and 20 control practices would likely have an MDE of about 24 percent, even if only a fraction of patients at the practice were included. For example, if the proportion of patients per practice who received a flu shot increased from 50 percent to 62 percent, a study with 20 intervention practices would be likely to detect it, as the increase of 12 percentage points is equivalent to 24 percent ( $0.12/0.50$ ), which is greater than the MDE.

**Evaluations can save money by collecting survey and chart review data on a sample of patients in a practice.** Depending on the degree of clustering, a study may be able to detect plausibly sized MDEs with only 20 to 100 patients per practice. Including more patients per practice only slightly improves the MDE. The small benefits this offers in terms of statistical power may not be worth the substantial costs associated with collecting survey or chart data from additional patients. Thus, studies should consider sampling patients when assessing binary or categorical outcomes based on surveys or charts, and focus their resources on increasing response rates among that sample. Most satisfaction and quality-of-care measures are intended to track all patients in the practices (not just the chronically ill), but patients with chronic illnesses can be oversampled for disease-specific measures. Although there may be minimal costs associated with acquiring and processing claims and administrative data for all patients in a practice, there are other reasons, as stated above, that cost and hospitalization data are best assessed in chronically ill patients.

**To design studies with adequate power to detect effects, for any given number of patients, it is clear that more practices with fewer patients per practice is preferable.** However, recruiting, transforming, and paying practices requires resources.



**While including more comparison practices reduces MDEs modestly without substantially increasing costs, finding suitable comparison practices will be difficult.** The modest improvement in MDEs will likely not be sufficient to detect plausible effects on cost and service use for all patients, but may be helpful for studies that measure these outcomes for the chronically ill.

If funders lack adequate resources to include enough practices to detect plausible effects on the outcomes of interest, they should consider fully the drawbacks and potential benefits of conducting different types of evaluations based on the goals for the study. Small studies can test how to implement a medical home and provide invaluable lessons. They may also be useful for determining what outcomes and approaches should be studied in future, larger studies. However, studies with insufficient statistical power may not provide the evidence about the effectiveness of medical homes needed to guide future decisions. If small studies correctly adjust for clustering, they are unlikely to find a statistically significant intervention-control difference—a result not of the intervention’s ineffectiveness, but simply of insufficient power. On the other hand, if such studies ignore clustering, the results will not be generalizable beyond the practices included in the study. There may be opportunities to undertake meta-analyses that combine the results of small evaluations or actually combine the data from several small evaluations into one analysis to increase the power; this will require that researchers specify common metrics and share data if they want to undertake one combined analysis. The Commonwealth Fund is currently convening evaluators of the medical home, and this is one topic of discussion (The Commonwealth Fund Web site).

## Implications

Many ongoing studies of the effectiveness of the medical home have fewer than 10 intervention practices and thus will likely be unable to detect the effects of the intervention on costs or hospitalizations across all patients in a practice. Fortunately, even in such small studies, it may be possible for researchers to detect effects on some outcomes and for some populations. First, researchers should include outcome measures with low variances, such as proportions that capture process measures (e.g., whether a patient received a vaccine) or satisfaction measures. Second, for outcome measures that have high variances, such as costs and hospitalizations, researchers should limit the population studied to those who are most likely to utilize services, such as those with chronic illnesses. Nonetheless, even though the study population for cost analyses must be limited to high-risk patients in order to detect effects of a plausible magnitude, the medical home model of care should still be provided to all patients. A third strategy is to consider transforming costs and service use outcomes to reduce variation and thereby reduce the MDE. The challenge of this approach is that it may make the impacts of the intervention appear smaller if the medical home reduced costs for the most expensive cases. In that case, it would not be wise to limit arbitrarily the maximum costs measured for an individual.

This paper raises some concerns about many recent and ongoing studies of medical home and medical home-like interventions that have fewer than 50 intervention practices yet aim to assess quantitative effects. These studies are likely to be underpowered for many outcomes (that is, they have too few practices to be likely to detect effects). Underpowered studies may not be able to demonstrate that an intervention improved outcomes, but this does not necessarily indicate that the medical home model did not work. These findings may be false negatives—that is, the study shows no statistically significant results, but the intervention actually worked.

Meta-analyses or combined analyses could overcome many of the challenges of evaluating the medical home. Such analyses would combine the results from many small studies to increase the power to detect effects. Collaboration among investigators (and funders) at the outset when designing evaluations will lay the foundation for such efforts, which would need to include common outcomes and metrics and, potentially, data-sharing agreements.

Our findings have implications for decisionmakers who are considering whether to fund medical homes or other practice-level interventions. First, decisionmakers should be wary of funding initiatives based on analyses that show significant results but are not adjusted for clustering, as such findings are likely to be false positives. Second, the lack of significant findings from underpowered studies (studies with too few practices) does not necessarily indicate that the medical home model does not work.

To help enable future evaluations to detect true effects of practice-level interventions, study designs should focus on increasing the number of practices (rather than the number of patients) in the sample. This may be challenging to do from an operational perspective, given the fixed costs involved in recruiting and transforming practices regardless of the number of patients in each practice, but it is critical from an evaluation perspective. Because many health-related outcome measures (such as health care costs and hospitalizations) have high variances, it is particularly important that medical home studies have a large number of practices. Moving forward, larger studies (with hundreds of practices) or thoughtfully executed meta-analyses may be needed to generate credible evidence about the medical home's effects on costs across all patients in a practice.

## References

- Agency for Healthcare Research and Quality. What is the PCMH? AHRQ's definition of the medical home. [http://www.pcmh.ahrq.gov/portal/server.pt/community/pcmh\\_\\_home/1483/what\\_is\\_pcmh](http://www.pcmh.ahrq.gov/portal/server.pt/community/pcmh__home/1483/what_is_pcmh). Accessed September 14, 2011.
- American Academy of Family Physicians, American Academy of Pediatrics, American College of Physicians, American Osteopathic Association. Joint principles of a patient-centered medical home. February 2007. [http://www.aafp.org/online/etc/medialib/aafp\\_org/documents/policy/fed/jointprinciplespcmh0207.Par.0001.File.dat/022107medicalhome.pdf](http://www.aafp.org/online/etc/medialib/aafp_org/documents/policy/fed/jointprinciplespcmh0207.Par.0001.File.dat/022107medicalhome.pdf). Accessed September 14, 2011.
- Ash AS, Ellis RP, Pope GC, et al. Using diagnoses to describe populations and predict costs. *Health Care Financ Rev* 2000;21(3):7-28.
- Bitton A, Martin C, Landon BE. A nationwide survey of patient centered medical home demonstration projects. *J Gen Intern Med* 2010;25(6):584-92.
- Bland JM. Cluster randomised trials in the medical literature: two bibliometric surveys. *BMC Med Res Methodol* 2004;4(1):21.
- Bloom H. Learning more from social experiments: evolving analytic approaches. New York: Russell Sage Foundation; 2005.
- Boult C, Reider L, Frey K, et al. Early effects of "Guided Care" on the quality of health care for multimorbid older persons: a cluster-randomized controlled trial. *J Gerontol A Biol Sci Med Sci* 2008;63(3):321-7.
- Boyd CM, Reider L, Frey K, et al. The effects of guided care on the perceived quality of health care for multi-morbid older persons: 18-month outcomes from a cluster-randomized controlled trial. *J Gen Intern Med* 2010;25(3):235-42.
- Campbell MK, Elbourne DR, Altman DG, CONSORT group. CONSORT statement: extension to cluster randomised trials. *BMJ* 2004;328(7441):702-8.
- Campbell MK, Mollison J, Grimshaw JM. Cluster trials in implementation research: estimation of intracluster correlation coefficients and sample size. *Stat Med* 2001;20(3):391-9.
- Campbell SM, Hann M, Hacker J, et al. Identifying predictors of high quality care in English general practice: observational study. *BMJ* 2001;323(7316):784-7.
- The Commonwealth Fund. The Patient-Centered Medical Home Evaluators' Collaborative. March 2011. <http://www.commonwealthfund.org/Content/Publications/Other/2010/PCMH-Evaluators-Collaborative.aspx>. Accessed September 14, 2011.
- Counsell SR, Callahan CM, Tu W, et al. Cost analysis of the Geriatric Resources for Assessment and Care of Elders care management intervention. *J Am Geriatr Soc* 2009;57(8):1420-6.
- Counsell SR, Callahan CM, Clark DO, et al. Geriatric care management for low-income seniors: a randomized controlled trial. *JAMA* 2007;298(22):2623-33.
- Dale S, Lundquist E. Revised power calculations for the MCMP demonstration. Memorandum to CMS. Princeton, NJ: Mathematica Policy Research; 2011.
- Donner A. A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. *Int Stat Rev* 1986;54(1):67-82.
- Feldman RD, Parente ST. Enrollee incentives in consumer directed health plans: spend now or save for later? *Forum for Health Economics & Policy* 2010;13(2) (Health Economics), Article 3. <http://www.bepress.com/fhpep/13/2/3/>. Accessed September 20, 2011.
- Gilfillan RJ, Tomcavage J, Rosenthal MB, et al. Value and the medical home: effects of transformed primary care. *Am J Manag Care* 2010;16(8):607-14.

- Goetzel RZ, Gibson TB, Short ME, et al. A multi-worksites analysis of the relationships among body mass index, medical utilization, and worker productivity. *J Occup Environ Med* 2010;52(Suppl 1):S52-8.
- Huang IC, Diette GB, Dominici F, et al. Variations of physician group profiling indicators for asthma care. *Am J Manag Care* 2005;11(1):38-44.
- Leff B, Reider L, Frick KD, et al. Guided care and the cost of complex healthcare: a preliminary report. *Am J Manag Care* 2009;15(8):555-9.
- Littenberg B, MacLean C. Intra-cluster correlation coefficients in adults with diabetes in primary care practices: the Vermont Diabetes Information System field survey. *BMC Med Res Methodol* 2006;6:20.
- McCall N, Cromwell J, Urato C. Evaluation of Medicare Care Management for High Cost Beneficiaries (CMHCB) Demonstration: Massachusetts General Hospital and Massachusetts General Physicians Organization (MGH). Research Triangle Park, NC: RTI International; September 2010.
- Murray DM. Design and analysis of group-randomized trials. New York: Oxford University Press; 1998.
- Ozminkowski RJ, Smith MW, Coffey RM, et al. Private payers serving individuals with disabilities and chronic conditions. Washington, DC: U.S. Department of Health and Human Services; January 2000.
- Peikes D, Peterson G, Schore J, et al. Effects of care coordination on hospitalizations and health care expenditures among high-risk Medicare beneficiaries: 11 randomized trials. Princeton, NJ: Mathematica Policy Research; 2011.
- Philipson T, Seabury S, Lockwood L, et al. Geographic variation in health care: the role of private markets. *Brookings Papers on Economic Activity Spring* 2010;:325-55.
- Pope G, Ellis R, Ash A, et al. Diagnostic cost group hierarchical condition category models for Medicare risk adjustment: Final report. Waltham, MA: Health Economics Research; 2000.
- Potiriadis M, Chondros P, Gilchrist G, et al. How do Australian patients rate their general practitioner? A descriptive study using the General Practice Assessment Questionnaire. *Med J Aust* 2008;189(4):215-9.
- Reid RJ, Coleman K, Johnson EA, et al. The group health medical home at year two: cost savings, higher patient satisfaction, and less burnout for providers. *Health Aff (Millwood)* 2010;29(5):835-43.
- 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;15(9):e71-87.
- Rittenhouse DR, Shortell SM, Fisher ES. Primary care and accountable care—two essential elements of delivery-system reform. *N Engl J Med* 2009;361(24):2301-3.
- Schochet PZ. Statistical power for random assignment evaluations of education programs. *J Educ Behav Stat* 2008;33(1):62-87.
- Short ME, Goetzel RZ, Pei X, et al. How accurate are self-reports? Analysis of self-reported health care utilization and absence when compared with administrative data. *J Occup Environ Med* 2009;51(7):786-96.
- Zhao Y, Ash AS, Ellis RP, et al. Predicting pharmacy costs and other medical costs using diagnoses and drug claims. *Med Care* 2005;43(1):34-43.

### AHRQ Definition of the Medical Home

The medical home model holds promise as a way to improve health care in America by transforming how primary care is organized and delivered. Building on the work of a large and growing community, the Agency for Healthcare Research and Quality (AHRQ) defines a medical home not simply as a place but as a model of the organization of primary care that delivers the core functions of primary health care (Agency for Healthcare Research and Quality).

The medical home encompasses five functions and attributes:

**Patient-centered:** The primary care medical home provides relationship-based primary health care that is oriented toward the “whole person.” Partnering with patients and their families requires understanding and respecting each patient’s unique needs, culture, values, and preferences. The medical home practice actively supports patients in learning to manage and organize their own care at the level the patient chooses. Recognizing that patients and families are core members of the care team, medical home practices ensure that they are fully informed partners in establishing care plans.

**Comprehensive care:** The primary care medical home is accountable for meeting the bulk of each patient’s physical and mental health care needs, including prevention and wellness, acute care, and chronic care. Comprehensive care requires a team of care providers, possibly including physicians, advanced practice nurses, physician assistants, nurses, pharmacists, nutritionists, social workers, educators, and care coordinators. Although some medical home practices may bring together large and diverse teams of care providers to meet the needs of their patients, many others, including smaller practices, will build virtual teams linking themselves and their patients to providers and services in their communities.

**Coordinated care:** The primary care medical home coordinates care across all elements of the broader health care system, including specialty care, hospitals, home health care, and community services and supports. Such coordination is particularly critical during transitions between sites of care, such as when patients are being discharged from the hospital. Medical home practices also excel at building clear and open communication among patients and families, the medical home, and members of the broader care team.

**Superb access to care:** The primary care medical home delivers accessible services with shorter waiting times for urgent needs, enhanced in-person hours, around-the-clock telephone or electronic access to a member of the care team, and alternative methods of communication such as email and telephone care. The medical home practice is responsive to patients’ preferences regarding access.

**A systems-based approach to quality and safety:** The primary care medical home demonstrates a commitment to quality and quality improvement by ongoing engagement in activities such as using evidence-based medicine and clinical decision-support tools to guide shared decisionmaking with patients and families, engaging in performance measurement and improvement, measuring and responding to patient experiences and patient satisfaction, and practicing population health management. Publicly

sharing robust quality and safety data and improvement activities is also an important marker of a system-level commitment to quality.

AHRQ recognizes the central role of health IT in successfully operationalizing and implementing the key features of the medical home. In addition, AHRQ notes that building a primary care delivery platform that the Nation can rely on for accessible, affordable, high-quality health care will require significant workforce development and fundamental payment reform. Without these critical elements, the potential of primary care will not be achieved.

## Appendix B

### Calculating Minimum Detectable Effects and Effective Sample Sizes

This appendix describes the factors used to calculate minimum detectable effects (MDEs) in practice-level interventions. Chapter 1 describes the general approach for calculating MDEs, Chapter 2 explains how to tailor the MDE calculation for a practice-level intervention, and Chapter 3 describes how to calculate effective sample sizes after accounting for clustering.

#### General Approach for Calculating MDEs

In general, the MDE of a research design depends on several factors:

- The policymaker's comfort level with the chance of erroneously concluding the medical home works when it really does not (the false positive rate, called the significance level, the Type I error rate, or  $\alpha$ ).
- The comfort level with incorrectly concluding the medical home does not work when it does (the false negative rate, or Type II error rate, called  $\beta$ , where  $(1-\beta)$  is the power level.
- The number of degrees of freedom (denoted as  $df$ ), a measure of the number of independent pieces of information that are analyzed in estimating the effects of an intervention on outcomes, in the statistical model used to estimate the intervention impact.
- The standard error of the impact estimate ( $SE$ ), the extent to which sample impact estimates could vary from the true program impact over repeated samples.

The coefficient of variation ( $CV$ ), a measure of the variation, or noise, in the outcome measure, defined as the standard deviation ( $\sigma$ ) of the outcome measure divided by the mean ( $\mu$ ).

The MDE of an experimental design, expressed in terms of percentage changes in the outcome measure,<sup>12</sup> can be written generally as:

$$1. \text{ MDE (\%)} = CV * M(\alpha, \beta, df) * \frac{SE}{\sigma}$$

In this equation,  $M$  is the standard error multiplier, a constant that depends on the study's chosen false positive rate or significance level, the statistical power level  $(1-\beta)$ , and the number of degrees of freedom. Researchers commonly set the significance level to 5 or 10 percent and the power level to 80 percent. A statistical power level of 80 percent implies that a study will fail to detect a true effect of a given size (and commit a Type II, or false negative, error) with a 20 percent probability. While 80 percent power and a 5 percent or 10 percent significance level are conventional, the chosen statistical power and significance level of a research design should be determined by the relative discomfort with making Type I and Type II errors in each specific context. Researchers sometimes relax statistical significance requirements in health care service interventions delivered at the practice or clinic level where adequate power is difficult to achieve and a Type I error would not be as problematic as in, for example, certain types of clinical trials.

<sup>12</sup> Equation 1 presents the MDE in percentage terms to facilitate comparisons across different study outcome measures.

## B. Tailoring MDE Calculations for Practice-Level Interventions

The medical home model is a practice-level intervention that involves changing the way all patients within a practice or clinic are served. As a result, studies generally designate a number of sites that will receive the intervention, and then select analogous units for the comparison group. When evaluating these interventions, researchers need to assess the extent of clustering in the data to accurately calculate the standard errors and p-values, which determine whether a finding is statistically significant. In common medical home interventions that target entire practices, the study design must account for clustering at the practice level, or the extent to which individual outcomes are correlated within practices. If alternative implementation models target different levels of organization, such as practice sites or individual clinicians or teams, clustering must be assessed at the same logical level.<sup>13</sup> It is important to design studies in which the intervention and comparison groups are constructed at the same level as the implementation model, to prevent cross-contamination or spillover effects (Bloom, 2005). Contamination might occur if some clinicians within a practice were in an evaluation's intervention group and some were in the comparison group, and the intervention clinicians shared with comparison clinicians ideas about ways practice patterns might be changed.

In many common study designs, such as the medical home, entire practices are selected into intervention and comparison groups, and mean patient outcomes are compared after the intervention begins. In this type of study, with one level of clustering (that is, patients are clustered within practices), MDEs can be expressed as<sup>14</sup>:

$$2. \text{ MDE (\%)} = CV * M(\alpha, \beta, df) * \sqrt{\left(\frac{ICC(1-R_G^2)}{P(1-P)G}\right) + \left(\frac{(1-ICC)(1-R_n^2)}{P(1-P)Gn}\right)}$$

Equation 2 illustrates how several facets of the research design influence the MDEs in clustered evaluations like the medical home. It is important to understand the benefits and drawbacks of manipulating these different study elements, especially relative to the costs of or savings from implementing associated changes. Below, we describe each factor in the equation and how it affects the MDE.

**Coefficient of Variation (CV).** As noted above, the CV is a measure of the variation, or noise, in the outcome measure (standard deviation divided by mean). Ideally, the CV will be low, because this leads to a smaller MDE. Intuitively, a lower CV means there is less random variation (or “noise”) in the outcome measure, which makes it easier to attribute differences in outcomes between intervention and comparison groups to the intervention itself; conversely, if the CV is high, it is difficult to distinguish the effect of the intervention from noise in the outcome.

---

<sup>13</sup> More generally, researchers should consider adjusting for clustering (by incorporating clustering terms into variance expressions for the impact estimate) at any level where either intervention-comparison assignment occurs or sampling occurs in the study. For example, if an evaluation samples cities, and then practices within cities, the study should account for clustering at both the city and the practice level.

<sup>14</sup> This formula would have to be modified to account for multiple levels of clustering (for example, if a researcher wanted to evaluate patient-level outcomes, but patients were clustered within physicians and physicians within practices). See Bloom (2005) for a derivation of this formula.



**2. The Standard Error Multiplier ( $M$ ).** This is a constant determined by the chosen statistical significance and power level of the study (as discussed above), as well as the regression model's degrees of freedom.<sup>15</sup> The number of degrees of freedom is defined as the number of independent pieces of information or observations that go into the calculation of a given statistic, minus the number of intermediate parameters used in the model. In clustered studies, individual observations are not independent of one another but rather correlated within groups (practices), so the number of independent observations is conservatively defined as the total number of study practices rather than the total number of study patients. Therefore, the number of degrees of freedom in clustered studies where intervention and control means are compared to assess program impacts is equal to the total number of study practices minus the number of practice-level covariates included in the impact estimation model, minus 2. Commonly used rule-of-thumb multipliers are 2.5 (two-tailed test: 10 percent significance; 80 percent power; at least roughly 30 degrees of freedom) or 2.8 (two-tailed test; 5 percent significance; 80 percent power; at least roughly 30 degrees of freedom). Increasing the degrees of freedom, as well as relaxing power and significance requirements, will lead to decreases in the study's MDE.

**The Intraclass Correlation Coefficient ( $ICC$ ).** The ICC is a measure of clustering within groups, such as practices. It is defined as the ratio of group-level outcome variance to total outcome variance. The MDE of a study will decrease as the ICC falls. If patients within practices tend to have similar outcomes and average patient outcomes tend to differ across practices, then the ICC will be high, and it will be difficult to tell whether outcome differences are due to the intervention or are simply the result of which practices were included in the intervention and comparison groups. Conversely, if patient outcomes are similar across practices, then the ICC (and therefore the MDE) will be relatively low. Patient-level data, with patients attributed to practices, are needed to estimate the ICCs for each study outcome measure (see Appendix F for sample code for calculating ICCs).

**The Number of Intervention and Comparison Practices or Groups ( $G$ ).** The MDE will decrease as the number of intervention and comparison practices rises. Adding more providers to the study sample increases the likelihood of intervention and comparison groups that are similar before the intervention begins. This makes it easier to attribute outcome differences between intervention and comparison groups to the intervention as opposed to noise in the data. Adding practices is the most effective way to lower the MDE.

**Number of Patients per Practice ( $n$ ).** The MDE will decrease as the number of patients per practice rises, but with diminishing returns. Increasing the number of study practices will improve precision more than adding additional patients per practice, even if the same total number of patients is included. In other words, it is better to have 20 practices of 1,000 patients each than 5 practices of 4,000 patients each.

**Proportion of Outcome Variance Explained by Regression Control Variables ( $R_n^2$  and  $R_G^2$ ).**

Including control variables in regression models used to estimate impacts can lower the MDE of a study because the control variables will help explain some of the variation in the outcome measure. While it is

---

<sup>15</sup> For a two-tailed t-test, the multiplier can be expressed mathematically as:  $M=(T_{\alpha/2} + T_{\beta})$ , where  $T_{\alpha/2}$  and  $T_{\beta}$  are critical values at which the t-distribution (with the associated model degrees of freedom) has a cumulative density of  $(1-\alpha/2)$  and  $(1-\beta)$ , respectively.

important to include control variables that reduce unexplained individual (patient)-level variance (that is, increase  $R_n^2$ ), far greater gains in precision can generally be achieved in practice-level interventions by including control variables that reduce unexplained group (practice)-level variance (increase  $R_G^2$ ). Pre-intervention measures of outcome variables and practice-level covariates such as practice size and patient demographics often serve as the best control variables if available.

**Proportion of Practices Allocated to the Intervention Group ( $P$ ).** Standard errors depend on intervention and comparison sample sizes, and MDEs are minimized for a given sample size when an equal number of practices are assigned to the intervention and comparison groups. However, in contexts where the costs of adding intervention practices are high relative to adding more comparison practices, increasing the relative size of the comparison group to a certain extent may still lead to important gains in precision.<sup>16</sup>

As an example of how to calculate MDEs based on the equation described above, assume that a study randomizes 30 practices assigned in equal proportions to the intervention or control group, and that each practice serves 2,000 patients on average. Further assume that (1) based on background research, the researchers estimate the CV of hospitalizations will be 2.0 and expect that the regression model used to estimate impacts will explain 15 percent of group-level variance with five group-level control variables (that is,  $R_G^2$  is equal to 0.15), and (2) the ICC for the outcome is 0.03. The MDE of this research design in a two-tailed test at the 90 percent confidence level ( $\alpha=0.10$ ) with 80 percent statistical power ( $\beta=0.20$ ) would be 30.3 percent.<sup>17</sup> In other words, the study could significantly detect a true intervention effect equal to 30.3 percent of the control group mean with a probability of 80 percent. To make the example even more concrete, if costs were the outcome measure, and the average patient cost per month were \$1,000, this study would detect an effect (a reduction or an increase) of \$303 (30.3 percent of \$1,000) most (80 percent) of the time. The study would be less likely to detect effects that were smaller than 30.3 percent.

$$3. \text{ MDE (\%)} = 2.0 * 2.57 * \sqrt{\left(\frac{0.03(0.85)}{0.5(0.5)*30}\right) + \left(\frac{0.97}{0.5(0.5)*30*2000}\right)}$$

As shown in the example above, MDEs in clustered studies can be quite large, and it is unlikely that an intervention could generate an effect of this size. This means that methodological decisions, especially those related to sample sizes, outcome measures, and study populations, become extremely important in ensuring that the research design has adequate statistical precision to ensure the study will generate useful findings.

<sup>16</sup> Bloom (2005) provides a broader discussion of the costs and benefits of unbalanced samples.

<sup>17</sup> The standard error multiplier ( $M=T_{\alpha/2} + T_{\beta}$ ) in this example can be calculated by looking up the critical t-values that correspond to a probability density of 0.05 ( $\alpha/2$ ) and 0.20 ( $\beta$ ) on a t-distribution reference table with 23 (number of clusters – cluster-level covariates – 2) degrees of freedom, or by plugging 0.05 and 0.20 into an inverse t-tail function such as `invttail()` in Stata. In this example, the critical t-values are  $T_{\alpha/2}=1.71$ ,  $T_{\beta}=0.86$ ; the multiplier ( $M$ ) can be found by adding these two terms together:  $1.71 + 0.86 = 2.57$ .

## Effective Sample Size in Clustered Design

Another way of thinking about clustering is in terms of how it changes the effective sample size. The effective sample size is determined by the actual sample size, the ICC, and the average number of patients per cluster ( $n$ ), according to the following formula:

- $$\text{Effective Sample Size} = \frac{\text{Actual Sample Size}}{(1 + \text{ICC} * (n - 1))}$$

If there is no clustering (that is, the ICC is equal to zero), the effective sample size is equal to the actual number of patients in the evaluation. Suppose there were 20,000 patients in total, spread across 20 practices (or 1,000 patients per practice). With no clustering, the effective sample equals the actual sample, 20,000. If there is maximum clustering (that is, the ICC is equal to 1), the effective sample size is  $20,000 / [1 + 1 * (1,000 - 1)] = 20$ , which is the number of practices—that is, the study effectively has only 20 unique observations. In practice, the amount of clustering tends to be closer to 0 than to 1. Suppose the ICC is 0.01; the effective sample size would then be  $20,000 / [1 + 0.01 * (1000 - 1)]$ , or  $20,000 / 10.99$ , or 1,819.8. That is, this clustered sample of size 20,000 has the equivalent power of a simple random sample of size 1,819.8. If the clustering is 0.1, the effective sample size falls to 198.2.

### **Explanation of Figure 1 on False Positive Rates When Clustering Is Ignored**

Figure 1 shows why studies need to take clustering into account. Decisionmakers are typically comfortable with some possibility of a false positive. In this situation, a false positive would be concluding that the medical home works when in fact it does not. By convention, we typically allow a 5 percent chance of a false positive (referred to as the alpha ( $\alpha$ ), Type I error rate, or significance level). In this graph, the horizontal line shows a 10 percent rate as the flat line. This indicates that decisionmakers would be willing to accept a higher (10 percent) false positive rate. But if the data are clustered—that is, the distribution of patient outcomes in one practice differs from the distribution in another—the chance of a false positive rises to levels decisionmakers will likely be uncomfortable with. A moderate level of clustering, if not accounted for, can lead to a 60 percent false positive rate, and if there is heavy clustering, the false positive rate could grow to 75 percent or more.

## Appendix D

### Sample Effect Sizes Found in Targeted Literature Review

We conducted a targeted literature review, examining selected health care interventions that were intended to improve quality of care and reduce health care costs, such as medical home or disease-management evaluations, to estimate plausible effect sizes. Table D.1 summarizes our findings.

**Appendix Table D.1. Examples of selected effects found in a targeted literature review**

Study (Author)	Population	Gross Savings (Without Costs of Intervention Unless Stated)	Reduced Hospitalizations
Group Health (Reid et al. 2010; 2009)	All, Privately Insured	Including intervention costs, 2%	6%
Geisinger (Gilfillan et al. 2010)	All, Medicare	NDE	18%
Geriatric Resources for Assessment and Care of Elders (GRACE) (Counsell et al. 2007)	All, Medicare	Including intervention costs, increased costs 28% in year 1, 14% in year 2, cost neutral in year 3.	NDE in years 1 and 2
GRACE (Counsell et al. 2007)	Chronically III, Medicare	Cost neutral in years 1 and 2, reduced costs 23% in year 3	NDE in year 1, 44% in year 2
Guided Care (Leff et al. 2009; Boyd et al. 2010)	Chronically III, Medicare	NDE	NDE
Medicare Care Management for High Cost Beneficiaries (CMHCB) (McCall et al. 2010)	Chronically III, Medicare	12.1%	NA
Medicare Coordinated Care Demonstration (Peikes et al. 2011)	Chronically III, Medicare	5.7%	10.7%

Note: Differences are statistically significant at the 10 percent level.

NA = not available; NDE = no detectable effect.

## Appendix E

### Inputs from the Literature for Calculating MDEs

To help inform assumptions underlying MDE calculations, we compiled CVs and ICCs from 18 published and unpublished studies. Within the health care literature, our search terms for ICCs included “practice-level interventions,” “clustered designs,” “intracluster correlation,” and “intraclass correlation.” We found that only a limited number of practice-level studies mentioned adjusting for clustering, and only a subset of those studies actually reported their ICCs. This infrequent reporting of ICCs is consistent with reviews of the health care literature that found that only about 20 percent of clustered randomized trials took clustering into account in calculating the study sample size needed to have adequate statistical power, and only about half account for clustering in the analysis (see, for example, Bland, 2004). Moreover, because health insurance cost data are often proprietary and many studies use health utilization measures (rather than costs) as outcome measures, only a handful of studies reported ICCs for costs. Because ICCs were rarely reported, we asked the lead authors of many practice-level studies to provide their ICCs; some of the ICCs in our tables are based on personal communications with the study authors and are not reported in the published paper or report.

We have more CV estimates than ICC estimates, partly because we drew CVs from a broader set of literature. For example, CVs are often reported in studies on risk-adjustment. Moreover, we did not limit our search for CVs to practice-level interventions, but included studies conducted at the patient or physician level. Therefore, we were able to report the CV estimates (but not the ICC estimates) for different patient populations, such as the chronically ill. Based on the ICC estimates we did obtain, it appears that ICCs do not systematically differ according to whether the study population included all patients or was limited to the chronically ill, but this should be confirmed using the study’s data.

### Ranges for Coefficient of Variation

Outcomes are more varied (and hence MDEs are bigger) when measured among all patients than among chronically ill patients. This occurs because all patients include both high-risk (or chronically ill) and low-risk patients, so outcomes (such as costs or hospitalizations) may range from zero to very high. As shown in Table E.1, costs are highly variable when measured for all patients, with CVs ranging from 2.17 to 5.17 within studies based on private payer (insurer) data for the general population. When the population is limited to those with chronic conditions or to the Medicare population, CVs become smaller. For example, for studies based on private payer data, CVs for patients with chronic conditions range from 1.8 to 2.74, depending on how narrowly *chronic condition* is defined. When the sample is limited to the general Medicare population, the average CV is about 2.46; the CV falls to about 1.85 when the Medicare population is further restricted to those with chronic illness.

CVs vary by outcome. The CVs for number of hospitalizations and number of emergency room visits are similar to those for costs, ranging from 2.43 to 6.0 for the privately insured general populations, and from about 2 to 3 when restricted to the chronically ill or Medicare populations. However, the CVs for number of hospital days tend to be so large (greater than 5 for all populations except the Medicare chronically ill), that it is unlikely that most studies will be able to reliably detect effects on the number of

hospital days. The maximum CV for a binary variable is 1, for a variable with a mean of 50 percent. Because study designers often adopt the conservative assumption that a binary variable will have a mean of 50 percent (and therefore a CV of 1), we did not systematically review the literature for the means of binary variables.

### **Range for ICCs**

Because ICCs vary by study, and because there are so few ICCs published in the literature, researchers are encouraged to use pre-intervention data from the practices in their sample to calculate ICCs for their planned analyses. (Sample code is in Appendix E.)

The degree of clustering appears to vary by outcome measure. Reported ICCs for health care costs are relatively low, ranging from 0.020 to 0.031 (Table E.3). Similarly, ICCs for health care service use (including number of emergency room visits and number of hospitalizations) range from 0.013 to 0.040. While ICCs for general satisfaction with health care tend to be low (ranging from 0.016 to 0.022), ICCs for specific satisfaction measures (access to care, coordination of care, etc.) tend to be much higher (ranging from 0.054 to 0.163). ICCs for quality-of-care process measures are also high (averaging 0.120) and have a wide range, from 0.058 to 0.25.

**Appendix Table E.1. Coefficients of variation (CVs) reported in the literature**

<b>Outcome Measure by Population Studied</b>	<b>CV</b>	<b>Studies Included</b>
<b>HEALTH CARE COSTS</b>		
All, Privately Insured	5.17	Ash et al. 2000
	2.17	Feldman and Parente 2010
	3.86	Zhao et al. 2005
Chronically Ill, Privately Insured	2.74	Philipson et al. 2010
	1.8 to 2.7	Ozminkowski et al. 2000 <sup>a</sup>
All, Medicare	2.79	Ash et al. 2000
	1.32	Counsell et al. 2009 <sup>b</sup>
	2.60	Pope et al. 2000
Chronically Ill, Medicare	1.86	Dale and Lundquist 2011 <sup>c</sup>
	1.47	McCall et al. 2010
	1.83	Philipson et al. 2010
	1.38	Peikes et al. 2011 <sup>c</sup>
<b>HOSPITALIZATIONS</b>		
All Privately Insured	3.5	Goetzel et al. 2010
	6.0	Short et al. 2009
Chronically Ill, Privately Insured	3.16	Goetzel et al. 2010 <sup>d</sup>
	3.19	Philipson et al. 2010
All Medicare	3.00	Counsell et al. 2007 <sup>a</sup>
Chronically Ill, Medicare	2.07	Dale and Lundquist 2011 <sup>c</sup>
	2.00	Philipson et al. 2010
	2.38	Leff et al. 2009
	1.49	Peikes et al. 2011 <sup>c</sup>
<b>HOSPITAL DAYS</b>		
All Insured	NA	
Chronically Ill, Privately Insured	5.80	Philipson et al. 2010
All, Medicare	5.78	Counsell et al. 2007 <sup>a</sup>
Chronically Ill, Medicare	2.92	Dale and Lundquist 2011
	2.93	Philipson et al. 2010
	3.37	Leff et al. 2009
<b>Outcome Measure by Population Studied</b>	<b>CV</b>	<b>Studies Included</b>



**Appendix Table E.1. Coefficients of variation (CVs) reported in the literature**

Outcome Measure by Population Studied	CV	Studies Included
EMERGENCY ROOM VISITS		
All, Privately Insured	2.43	Goetzel et al. 2010 <sup>d</sup>
	3.23	Short et al. 2009
Chronically Ill, Privately Insured	2.13	Goetzel et al. 2010
	2.28	Littenberg and MacLean 2006
All, Medicare	2.00	Counsell et al. 2007 <sup>a</sup>
Chronically Ill Patients, Medicare	2.77	Dale and Lundquist 2011
	2.73	Leff et al. 2009

<sup>a</sup>Reports CVs separately for a wide range of chronic conditions.

<sup>b</sup>Population is Medicare beneficiaries with incomes below 200 percent of the poverty line.

<sup>c</sup>CVs provided by study authors.

<sup>d</sup>The “obese” subgroup from this study is considered chronically ill for the purposes of this table.

NA = not available.

**Appendix Table E.2. Definitions of chronically ill used in selected studies**

<b>Study</b>	<b>Definition of Chronically Ill</b>
Dale and Lundquist 2011	Medicare beneficiaries with coronary artery disease, chronic heart failure, diabetes, Alzheimer's disease, or other mental, psychiatric, or neurological disorders; any chronic cardiac/circulatory disease, such as arteriosclerosis, myocardial infarction, or angina pectoris/stroke; any cancer; arthritis and osteoporosis; kidney disease; and lung disease, according to Medicare claims data
Goetzel et al. 2010	Obese adults (BMI greater than 30)
Leff et al. 2009	Medicare beneficiaries aged 65 or older in the top quartile of risk of using health services heavily during the following year (hierarchical condition category [HCC] score of 1.2 or higher)
Littenberg and MacLean 2006	Adults with diabetes
McCall et al. 2010	Medicare beneficiaries with HCC scores greater than or equal to 2.0 and annual costs of at least \$2,000 or HCC risk scores greater than or equal to 3.0 and a minimum of \$1,000 in annual medical costs (in 2005)
Ozminkowski et al. 2000	Individuals with malignant neoplasm, stroke, heart failure, psychiatric, diabetes, arthritis, seizures, COPD, asthma, or ulcerative colitis, according to private insurance claims data
Peikes et al. 2011	Medicare beneficiaries who met each program's eligibility conditions and also had either (1) CAD, CHF, or COPD and a hospitalization in the prior year, or (2) two or more hospitalizations in the prior 2 years
Philipson et al. 2010	Adults with heart disease

**Appendix Table E.3. Intracluster correlation coefficients (ICCs) reported in the literature**

<b>Outcome Measure</b>	<b>ICC</b>	<b>Study</b>	<b>Population</b>
Costs	0.021	Campbell MK et al. 2001	Patients with urology problems
	0.031	Dale and Lundquist 2011 <sup>a</sup>	Chronically ill Medicare
Hospitalizations	0.025	Dale and Lundquist 2011 <sup>a</sup>	Chronically ill Medicare
	0.014	Huang et al. 2003	Asthma patients with managed care
	0.030	Leff et al. 2009 <sup>a</sup>	High-risk Medicare
ER Visits	0.020	Dale and Lundquist 2011 <sup>a</sup>	Chronically ill Medicare
	0.040	Huang et al. 2003	Asthma patients with managed care
	0.013	Leff et al. 2009 <sup>a</sup>	High-risk Medicare
	0.015	Littenberg and MacLean 2006	Adults with diabetes
Satisfaction With Overall Care	0.022	Campbell SM et al. 2001	All patients
	0.019	Dale and Lundquist 2011 <sup>a</sup>	Chronically ill Medicare
	0.016	Potiriadis et al. 2008	All patients
Satisfaction With Access to Care	0.079	Campbell SM et al. 2001	All patients
	0.053	Dale and Lundquist 2011 <sup>a,b</sup>	Chronically ill Medicare
	0.163	Potiriadis et al. 2008	All patients
Quality-of-Care Process Measures	0.186	Campbell SM et al. 2001	All patients
	0.069	Dale and Lundquist 2011 <sup>a,c</sup>	Chronically ill Medicare
	0.058	Littenberg and MacLean 2006	Diabetes patients

<sup>a</sup>ICCs provided by study authors.

<sup>b</sup>Averaged the ICCs from 2 access measures related to access to care from this study.

<sup>c</sup>Averaged the ICCs from 4 diabetes process-of-care measures from this study.

## Appendix F

### Sample Code to Calculate the ICC

The ICC of a given outcome measure can be estimated using one of several computerized statistical packages that support Analysis of Variance (ANOVA) and/or General Linear Mixed Model (GLMM) commands.<sup>18</sup> Prior to analysis, data should be organized at the patient level so that each patient in the study sample has one record. That record should contain a practice identifier variable (called *practice\_id* in the code below), so that each patient outcome (called *outcome\_var* in the code) can be linked to the practice with which that patient was associated during the study. For illustrative purposes, sample SAS and Stata code used to estimate the ICC with the ANOVA method is provided below for a sample data set and outcome measure.<sup>19</sup>

### SAS Approach

#### Code

```
PROC ANOVA data = outcome_data;  
  CLASS practice_id;  
  MODEL outcome_var = practice_id;  
RUN;
```

---

<sup>18</sup> The PROC MIXED command can be used to estimate the ICC using the GLMM in SAS; the xtmixed command can be used to estimate the ICC using the GLMM in Stata. By default, both commands use Restricted Maximum Likelihood algorithms to estimate the between-practice and within-practice outcome variance components.

<sup>19</sup> Donner (1986) gives a thorough discussion of how to calculate the ICC using the means squared results from the ANOVA model.

## Output

The ANOVA Procedure					
Dependent Variable: <i>outcome_var</i>					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1187	8785.0977	7.4011	8.16	<.0001
Error	314844	285435.0576	0.9066		
Corrected Total	316031	294220.1552			
R-Square	Coeff Var	Root MSE	Mean		
0.029859	208.0339	0.952151	0.457690		
Source	DF	Anova SS	Mean Square	F Value	Pr > F
<i>practice_id</i>	1187	8785.097674	7.401093	8.16	<.0001

## Calculating the ICC

To estimate the ICC, first calculate the between-practice outcome variance ( $\sigma_B^2$ ) and the within-practice outcome variance ( $\sigma_W^2$ ) using the formulas below. The Mean Square of the Model ( $MS\alpha$ ) (shown as the first boxed number, 7.4011) and the Mean Square Error ( $MSE$ ) (shown as the second boxed number, 0.9066) are provided by the PROC ANOVA output in the SAS LST file. The denominator for the formula for the between-practice variance is equal to the average practice size ( $n$ ), or the total number of patients divided by the total number of practices.<sup>20</sup>

$$\sigma_W^2 = MSE \quad \sigma_B^2 = \frac{MS\alpha - MSE}{n}$$

$$\sigma_W^2 = 0.9066 \quad \sigma_B^2 = \frac{7.4011 - 0.9066}{265.64} = 0.02445$$

Once these two variance components have been calculated, the ICC can be estimated as the between-practice outcome variance divided by the sum of the between-practice and within-practice outcome variances.

$$ICC = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2}$$

$$ICC = \frac{0.02445}{0.02445 + 0.9066} = 0.02626$$

<sup>20</sup> If practices differ in the number of patients they serve, this formula should be used to calculate a weighted average practice size:

$$n = \frac{N - \sum_{i=1}^k \frac{n_i^2}{N}}{k-1}$$

Where:  $N$  = total number of patients

$n_i$  = number patients in practice  $i$

$k$  = number of practices

## Stata Approach

### Code

```
lone way outcome_var practice_id
```

### Output

One-way Analysis of Variance for *outcome\_var*

Source	SS	df	MS	F	Prob > F
Between <i>practice_id</i>	8785.0977	1187	7.4010932	8.16	0.0000
Within <i>practice_id</i>	285435.06	314844	.90659202		
Total	294220.16	316031	.93098511		

Number of obs = 316032

R-squared = 0.0299

Intraclass Asy.  
correlation S.E. [95% Conf. Interval]

0.02626 0.00181 0.02270 0.02982

Estimated SD of *practice\_id* effect .1563601

Estimated SD within *practice\_id* .9521513

Est. reliability of a *practice\_id* mean 0.87751

(evaluated at n=265.64)

### Calculating the ICC

The ICC can be read directly from the console output (0.02626, shown boxed above).





AHRQ Publication No. 11-0100-EF  
October 2011