White Paper

Methodological Considerations in Generating Provider Performance Scores for Use in Public Reporting

A Guide for Community Quality Collaboratives

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

Contract No. HHSA290200810037C

Prepared by:
Mark W. Friedberg, M.D., M.P.P., RAND Corporation
Cheryl L. Damberg, Ph.D., RAND Corporation

With assistance from:
Elizabeth A. McGlynn, Ph.D., RAND Corporation
John L. Adams, Ph.D., RAND Corporation

AHRQ Publication No. 11-0093
September 2011

Acknowledgments

The authors gratefully acknowledge the following individuals who generously contributed their time and expertise to the development and review of this report:

- Peggy McNamara, John Bott, and Katherine Crosson, Agency for Healthcare Research and Quality
- Christine Amy, Pennsylvania Aligning Forces for Quality—South Central Pennsylvania
- Jim Chase, Minnesota Community Management
- Nancy Clarke, formerly with the Oregon Health Care Quality Corporation;
- R. Adams Dudley, University of California-San Francisco
- Marc Elliott and Eric C. Schneider, RAND Corporation
- Roberta Esmond, Renee Frazier, and Jill Nault, Healthy Memphis Common Table
- Mary Gordon and Chris Queram, Wisconsin Collaborative for Healthcare Quality
- Judith Hibbard, University of Oregon
- Aparna Higgins, America's Health Insurance Plans
- Melinda Karp, Massachusetts Health Quality Partners
- Susan McDonald, formerly with the Minnesota Department of Human Services
- Ateev Mehrotra, RAND Corporation and the University of Pittsburgh
- Betsy Mulvey, New York Quality Alliance
- Carolyn Pare, Buyers Health Care Action Group (Minnesota)
- Devorah Rich, formerly with Greater Detroit Area Health Council
- Dana Richardson, Wisconsin Hospital Association
- Patrick Romano, University of California-Davis
- Natasha Rosenblatt, Puget Sound Health Alliance
- Dana Gelb Safran, Blue Cross Blue Shield of Massachusetts
- Dale Shaller, Shaller Consulting Group
- Mark Sonnenborn, Minnesota Hospital Association
- Ted van Glahn, Pacific Business Group on Health (California)

Friedberg MW, Damberg CL. Methodological considerations in generating provider performance scores for use in public reporting: a guide for community quality collaboratives. Rockville, MD: Agency for Healthcare Research and Quality; 2011. AHRQ Pub. No. 11-0093.

Foreword: Continued National Dialogue on Methodological Decisions in Generating Provider Performance Scores

For the past 3 years, the Agency for Healthcare Research and Quality (AHRQ) has provided technical assistance to 24 multistakeholder community quality collaboratives, which we refer to as Chartered Value Exchanges (CVEs). These CVEs support an agenda of quality transparency via public reporting of physician and hospital performance. At a recent national meeting of CVEs, community leaders expressed concern that currently two organizations could use the exact same dataset to produce a public report, but the two reports could produce and release diverging provider performance scores. Score differences could result from the way one or more measurement and data collection decisions are made.

To set the stage for continued regional and national dialogue, AHRQcommissioned RAND Corporation's Mark Friedberg and Cheryl Damberg to isolate and examine the set of decisions that collaboratives and other report sponsors face in the steps leading to the release of a public report.

These decisions, 20 in all, are grouped in this white paper in the following six categories:

1. Negotiating consensus on goals and value judgments of performance reporting.
2. Selecting measures that will be used to evaluate provider performance.
3. Identifying data sources and aggregating performance data.
4. Checking data quality and completeness.
5. Computing provider-level performance scores.
6. Creating performance reports.

For each decision, optional decision paths are laid out and the relative pros and cons of each are examined. RAND developed this paper in partnership with a panel of representatives from nine community quality collaboratives. These individuals provided local perspectives and real-world vignettes to illustrate optional paths for each decision.

Our overall aim in commissioning this paper was to produce a useful resource for community collaboratives and regional and national policymakers as together we build a network of quality transparency that supports consumer, provider, and payer engagement in quality and, ultimately, quality improvement. I hope this white paper informs ongoing deliberations, and I welcome your feedback (peggy.mcnamara@ahrq.hhs.gov).

I thank Mark Friedberg, Cheryl Damberg, and their colleagues at RAND Corporation for their comprehensive and scholarly approach and for their timeliness in developing this important resource for report sponsors across the country. I also thank the nine CVE representatives who shared their perspectives and provided case examples used in the paper and the impressive list of experts who provided critical feedback on an earlier draft.

This white paper is the latest in a series of coordinated efforts by AHRQ to support and enhance ongoing local and national dialogue related to data, quality measurement, and reporting. Visit www.ahrq.gov/qual/value/localnetworks.htm for the menu of related AHRQ resources.

Peggy McNamara
Senior Fellow
Agency for Healthcare Research and Quality
September 2011

Executive Summary

Public reports of health care providers' performance on measures of quality, cost and resource use, patient experience, and health outcomes have become increasingly common. These reports are often intended to help patients choose providers and may encourage providers to improve their performance.

At the July 2009 National Meeting of Chartered Value Exchanges (CVEs) hosted by AHRQ, CVE stakeholders identified a dilemma: Two organizations could, by making different methodological decisions, use the exact same data to produce divergent public performance reports that send conflicting messages to patients and providers. At the request of CVEs and in response to this dilemma, AHRQ commissioned RAND Corporation to develop a white paper to identify the key methodological decision points that precede publication of a performance report and to delineate the options for each. Our overall aim in developing this white paper is to produce a resource that is useful to CVEs and other community collaboratives as they consider the range of available methodological options for performance reporting.

Many methodological steps underlie the construction of provider performance scores for public reporting. These steps include data aggregation, measure selection, data validation, attribution of data to providers, categorization of providers by levels of performance, and assessment of the likelihood of misclassifying a provider's "true" performance. The purpose of this white paper is to review a number of the key methodological decision points CVEs and other community collaboratives may encounter when generating provider performance scores. The paper also discusses the advantages and disadvantages associated with various choices for each of these decision points. While the discussion focuses on analytic methods, there are rarely "right" answers. At each decision point, methodological considerations will be balanced by other stakeholder goals and values.

We recognize that CVEs and other community collaboratives may approach the process of developing provider performance reports in a variety of ways and may start at various points along the continuum of steps in constructing performance scores. Thus, while this paper can be read from front to back, it is written so that the reader can skip straight to any topic of interest.

In constructing provider performance reports for public reporting, a key concern, particularly among providers, is the possibility of generating performance scores that do not reflect the provider's "true" performance. In the lexicon of methodologists, this possibility is called the risk of "misclassifying" a provider (e.g., scoring a 4-star provider as a 1-star provider). Some degree of misclassification is always possible in any real-world report of provider performance. But the methodological decisions that a CVE makes can help to determine the frequency and magnitude of provider performance misclassification.

This report is intended to help CVEs understand different types of measurement error, how sources of error may enter into the construction of provider performance scores, and how to mitigate or minimize the risk of misclassifying a provider. Again, the methods decisions generally involve important tradeoffs. There are rarely clear "right answers," and value judgments underlie most decisions.

To illustrate some of the ways CVEs and other community collaboratives are approaching the methodological decision points discussed in this paper, we interviewed the leaders of nine such organizations. Quotes from these leaders are included throughout the paper, following many of the discussions about methods. The contents of these leadership interviews are also synthesized at the end of the paper in a section titled "Summary of methodological decisions made by a sample of CVE stakeholders."

Our report focuses on the steps involved in producing the comparative performance scores for public reporting. An equally important step and one that has a different set of methodological considerations (such as a report's understandability to consumers) is the design of provider performance "report cards." For guidance on the *design* of performance reports, we direct you to separate documents by Drs. Judith Hibbard and Shoshanna Sofaer that were sponsored by AHRQ as part of the "Best Practices in Public Reporting" series. *How To Effectively Present Health Care Performance Data to Consumers* and *Maximizing Consumer Understanding of Public Comparative Quality Reports: Effective Use of Explanatory Information*[1-2] are available online (www.ahrq.gov/qual/value/localnetworks.htm). AHRQ's "Talking Quality" Web site (www.talkingquality.ahrq.gov/default.aspx) and "Model Public Report Elements: A Sampler" (www.ahrq.gov/qual/perfmeasguide/perfmeaspt5.htm) also provide guidance on the design of performance reports.

Table of Contents

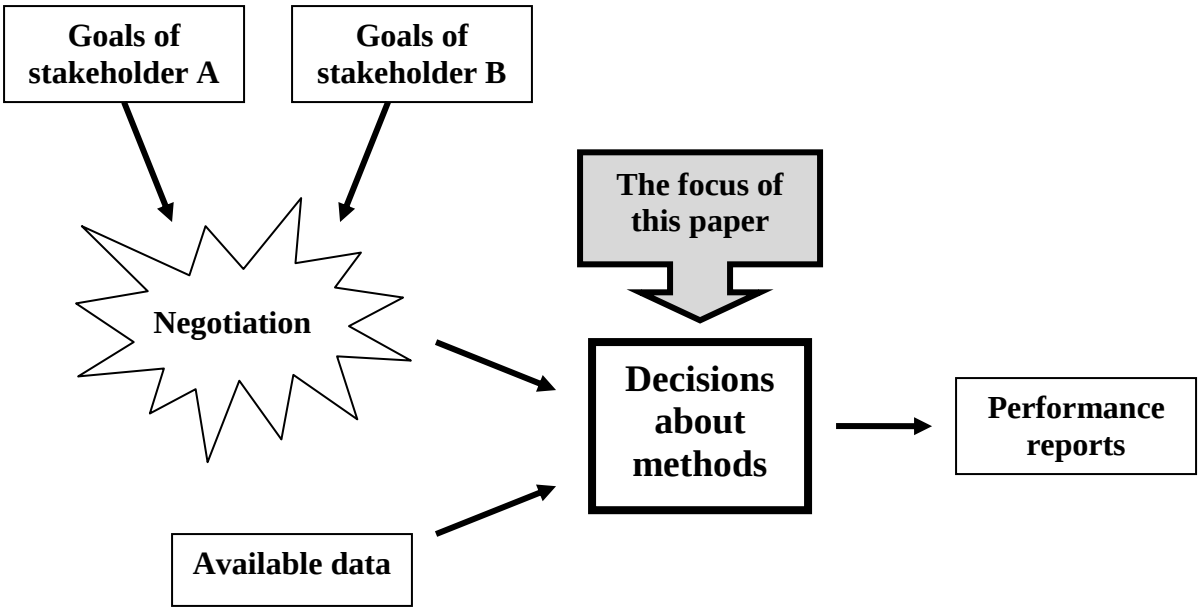
Figures and Tables

Introduction

This paper is intended for use by Chartered Value Exchanges (CVEs), community collaboratives, and other organizations interested in creating public reports on the performance of health care providers in their communities. This paper was written in response to a dilemma identified at the July 2009 National Meeting of CVEs hosted by AHRQ: Two organizations could use the exact same data to produce divergent public performance reports that send conflicting messages to patients and providers. At the request of CVE stakeholders, AHRQ commissioned RAND Corporation to develop a white paper to identify the key methodological decision points that precede publication of a performance report and to delineate the options for each. Our overall aim in developing this white paper is to produce a useful resource for CVEs and other collaboratives as they consider the range of available methodological options.

While decisions about methods are important, this paper also emphasizes the important roles that other factors (e.g., the goals of community stakeholders) play in determining how performance reports can be created. Figure 1 presents a simplified illustration of where methodological decisions fit into the overall process of creating performance reports.

Figure 1. The role of methodological decisions in creating performance reports



Moving from left to right, the figure shows that different stakeholders in a community (such as providers, patient advocates, and employers) may have different goals and priorities in creating a performance report. Some stakeholders may prioritize the inclusion of as many providers as possible in the report. Other stakeholders may prioritize the accuracy of reported performance. Negotiations between these stakeholders will ideally produce a set of criteria on which most CVE stakeholders can agree, and these criteria are likely to evolve as the CVE gains experience and more data become available. For example, these criteria might be (1) that the report should contain at least 75 percent of providers in the local area and (2) that no more than a small amount of provider performance misclassification should be present.

Guided by the results of negotiations between stakeholders and the available data, a CVE can begin to make decisions about methods for producing provider performance reports. These decisions, which are the focus of this paper, generally do not have "right answers" based on methodological criteria alone. Therefore, this paper is designed to help CVEs consider options at each methodological decision point and understand the advantages and disadvantages associated with these options. CVEs can reopen stakeholder negotiations and obtain new data if no methodological option produces a performance report that is satisfactory to all stakeholders.

This paper's lists of advantages and disadvantages of each option are unlikely to be exhaustive. Readers may think of new advantages and disadvantages for many of the options. Moreover, the relative importance of each advantage and disadvantage will probably differ among CVEs. Therefore, the lists of advantages and disadvantages in this paper should serve as *starting points* for discussion.

Types of Measures, Providers, and Data

The methodological considerations reviewed in this paper may apply to a wide variety of performance measures, including measures of quality, costs, patient experience, and health outcomes. These methodological considerations also may apply to reports that focus on different kinds of providers, including individual physicians and other practitioners, small practices, large provider groups, and hospitals. Some of the methodological considerations are most applicable to certain types of data (such as health plan claims) and less applicable to other types of data (such as patient surveys). Similarly, some methodological considerations matter more when the providers being measured serve relatively small patient populations (e.g., individual practitioners). When a particular methodological choice pertains mainly to one type of data or one type of provider, we identify these situations.

Definition of "Provider"

Throughout this document, the word "provider" is intended to be flexible in its meaning. "Providers" may refer to individual health care practitioners (physicians, nurses, therapists, pharmacists, etc.), practices or clinics (i.e., collections of practitioners who provide care together at a single address), or larger health care organizations (physician groups, hospitals, nursing homes, etc.). Throughout the paper, we mention particular types of providers as illustrative examples of larger methodological points. However, all of the issues discussed in this paper can apply to multiple types of providers.

How This Paper Is Organized

This paper begins with a discussion of performance misclassification, which is a fundamental, overarching methodological issue in any report of provider performance. Performance misclassification is defined and briefly discussed in the next section. Readers who are interested in more detailed information about the ways performance can be misclassified are encouraged to consult two appendixes to the report:

- Appendix 1: Systematic performance misclassification, and
- Appendix 2: Performance misclassification due to chance.

The remainder of this paper follows a series of six general steps that a CVE or other collaborative is likely to encounter when creating a performance report:

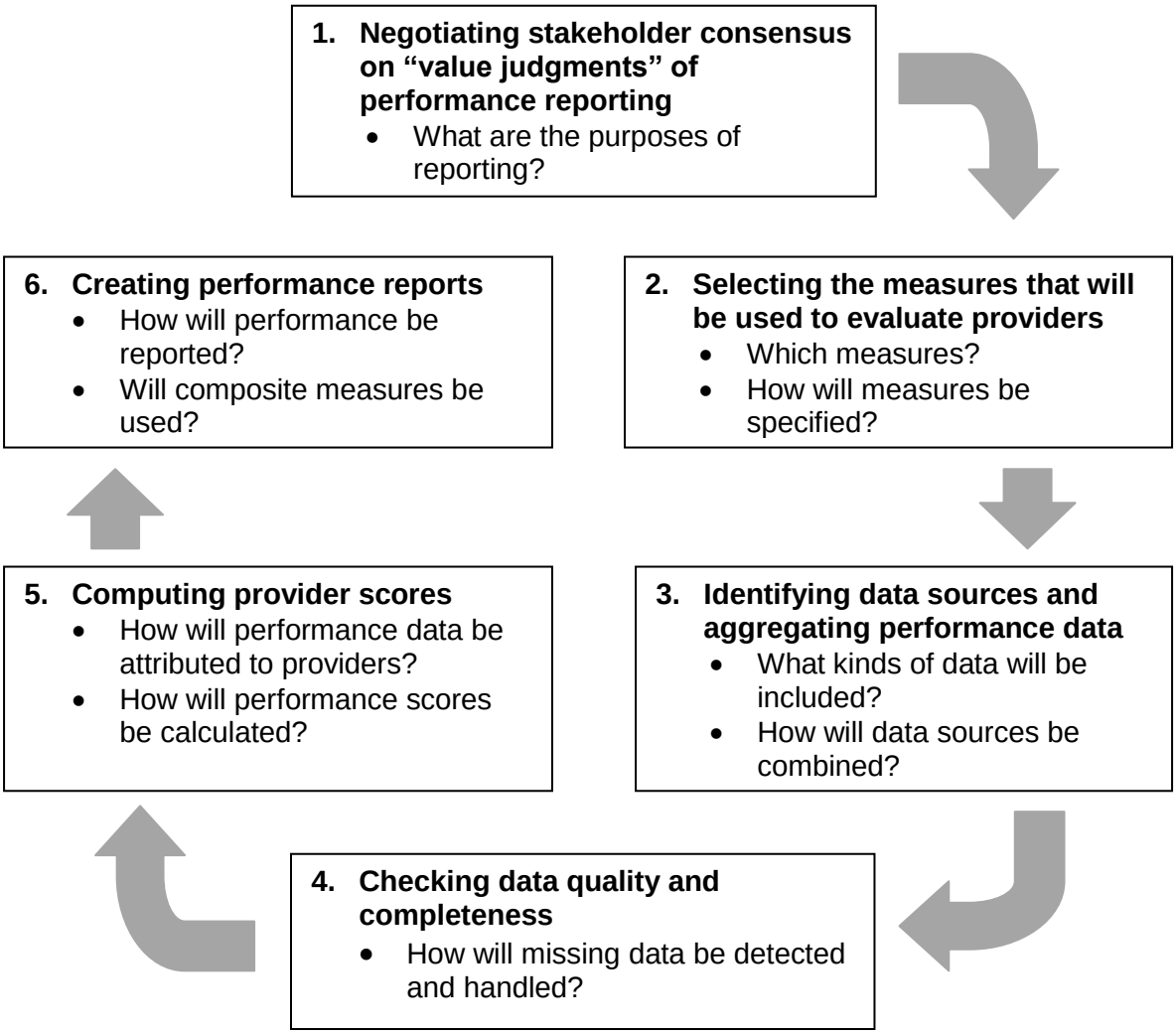
1. Negotiating stakeholder consensus on "value judgments" of performance reporting.
2. Selecting the measures that will be used to evaluate providers.
3. Identifying data sources and aggregating performance data.
4. Checking data quality and completeness.
5. Computing provider scores.
6. Creating performance reports.

Figure 2 shows this series of steps and some of the methodological decision points that may occur in each step. For each decision point, we present options. For most decision points, we also present examples of how a sample of CVEs (or stakeholder organizations) has chosen among the available methodological options. We focus on the reasoning behind these choices so that readers can get a sense of whether a given option may be preferable in their communities.

CVEs vary in the length of their reporting experience. Some CVEs have already produced multiple public reports and others have only recently begun to gather performance data for the first time. Therefore, this paper is organized so that it can be read from front to back, or readers can go directly to the section that pertains to a current report-making step. The full list of methodological questions addressed in this paper is available in the Table of Contents.

This paper concludes with a summary of methodological decisions made by a sample of stakeholders from nine CVEs. These nine CVEs may not be nationally representative. However, the choices and reasoning expressed by their stakeholders may be useful as a reference point for discussions about performance reporting methods.

Figure 2. Key operational tasks in generating performance reports



Overarching Methodological Issue: Performance Misclassification

A. What is performance misclassification?

The misclassification of provider performance is an overarching methodological issue in creating performance reports. Performance misclassification refers to reporting a provider's performance in a way that does not reflect the provider's *true* performance. For example, a report may contain three performance categories (e.g., bottom quartile, middle two quartiles, and top quartile), and for a given provider, performance may be reported as being in category 1 when true performance is in category 2.

Misclassification is a familiar concept in legal proceedings. Courts are imperfect: they sometimes convict the innocent and acquit the guilty. However, despite the presence of this misclassification, courts are generally believed to serve a useful social function.

Misclassifying providers is distinct from displaying performance results in a way that is difficult to understand and that confuses patients and providers. Even if a report is perfectly clear, and each patient and provider thoroughly understands its contents, performance misclassification can still lead to suboptimal results. For example, patients may go to truly low-performing providers, thinking they are high performing (as shown in Figure 3). Performance misclassification also may lead some low-performing providers to falsely believe that they have high performance, discouraging efforts to improve.

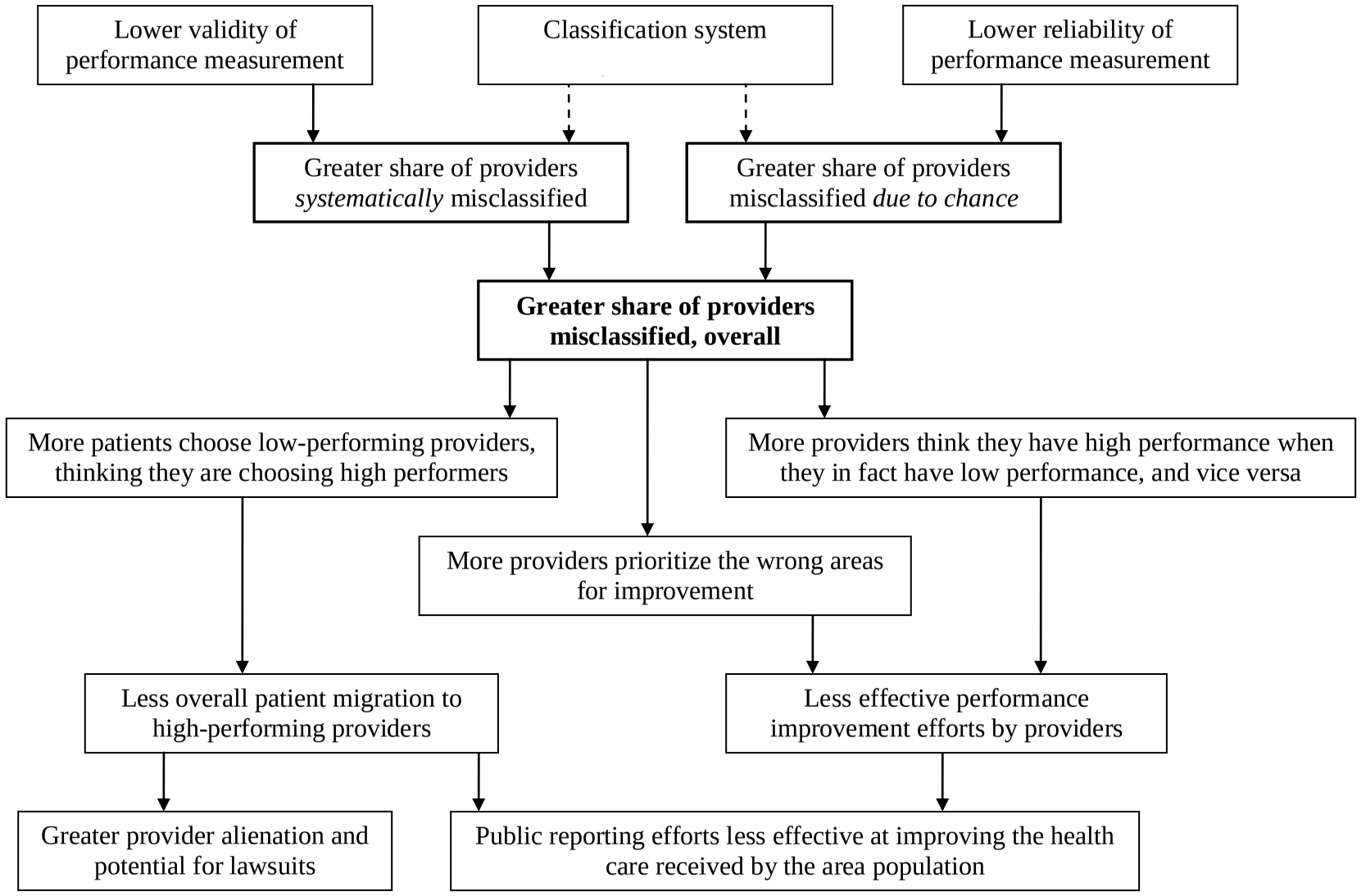
B. Why is performance misclassification important?

In a sense, all reports of provider performance classify the providers. For example, providers can be classified relative to each other or relative to a specified level of performance (e.g., above or below national average performance). Provider rankings are also a kind of classification system, since each rank is a class. Even reports that show performance scores and confidence intervals enable users to classify providers by comparing their performance. For example, report users might be able to see whether a provider's performance is different from the average performance. Alternatively, report users might just rank providers' performance scores, ignoring the confidence intervals.

Fundamentally, reports that misclassify the performance of too many providers (and misclassify them by too great an amount) may prevent the reports from having their best possible impact on health care received by a Chartered Value Exchange's (CVE) local patient population. For example, as shown in Figure 3, if greater shares of providers are misclassified, then more patients may choose low-performing providers, erroneously believing that they are high performing. Performance misclassification may even cause patients to leave high-performing providers, disrupting clinical relationships.

Higher rates of misclassification also will lead more providers to receive the wrong messages from performance reports. More low-performing providers may not attempt to improve, mistakenly believing that they are high performing. Providers also may prioritize the wrong areas for improvement, devoting scarce resources to areas in which they are truly doing fine and ignoring areas in which they could truly improve—again, because the report has misclassified their performance. Finally, high degrees of performance misclassification may threaten the stakeholder coalitions that are central to the success of a CVE.

Figure 3. Practical consequences of performance misclassification



C. What causes performance misclassification?

There are two general types of performance misclassification, and they have different causes. The first type is **systematic performance misclassification**. This kind of misclassification occurs when, for example, provider performance ratings are influenced by something beyond the provider's control (such as unusually high numbers of older patients). If providers are measured on the mortality rates of their patients, then the *measured* performance of providers with older patient populations will **systematically** look worse than their *true* performance. This is because older patients tend to have higher mortality rates than younger patients, all other things being equal. Most CVE stakeholders will agree that this kind of systematic performance misclassification is undesirable.

Baseball presents a useful analogy for thinking about systematic misclassification. Batters in an unusually competitive part of the league may face unusually skilled pitchers. If we only look at their batting averages, without paying attention to the pitchers they faced, the *measured* performance of these batters will systematically look *worse* than their *true* performance. Similarly, batters in less competitive parts of the league may face relatively unskilled pitchers (so getting a hit is relatively easy), and their *measured* performance will systematically look *better* than their *true* performance.

The second type of performance misclassification is **misclassification due to chance**. This kind of misclassification occurs because any time performance is measured, there will always be some amount of random measurement error. The unavoidable presence of measurement error means that for every provider in a report, any report that contains more than one category (i.e., a report that enables any kind of comparison between providers) will have some risk of misclassification due to chance.

Many CVE stakeholders may already have discussed misclassification due to chance without realizing it. Debates over "minimum sample sizes" (or how many patients need to be included in a performance measure before it can be reported) are an intuitive way to think about misclassification due to chance. Stakeholders would be right to wonder, "If a given provider has only had a handful of patients, how can we really know anything about the provider's true performance?"

Baseball also presents a useful analogy for thinking about misclassification due to chance. Suppose a rookie has an especially good first game, getting a hit in 3 of 4 times at bat. Based on this first game, his or her batting average will be 0.750. Are we then to assume this player is the greatest hitter of all time? If we only look at the batting average without paying attention to sample size (n = 4), we would have no other choice.

Most people will intuitively agree that induction into the Baseball Hall of Fame would be premature. Regardless of skill, the rookie was probably lucky in the first game. Baseball players' batting performance can vary dramatically from game to game, and the observed batting average of 0.750 far exceeds the full range of batting averages normally seen, even among great players. Therefore, classifying the rookie as "greatest ever" would run a very high risk of misclassification due to chance.

It is fair to ask: “How many times at bat would be needed before we know how good a baseball player really is?” In other words, how many observations are needed before we feel reasonably confident about predicting the player’s future performance? Thirty times at bat? One hundred times? A season? Multiple seasons? The best answer may depend on the purposes for which the performance data will be used and on the player’s calculated risk of performance misclassification.

If performance data will be used to decide whether to include a player in a team’s starting lineup for just a few games, then a relatively high risk of misclassification may be tolerable. On the other hand, if performance data will be used to offer the player a multiyear contract, then team managers may be willing to accept only a small risk of misclassification. When millions of dollars and multiple seasons are on the line, they will probably want as much certainty about future performance as possible.

The statistical issues in this baseball example are similar to the methodological issues facing CVEs, community collaboratives, and other organizations interested in creating public reports of performance of health care providers. Readers interested in more detailed information about the ways performance can be misclassified are encouraged to consult two appendixes to the report: Appendix 1: Systematic performance misclassification, and Appendix 2: Performance misclassification due to chance.

Decisions Encountered During Key Task #1: Negotiating Consensus on Goals and "Value Judgments" of Performance Reporting

Chartered Value Exchanges (CVEs) have multiple stakeholders, including patients, providers, health plans, employers, government agencies, and community groups. These stakeholders may have differing ideas and concerns about measuring and reporting provider performance. Because generating performance reports may require considerable time, effort, and financial resources, CVEs may find it beneficial to include all potential stakeholders in early and ongoing discussions concerning the "value judgments" of performance reporting. The value judgments will affect how stakeholders choose among the various options at each methodological decision point.

These value judgments are decisions for which there are no clearly right or wrong answers (or at least, no right or wrong answers from a methodological standpoint). Where possible, it is advisable to identify and address areas of disagreement among CVE stakeholders on these value judgments before resources are devoted to generating performance reports. By negotiating consensus among stakeholders early in the process and periodically revisiting this consensus, a CVE can establish good working relationships and approach problems in a neutral environment (i.e., an environment in which providers do not yet know their performance on a public report).

Examples: Negotiating consensus

Massachusetts Health Quality Partners (MHQP; www.mhqp.org) **brought providers into the reporting process early, years before reports were generated.** According to Melinda Karp, MHQP Director of Strategic Planning and Business Development, an important priority was to convince providers that MHQP's goal was to "do something *with* the providers, not do something *to* the providers." In addition to providers, health plans were brought to the table years before MHQP's first public reports were released.

The **California Cooperative Healthcare Reporting Initiative** (CCHRI), through the California Physician Performance Initiative (CPPI), is constructing individual physician performance scores on 17 measures of ambulatory quality. CCHRI has formed a Physician Advisory Group to review and provide input on an array of methods issues, including measure selection, attribution, and reliability of results for use by stakeholders. The CPPI project is adhering to the principles outlined in the Patient Charter for Physician Measurement (Consumer-Purchaser Disclosure Project, http://healthcaredisclosure.org/activities/charter), with the following negotiated criteria:

- Physicians must have an opportunity to correct their performance data.
- Performance reports must exceed a minimum reliability threshold (in order to limit the risk of misclassification due to chance).
- Performance must be reported in categories rather than as absolute values.
- Consumers must be given a way to understand the performance data and their limitations.

A. What are the purposes of publicly reporting provider performance?

This document is intended for use by CVEs interested in creating public reports of provider performance. These public reports may include measures of quality, costs (or efficiency), patient experience, or other types of performance measures. **Throughout this document, the word "provider" is intended to be flexible in its meaning.** "Providers" may refer to individual health care practitioners (physicians, nurses, therapists, pharmacists, etc.), practices or clinics (i.e., collections of practitioners who provide care together at a single address), or larger health care organizations (physician groups, hospitals, etc.).

Public reporting is not the only activity CVEs may undertake to improve health care in their local areas. CVEs also can engage in confidential reporting in which each provider's performance data are shared only with the provider. When the provider is an organization, this usually means sharing the data with organizational leaders, who may then decide whether and how to internally disseminate the data. This form of reporting can provide useful guidance to providers trying to improve their performance. For example, these providers may want to know how well their improvement initiatives are working.

Confidential reporting also can motivate providers to improve by appealing to a sense of professionalism. However, because confidential performance reports are not released to the public, they cannot be used by patients to select a provider. Therefore, a CVE's decision about whether to produce public or confidential performance reports (or produce both a public and a confidential report) may depend on the goals of CVE stakeholders.

If CVEs choose to publicly report provider performance, it may be advisable to reach early consensus on the purposes of these reports. This is a critical first step because the purposes of reporting will affect later methodological decision points. Based on Berwick[3] and Hibbard,[4] reporting has at least three general purposes:

- To help patients choose providers.
- To motivate performance improvement.
- To empower patients to act as "co-producers" of their health care.

Below, we discuss the advantages and caveats associated with these purposes of publicly reporting provider performance. **It is important to note that the potential purposes of public reporting are not mutually exclusive.** By using the same performance data in different ways, a CVE may be able to produce different reports to achieve different purposes.

Recognizing that different audiences may have different needs, a CVE could produce one report for patients and a second report for providers. For example, the kinds of performance reports that are most useful to patients may not be the most useful to providers seeking to improve (e.g., there may not be enough detail to provide guidance on improvement efforts).[3] Similarly, if reports are sufficiently detailed to guide providers' improvement efforts, they may be too detailed for many patients to easily understand.[5]

1. **Option 1: To help patients choose providers.** The goal of helping patients become better informed consumers of health care is a commonly cited reason for public performance reporting. If this option is chosen, then performance reports should be

designed with the patient in mind. They should be readily understandable to an audience that may not have medical or statistical expertise.[5] For guidance on which kinds of reporting formats might be preferable for helping patients choose providers, refer to papers by Drs. Hibbard and Sofaer[1-2] and to AHRQ's "Talking Quality" Web site (www.talkingquality.ahrq.gov/default.aspx) and "Model Public Report Elements: A Sampler" (www.ahrq.gov/qual/perfmeasguide/perfmeaspt5.htm).

Advantages:

- Patients may choose better performing providers.
- If providers believe patients are using public performance reports to make health care choices, providers may be motivated to improve.

Caveats:

- Historically, patients have not prioritized publicly available performance information when choosing a provider.[6-8] Anecdotal information from family and friends may be more heavily used by patients, even when performance data are available.
- Due to data limitations, it may not be possible to produce the performance reports that patients would find most useful or make them available at the right moment in the health care decisionmaking process. For example, a report of individual practitioner performance, rather than organizational performance, may have the best fit with how patients view their health care. However, publicly reporting the performance of individual practitioners may not be possible, especially when a CVE also wants to limit the amount of performance misclassification due to chance.

2. **Option 2: To motivate providers to improve.** Enabling patients to choose providers based on their performance may motivate improvement efforts. If providers believe patients use public reports, then providers who want to attract and keep patients will be motivated to attain high performance. However, even if providers do not believe patients use public reports when seeking health care, these reports can have a powerful motivating effect. Providers may want to do well—out of a sense of professionalism, competition, or "peer pressure"—in the eyes of their colleagues, other health care organizations, and the general public.[9] In addition, performance reports that present detailed performance information can help guide providers in their improvement efforts (e.g., by showing them exactly which measures need the most improvement).[3]

   Advantages:

   - Providers may improve their performance.
   - Providers may get guidance in their improvement efforts, especially when reports give detailed performance information.

   Caveats:

   - Some providers with poor performance may criticize the report rather than engage in improvement efforts.[6, 10]

- There is some evidence that publicly reporting performance may not always spur performance improvement.[11]

3. **Option 3: To empower patients to "co-produce" their health care.** Patients who are empowered to be more active participants in their own health care may have better outcomes of care.[4] Public reports of provider performance may raise patients' awareness that there is substantial variation in performance on important measures of health care quality. Regardless of whether they use performance information to choose a provider, patients may be motivated to ask for the health care services included in performance reports (especially if they note that their own provider's performance is not perfect). As with reports aimed at informing patients' choice of provider, reports aimed at empowering patients should be understandable by (and educational for) those who may not have medical and statistical expertise.

   Advantages:

   - Empowered patients may receive better care.

   Caveat:

   - Patient empowerment may not require public performance reporting. Other means of patient education may be more efficient.

Thoughts on the purposes of public reporting

- Nancy Clarke, formerly Executive Director of the **Oregon Health Care Quality Corporation** (q-corp.org), describes the organization's main purposes in public reporting as motivating quality improvement and making the patient a partner in quality improvement. Due to a shortage of primary care providers (PCPs), the "shopping model for consumers driving markets doesn't have much traction [in Oregon]." These thoughts were echoed by Christine Amy, Project Director of **Aligning Forces for Quality-South Central Pennsylvania** (www.aligning4healthpa.org): "There aren't enough PCPs in the area, so labeling a provider as 'great' isn't relevant to patient choice when the provider is closed to new patients. The purposes of public reporting are to motivate and guide providers and to use the reports as a teaching tool to help patients be better partners in their own care."
- Devorah Rich, formerly Project Director of the **Greater Detroit Area Health Council** (www.gdahc.org), describes an evolution in the purposes of reporting: "Ideally, people originally thought it would engage the consumer, but it's turned out to actually motivate the physicians very powerfully. The physicians pay a lot of attention to our reports. You don't get through medical school without being competitive." There has been less evidence of consumer engagement with the reports, and this is felt to be due to reporting performance at the physician organization level (rather than the individual physician level).
- Renee Frazier, Executive Director of the **Healthy Memphis Common Table** (www.healthymemphis.org) explains that the main purposes of public reporting are to motivate provider improvement and to empower patients: "Knowing the indicators (and the reasons for them) helps individuals to understand the most important care they should be receiving. It also helps to know what services to ask for if you are not already receiving them from your doctor."

- Jim Chase, Executive Director of **Minnesota Community Measurement** (www.mnhealthscores.org; a member of the Minnesota CVE) notes that while performance reporting has mostly motivated providers to improve, rather than guiding patients' choice of provider, reporting has been tied to explicit incentives aimed at providers. The performance scores in public reports also have served as the basis for pay-for-performance and provider tiering programs: "We've learned that [patients and providers] don't just go out and use the information. There's an evolution, and incentives like pay-for-performance and tiering can make the information more relevant."
- Susan McDonald, formerly with the **Minnesota Department of Human Services,** credits public purchasers' use of the performance reports with catalyzing provider improvement efforts, and Carolyn Pare, President and Chief Executive Officer of the **Buyers Health Care Action Group** (also a member of the Minnesota CVE) further notes the crucial roles played by purchasers and quality improvement organizations in helping providers make the best use of performance reports: "While critically important, standard measurement, data collection, and reporting in and of itself would not have changed things in Minnesota."

B. What will be the general format of performance reports?

Performance reports can vary widely in their general formats. They can be complex, with detailed reports of measure-by-measure performance rates and statistical confidence intervals, or they can be much simpler, displaying categories of overall performance on a composite measure (e.g., "a 3-star hospital on pneumonia"). For CVEs, it may be advisable to negotiate the general format (or formats, if multiple reports are planned) *before* providers know exactly how their own performance will appear. At this stage, a scan of existing reports (including reports that are on paper and on the Internet) may be useful to help stimulate discussion.

The decision about which general format to use will probably be heavily influenced by the purpose of public reporting. In general, reports that are aimed at a patient audience will need to have a simpler reporting format that is more usable by this audience.[i] Such reports may present only a few categories of performance (e.g., a 4-star scale) or may rank providers to enable quick ascertainment of the highest and lowest performers. In addition, **reports that are based on *relative* provider performance may be most informative to patients who are trying to choose the highest performing providers**. Reports of relative provider performance focus on enabling comparisons between providers within a given market area (i.e., the market area theoretically accessible to the patient), rather than comparing providers to an external performance threshold, such as a national benchmark.

On the other hand, reports that are aimed at a provider audience (to motivate and guide improvement) may require more reporting formats that display more detailed information. Relative performance may be presented in such reports to enhance their ability to motivate improvement, but absolute performance (with numerators, denominators, and other "raw scores") is likely to be most useful in guiding improvement efforts.

[i] See reports by Hibbard and Sofaer for more detailed guidance on which kinds of reporting formats might be preferable for purposes such as helping patients choose their providers.[1-2]

Many options and combinations of options are available for the general format of performance reports. **Each reporting format may be more appropriate for some audiences and less appropriate for others.** We present three examples here.

1. **Option 1: Simplified reports of relative provider performance.** This option is attractive when the purpose of reporting is to inform patients' choices of health care providers. These reports generally present only a few categories of performance, measures are aggregated when possible, and providers may be ranked. Raw performance rates and scientific depictions of statistical uncertainty are rarely included in such reports.

   Advantages:

   - Enables patients (who may lack medical or statistical expertise) to more easily interpret performance differences among providers.

   Disadvantages:

   - May oversimplify the full range of provider performance. For example, the "1-star" category for provider performance may include a wide range of actual performance levels.
   - May obscure the representation of statistical uncertainty. If patients do not understand the degree of statistical uncertainty in a performance report, small differences in performance may be interpreted as meaningful when in truth they are not.
   - May not contain enough detailed information to guide provider improvement efforts.

Option 1 Examples: Simplified reports of relative provider performance

The **Oregon Health Care Quality Corporation** (q-corp.org), which received guidance from a "consumer plain language" expert, reports clinic performance in three categories: "better" (clinic absolute score is higher than one standard deviation above the statewide score), "average," and "below" (clinic absolute score is lower than one standard deviation below the statewide score). In addition, the CVE confidentially provides detailed performance data to each clinic.

The **Puget Sound Health Alliance** (www.wacommunitycheckup.org) reports provider performance in three categories: above regional average, at regional average, and below regional average. However, users can select a provider's name in the Web-based report to access numeric performance scores and statistical confidence intervals.

The **Healthy Memphis Common Table** (www.healthymemphis.org) reports provider performance using a star system: providers get 1 star for performance that exceeds the 75th percentile in Shelby County and 2 stars for performance exceeding the 90th percentile. This reporting format was felt to be consistent with the literacy level of the patient community (i.e., consistent with a fifth grade level of literacy).

2. **Option 2: Simplified reports of absolute provider performance.** Rather than showing how providers compare with each other, performance reports can show simplified categories of absolute performance. For example, if the range of possible scores on a performance measure is 0-100, such a report could tell patients whether a given provider scored above 80 or below 80 (regardless of how many providers score above or below 80). This approach is attractive when CVE stakeholders can agree on an absolute performance threshold above (or below) which there are no truly meaningful differences in performance.

   Advantages:

   - Enables patients (who may lack medical or statistical expertise) to understand performance information when choosing providers.
   - May set clear performance goals for providers. By comparing their absolute current scores to the performance thresholds that define the reported performance categories, providers can gauge how much they need to improve to get into a higher category.

   Disadvantages:

   - The representation of statistical uncertainty may be challenging.
   - Reports may not contain enough detailed information to guide provider improvement efforts.

   Caveat:

   - If all providers in a CVE's area are in the same performance category, then the report will not be useful in choosing a provider. This is not necessarily a bad thing. If, for example, all providers score in the highest category, then patients can choose providers on attributes such as convenience and be reasonably confident that they will get high-performing providers.

3. **Option 3: Detailed reports of absolute provider performance.** This option is attractive when the purpose of reporting is to guide providers' efforts to improve performance. These reports may present data that are as detailed as possible as well as data that are somewhat more aggregated (to enable providers to prioritize their efforts). These reports also may contain explicit improvement strategies and identify high-performing providers who can share best practices.

   Advantages:

   - Reports may give providers useful guidance in their improvement efforts.

   Disadvantages:

   - Data complexity may make these reports less accessible to patients who are trying to choose a provider.

Option 3 Examples: Detailed reports of absolute provider performance

Organizations leading the **Minnesota Healthcare Value Exchange** (www.mnhealthscores.org/ and www.mnhospitalquality.org/) report numeric performance scores for each provider in its reports. These reports display the providers in the rank-order of their scores. There is no representation of statistical uncertainty in the public reports. However, providers receive even more detailed reports of their own scores with statistical confidence intervals.

Aligning Forces for Quality-South Central Pennsylvania (www.aligning4healthpa.org) displays provider performance on each measure of diabetes care quality as an absolute percentage, with national and community average scores included as benchmarks. These performance scores are initially sorted according to provider name (in alphabetical order), but providers also can be sorted by performance rank (with a single user action). Currently, no representation of statistical uncertainty is included in these performance reports.

The **Wisconsin Healthcare Value Exchange** generally reports absolute performance scores, consistent with the primary purpose of enabling provider groups to compare their performance to benchmarks. The ambulatory Web site (www.wchq.org) is "not really designed for consumers," and Web site user tracking statistics confirm that the site is most often visited by Wisconsin health care providers.

C. What will be the acceptable level of performance misclassification due to chance?

It is impossible to know exactly which providers are misclassified due to chance alone. However, it *is* possible to know, for each provider, the *risk* (i.e.,probability) that performance is misclassified. CVE stakeholders can therefore negotiate a maximum acceptable risk of performance misclassification due to chance, and this negotiation can take place before performance reports are created (i.e., before providers know exactly how their performance will appear). This negotiation can be more useful and concrete if there is general agreement about the format of a performance report.

For example, if a CVE has provisionally decided on a 4-star scale for reporting, stakeholders can address such questions as:

- What is the maximum acceptable risk that a *true* 4-star provider will be misclassified as a 3-star provider? What about being misclassified as a 2-star provider?
- What is the maximum acceptable risk that a *true* 2-star provider will be misclassified as a 3-star provider? Or a 4-star provider? Or a 1-star provider?

There is no "right answer" to the acceptable risk of misclassification due to chance. How much risk is acceptable may vary by CVE, depending on exactly which measures will be reported and on how performance reports will be used. Patients have a wide range of opinions about the acceptable level of misclassification risk. In a 2006 survey, most patients thought a risk of misclassification greater than 5 percent but not greater than 20 percent would be acceptable.[12]

Other CVE stakeholders may have different opinions about how much misclassification they think is reasonable in a performance report. The important thing is to engage CVE stakeholders in discussions about misclassification, to acknowledge its existence as a limitation of any performance report, and to begin to achieve consensus on how much misclassification risk is acceptable. This level of risk always can be revisited at later stages (especially, once more is known about how the factors that determine misclassification interact with each other in a given performance report; Appendix A discusses this issue further).

To decide on an acceptable amount of misclassification due to chance, CVE stakeholders may want to think about the goals of performance reporting:

- If the goal of the performance report is to help patients choose higher performing providers, reports that have too high a rate of misclassification can mislead too many patients.
- If the goal of the performance report is to motivate providers to improve, an excessive rate of misclassification will falsely reassure too many low-performing providers who are misclassified as high performing. It also can generate concern among high-performing providers who are classified as low performers.
- If the goal of the performance report is to reward high performance, an excessive rate of misclassification will result in too many low performers being rewarded and too many high performers not receiving a reward.

The acceptable risk of performance misclassification due to chance can take many values. We present two polar extremes to illustrate the tradeoffs.

1. **"Extreme" Option 1: Set a very low level of acceptable misclassification risk due to chance.** An example of a very low level of risk is "less than 1% of all true 4-star providers will be misclassified as 3-star providers, and less than 0.1% will be misclassified as 2-star providers." An example of a current report that uses statistical confidence intervals to limit the risk of misclassifying average performers as above or below average is the Hospital Compare report of hospitals' 30-day mortality rates (www.hospitalcompare.hhs.gov).[13] For the vast majority of hospitals, Hospital Compare classifies their mortality performance as average ("No different than the U.S. national rate").

   Advantages:

   - The risk that a provider's performance will be misclassified due to chance will be low.
   - If statistical confidence intervals are used to set a low level of misclassification risk, then the probability of one type of misclassification (classifying providers as below or above average when they truly have average performance) will be limited to the level of confidence (usually 5%). Using confidence intervals to limit misclassification risk is discussed in more detail in the section on Task #5.

Disadvantages:

- When sample sizes are small (or when between-provider differences in true performance are minimal), it may not be possible to include a large proportion of providers in the report. Or it may not be possible to report performance on measures that are important to stakeholders. These problems are especially likely when reporting the performance of individual clinicians.
- If statistical confidence intervals are used to set a low level of misclassification risk, then nearly all providers may be classified as having average performance. Therefore, there will be a higher risk of misclassifying truly above or below average providers as average performers.

2. **"Extreme" Option 2: Set a very high level of acceptable misclassification risk due to chance.** An example of a very high level of risk is"up to 40% of all true 4-star providers will be misclassified as 3-star providers."

   Advantages:

   - Even when sample sizes are not large, it may be possible to report the performance of nearly all providers on nearly all measures.

   Disadvantages:

   - The performance report may misclassify the performance of many providers on many measures solely due to chance. The potential consequences of this performance misclassification are shown in Figure 3.

Example: Talking with stakeholders about misclassification risk

Even though misclassification risk is a fundamental and important methodological issue, more tangible approaches to discussing the subject may help engage stakeholders. In interviewing CVE stakeholders, we found that most do not currently engage stakeholders in conversations that are explicitly about misclassification risk. Instead, CVEs discuss more "tangible" topics that are fundamentally about misclassification risk…without actually mentioning the words "misclassification risk." *These discussions may combine the statistical theory-based concerns outlined in this report with the political realities in which each CVE operates.*

As Nancy Clarke, formerly Executive Director of the Oregon Health Care Quality Corporation (q-corp.org) explains:

> If we held a meeting on "risk of misclassification," no one would come. But when we have meetings on "tradeoffs: what's fair to providers and fair to consumers," plenty of people come. We had sequential meetings, each with a white paper that combined the statistical and the political: "What's big enough for clinic size?" "What's big enough for number of cases?" "How do we put data into buckets to show the public?" "What's a fair benchmark?" etc. EVERYBODY comes to those meetings.

Decisions Encountered During Key Task #2: Selecting the Measures That Will Be Used To Evaluate Provider Performance

A. Which measures will be included in a performance report?

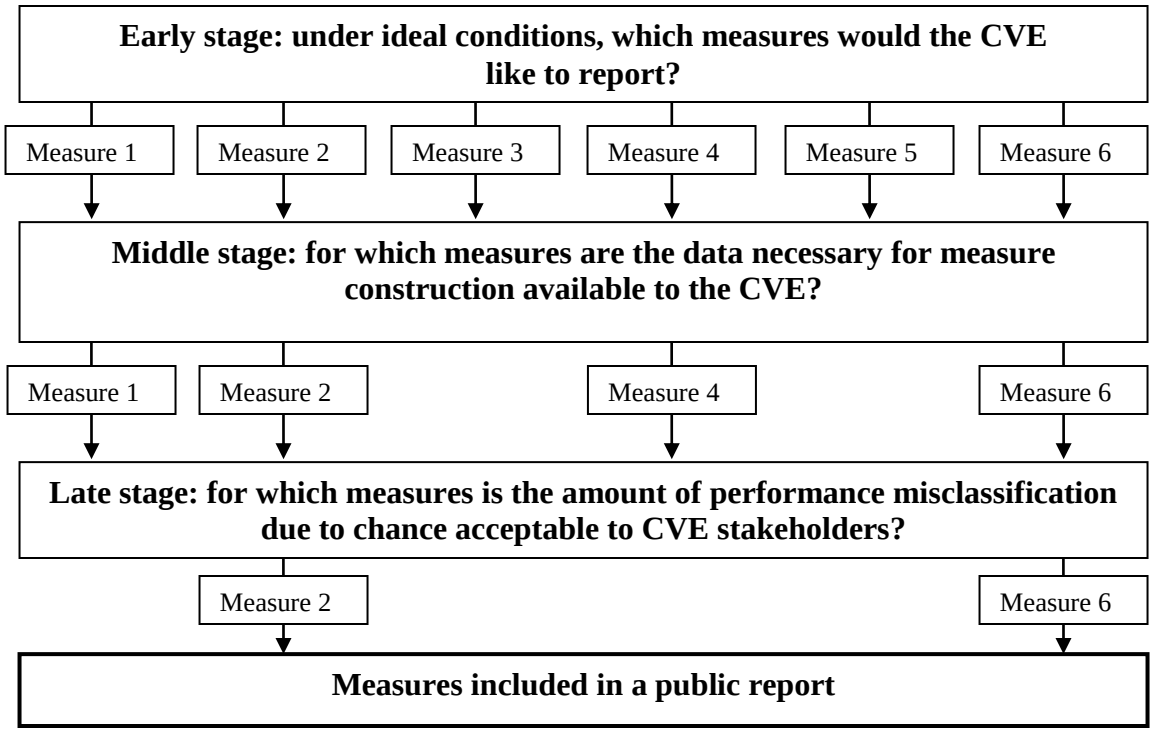
There are many things a Charted Value Exchange (CVE) may want to consider when choosing performance measures for a report. For the purposes of this paper, we divide the measure selection process into an "early stage," a "middle stage," and a "late stage."

- **The early stage** occurs before a CVE knows exactly what kinds of performance data are available (Key Task 3) and before the quality and completeness of these data are known (Key Task 4). In the early stage of measure selection, a CVE may want to be as inclusive as possible: *Under ideal conditions, what performance measures would CVE stakeholders like to report?* A separate AHRQ publication titled *Selecting Quality and Resource Use Measures: A Decision Guide for Community Quality Collaboratives*, provides a broader and complementary discussion of how CVEs might engage in early-stage quality measure selection.[14] The process outlined in this complementary decision guide is intended to assist CVEs in choosing measures with good intrinsic properties (i.e., those measures that cover the desired domains of performance and meet standards of importance, scientific acceptability, and usability). Once these early-stage standards have been met, a CVE can proceed to the middle and late stages, which depend on how *local* factors interact with the intrinsic characteristics of the measures.
- **In the middle stage** of measure selection, a CVE may discover *local* data limitations that prevent the construction of certain performance measures. In the decision points of Key Tasks 3 and 4, we provide some options for addressing these data problems. However, data problems may still make it impossible to report all of the measures identified in the early stage of measure selection. Some measures may need to be set aside at this point. This middle stage of measure selection will be easier if CVE stakeholders can reach an earlier consensus on criteria for setting measures aside.
- **In the late stage** of measure selection, the remaining measures can be calculated for a "mockup" performance report, and the *local* risk of misclassification due to chance can be calculated. For a detailed discussion of misclassification risk, refer to Appendixes 1 and 2. Even though a CVE can find many ways to limit the risk of misclassifying provider performance, some measures may not be publicly reported due to excessive misclassification risk for a large number of providers. When this is the case, additional measures may need to be set aside.

Figure 4 illustrates the way a measure selection process might occur. In a way, the middle and late stages of measure selection are "filters" on the early stage. Of course, this way of thinking about measure selection is simplified and does not include every factor that might influence which measures make it into a performance report. But from a purely methodological point of view, this approach illustrates some of the major considerations.

A key take-away point is that a measure may satisfy **general requirements,** but **local factors** particular to each CVE are also crucial determinants of which measures may be included in a performance report. (Examples of general requirements include the National Quality Forum's criteria of importance, scientific acceptability, usability, and feasibility.[15] Examples of local factors include negotiated "value judgments," data availability, and misclassification risk in the local provider community.) Whether performance measures are right for reporting in "your backyard" can only be determined in later stages of the measure selection process.

Figure 4. Illustration of a measure selection process



In the early stage of measure selection, CVEs may want to choose a broad set of measures that:

- Measure care for conditions that are common in the population,
- Measure care for conditions that are important to members of the population,
- Measure outcomes of care, such as patient mortality,
- Measure costs, utilization, or efficiency of care,
- Measure processes of care, such as checking cholesterol in patients with diabetes,
- Measure patient experience or patient satisfaction,
- Measure coordination of care,
- Are known to have room for performance improvement,
- Come with "prepackaged" risk-adjustment methods,
- Are relatively easy for the target audience to understand,
- Are part of consensus-based measure sets, such as being certified by the National Quality Forum, or
- Are part of a local, regional, or national improvement effort.[16]

Other considerations may come into play when choosing measures in the early stage, and lists of existing measures may help and provide inspiration. Additional guidance on making initial choices about performance measures is available from a variety of sources.[15, 17-20]

Example: Choosing performance measures

The **Wisconsin Healthcare Value Exchange** ambulatory performance report (www.wchq.org) initially started with diabetes quality of care measures because this health condition was a high public health priority in Wisconsin. Measures of preventive care quality were added next, because these measures were "in demand." The selection of ambulatory measures has been "organic" over the years, without a formal, standardized process. Existing measures, national measurement trends, and member preferences have been taken into consideration in measure selection.

For hospitals (www.wicheckpoint.org), however, measures were chosen in a well-defined selection process, guided by a board of directors and steering committee. The workgroup members think "big picture," starting with ideas about what people would want to improve. They then proceed through 20 criteria, such as:

- Are there existing measures?
- Is there evidence that the measure being considered actually reflects the clinical practice that we're trying to improve?
- How does the measure align with national priorities?
- How does the measure align with State priorities?

B. How will the performance measures be specified?

Once performance measures are agreed on in the early stage of selection, the next step is to decide exactly how each measure will be specified. Here, specification refers to the exact ways raw data about patient care (e.g., data that come directly from administrative sources or medical records) are used to construct performance measures. For example, a measure may consist of a numerator and a denominator. The numerator measures the number of times a clinical service (e.g., an immunization) *is* provided, while the denominator measures the number of times a clinical service *should be* provided (e.g., the number of patients who should be immunized in a given year).

The measure specifications are the criteria for determining which patients are eligible for the service and which clinical services are received by these patients. Eligible patients are counted in the denominator, and services received are counted in the numerator. The specifications of a measure are the "DNA" of a measure, and small changes in specifications can have large effects on a performance report.

Whether a CVE can develop its own measure specifications depends on how "raw" the available performance data are. If these data are claims, or unprocessed clinical data or survey responses, then a CVE could construct performance measures according to the CVE's own specifications

(e.g., the CVE can decide which patients count toward the denominator and which services count toward the numerator). However, if a third party already has constructed "prescored" performance scores (e.g., Leapfrog measures), then the task of specification already has been performed by the body that constructed the measure.

Here are some options a CVE may consider in deciding how a performance measure will be specified. This is not an exhaustive list, but it illustrates the pros and cons of some commonly available options.

1. **Option 1: Use measure specifications that are endorsed by national bodies.** Many nationally endorsed performance measures, such as those developed by the National Committee on Quality Assurance and AHRQ, come with detailed specifications. These measures also may have national performance benchmarks. Altering the specifications of these measures may invalidate comparisons to these benchmarks.

   In addition, some nationally endorsed measure specifications may come with established methodologies to adjust for differences in case mix among providers. Case mix adjustment is discussed in more detail in the section on Task #5. In a nutshell, case mix adjustment refers to statistical techniques that are intended to ensure that performance comparisons do not systematically misclassify providers. In other words, case mix adjustment seeks to avoid "comparing apples to oranges."

   Advantages:

   - May allow valid comparisons to national benchmarks and to performance scores reported by other reporting programs.
   - May already have case mix adjustment methodologies developed.

   Disadvantages:

   - Nationally endorsed measures may not optimally address a CVE's local priorities.
   - Nationally endorsed measures may not be usable "off-the-shelf," since all data elements may not be available for their construction. It is common for local collaboratives to slightly modify national specifications (e.g., using data that *are* available to identify patients for the denominator).
   - When the scientific evidence behind a measure changes, nationally endorsed measure specifications may be slower to incorporate the new scientific evidence than CVE stakeholders would like.

2. **Option 2: Use locally modified measure specifications.** When CVEs examine the specifications of existing performance measures, stakeholders may want to consider modifying these specifications. Reasons to modify these specifications may include wanting to take advantage of data that are available locally but are rarely available nationally, such as data about a clinically important comorbidity. Also, data elements that are included in the national specifications may not be available locally, precluding precise adherence to the national specifications. However, modifying measure

specifications involves important tradeoffs. Comparisons to national benchmarks may not be valid, and new case mix adjustment methodologies may need to be developed.

In general, performance measurement experts advise against modifying nationally endorsed measure specifications unless modification is unavoidable. As an alternative to modifying measure specifications in its own performance reports, a CVE may choose to convey ideas for measure modification to the measure developer. The goal is to improve subsequent revisions of the nationally endorsed measure.

Advantages:

- Modified measures may better address a CVE's local priorities.

Disadvantages:

- Valid comparisons to national benchmarks probably will not be possible.
- Deviating from nationally endorsed specifications may open the way for constant negotiation over further changes.
- New case mix adjustment methodologies will need to be developed, requiring the assistance of a statistician with expertise in performance measurement (see the section on Task #5).

3. **Option 3: Use measure specifications that are included in proprietary software packages.** Proprietary software packages are available that compute measures of provider performance using locally obtained data (usually administrative data such as health plan claims). Some software packages are widely used, so performance comparisons to external benchmarks may be possible. The software packages already may incorporate case mix adjustment methodologies. However, because these software packages are proprietary, it may not be possible to know exactly how the measures that they generate are specified. Information about how the performance measures are calculated may be available from the software vendors.

   Using proprietary software to construct performance measures also may raise concerns about systematic performance misclassification and the risk of misclassification due to chance. Case mix adjustment methods included in the software may be inadequate, allowing systematic performance misclassification. The software may not generate performance data with enough detail to calculate the risk of performance misclassification for each provider (i.e., enough detail to calculate within-provider measurement error; see Appendix 2 for a more detailed discussion). In that case, it may be impossible to know whether the overall risk of misclassification is acceptable to CVE stakeholders. This could happen if the software provides each provider's score on a performance measure without indicating how much uncertainty there is about that score.

   Advantages:

   - May allow comparisons to external benchmarks.
   - May already incorporate case mix adjustment methodologies.

- Relatively easy to use.
- No need for measure development.

Disadvantages:

- Proprietary software packages may function like "black boxes" that turn raw performance data into performance scores. In other words, such packages may not reveal detailed measure specifications to CVEs. This lack of transparency can make it difficult to really understand what is being reported, undermining stakeholder trust (especially among providers). Moreover, it may be impossible to assess the construct validity of "black box" measures—one of the most basic requirements of a valid performance report. Construct validity is discussed in Appendix 1.
- If performance data are not generated with the right level of detail, assessing the risk of misclassification due to chance may be difficult.
- CVEs may need to check the performance of the case mix methodologies included in the software. To detect systematic performance misclassification due to inadequate case mix adjustment (a threat to the validity of performance reports, as discussed in Appendix 1), a CVE will need the assistance of a statistician.

Example: Deciding how performance measures will be specified

For reports of hospital performance (www.wicheckpoint.org and www.wchq.org), the **Wisconsin Healthcare Value Exchange** uses the following strategy for determining measure specifications:

- If there is a nationally endorsed measure, use its specifications.
- If there is no nationally endorsed measure, use regionally endorsed measure specifications.
- If there are no nationally or regionally endorsed measures, then the "last-case scenario" is for the CVE to design and test its own measure.

For "HEDIS-like"* measures of ambulatory provider performance (www.wchq.org), the CVE tries to stick as close as possible to the national measure specifications, which are intended for use with claims data. However, because the CVE obtains performance data directly from providers rather than from health plan claims, the measure specifications must be translated for this alternative data source. The main goal is to try to capture "the essence" of the denominator that might be applied to claims data.

As a side benefit to providers, the list of patients included in each measure denominator can also be used as a patient registry, and this functionality makes the reporting effort very well accepted by providers.

*HEDIS is the Healthcare Effectiveness Data and Information Set of the National Committee for Quality Assurance.

C. What patient populations will be included?

In the early stage of measure selection and specification, CVE stakeholders also may want to consider which patient populations they would like to include in measuring provider performance. This "included" patient population consists of all the patients whose care generates the performance data that will be used to create performance measures.

The choice of patient population is important for at least two major reasons. First, reported provider performance will have the greatest meaning for patients who belong to the population contributing performance data. For example, if only patients who are Medicare beneficiaries generate performance data, then performance reports will have the most meaning for patients age 65 and older. Provider performance for these patients may or may not accurately indicate how well a provider delivers care to a much younger population.

Second, due to segmentation of the U.S. health care system, certain patient populations will require different data sources than others. This segmentation is a particular concern when constructing performance measures based on health plan claims data. For example, if a CVE wants to include patients age 65 and older, the CVE generally will need to access performance data from traditional fee-for-service Medicare or from a Medicare Advantage plan. If a CVE wants to include patients from vulnerable sociodemographic groups, performance data from Medicaid or uninsured patients may be needed. Some data sources may be difficult or impossible to access.

Even if all potential sources of performance data are available, not all patients captured in these data at any given time can be included in some performance measures. In their specifications, some performance measures have "continuous enrollment criteria," usually meaning that to be included in a measure, a patient must be a member of the same health plan for at least 1 or 2 years. Similarly, some measures may require that a patient receive care from a given provider for 1 or 2 years. However, a significant percentage of patients may switch health plans or providers from year to year, becoming ineligible for inclusion in a measure.

Care for patients who switch plans or providers may differ from care for patients who do not switch. Performance reports may therefore be less meaningful for patients who switch than for those who stay in the same health plan and maintain the same provider. We discuss a way to quantify the extent of this potential problem in the section on Task #4, in the bullet point titled "Compute overall number of patients who qualify for a measure."

Here are two "extreme" options for deciding which patient populations to include in a performance report. These "extreme" options are intended only to illustrate the tradeoffs that may be involved.

1. **"Extreme" Option 1: Include all patient groups present in a CVE's local area.**

   Advantages:

   - May maximize the usefulness of performance reports to a broad population. May also enhance the usefulness of performance reports to providers who may otherwise

receive multiple competing reports, each representing the care delivered to a different patient population.
- *May* reduce the risk of misclassification due to chance because the number of observations for each provider is increased. However, including more patient groups does not guarantee more reliable performance estimates.

Disadvantages:

- Population-based data are generally segmented into different sources (e.g., Medicare, Medicaid, commercial insurance). Obtaining and pooling data from multiple sources may be difficult. For some data sources, legal restrictions on data use may be a barrier to their inclusion in a report.

2. "Extreme" Option 2: Only include patient groups for which performance data are readily available.

Advantages:

- May be easier to generate performance reports.

Disadvantages:

- May limit the usefulness of performance reports, especially for patients from populations whose care is not reflected in the reports.
- May have a high risk of performance misclassification due to chance related to lower numbers of observations per provider.
- May have dissimilar populations of patients included in performance reports and using the reports. This dissimilarity raises the possibility that from the point of view of patients using the reports, providers will be systematically misclassified, resulting in "selection bias" (a threat to the validity of performance reports; briefly discussed in Appendix 1).

Examples: Patient populations included in ambulatory care* performance reports

Most of the CVE stakeholders we interviewed were reporting provider performance using claims-based performance measures. Therefore, these CVEs could include only the patient populations for whom claims data were available (typically commercially insured patients, plus Medicaid enrollees in some cases). However, **Aligning Forces for Quality-South Central Pennsylvania** (www.aligning4healthpa.org) relies on provider medical record reviews (rather than claims) to generate performance data and can therefore include all patients regardless of health plan coverage. Similarly, the **Wisconsin Healthcare Value Exchange** (www.wchq.org) uses provider electronic health record data, including clinical data and lab results, as the basis for most performance measures.

Organizations leading the **Minnesota Healthcare Value Exchange** (www.mnhealthscores.org/ and www.mnhospitalquality.org/) also report some measures of ambulatory care quality based on provider medical records; for these measures, all patient populations are included. The same is true for measures based on claims data, with one exception. Because Medicare claims data are unavailable for performance reporting purposes, no patients with Medicare fee-for-service coverage can be included in the claims-based measures.

*For *hospital* performance reports, all-payer State hospital discharge databases (depending on the State) may enable reporting that includes patients covered by Medicare fee for service, Medicaid, and commercial insurance, as well as uninsured patients. (For a list of State data contacts, go to www.hcup-us.ahrq.gov/partners.jsp).

Decisions Encountered During Key Task #3: Identifying Data Sources and Aggregating Performance Data

A. What kinds of data sources will be included?

There are a few key types of performance data that a Chartered Value Exchange (CVE) may want to collect, and there are different ways to compile these data. The basic types of performance data include:

- Administrative data (e.g., claims, hospital discharge data, prescription fills, laboratory services).
- Medical record data (both paper and electronic).
- Clinical registry data.
- "Hybrid" data (i.e., administrative data that are combined with selected medical record data to improve accuracy).[21]
- Data from patient experience surveys.

The definitions, advantages, and disadvantages of using all of these types of data are discussed in more detail in a separate AHRQ decision guide, *Selecting Quality and Resource Use Measures: A Decision Guide for Community Quality Collaboratives*.[14] However, **from a methodological point of view, a key decision is the degree to which performance data will be *processed* before reaching a CVE.**

To construct performance measures, **"raw" sources of data (e.g., health plan claims, hospital discharge data) must be converted into a format that is ready for measure specifications to be applied.** This conversion (also known as "data cleaning") can be very cumbersome, especially when a CVE does not already have in-house expertise in processing a particular data source. **One approach to dealing with raw data sources is to contract with a data management vendor.** Guidance on selecting and interacting with a vendor is available in the decision guide *Selecting Quality and Resource Use Measures* mentioned above.

Two general models or approaches to data aggregation can be followed:

1. An "**aggregated data model**" where more detailed raw data are aggregated by a CVE to produce performance measures.
2. A "**distributed data model**" where the entity or entities that provide the data (usually health plans) retain many key data elements (especially those that constitute personal health information) and may process the data into provider scores (or numerators and denominators). For example, a health plan might process its own raw claims, apply measure specifications provided by a CVE to these processed claims, and then report measured performance (e.g., numerators and denominators on diabetic eye exams) to the CVE for each provider. This way, the CVE never has to deal directly with raw performance data. However, the CVE still has the freedom to combine these measured

performance statistics with other data sources in its own report. In a distributed data model, a CVE may also be able to construct some types of composite measures.[ii]

A third alternative is to use "prescored" data generated by another performance reporting organization. Examples of prescored data include hospital safety ratings by Leapfrog or categories of hospital death rates from the Centers for Medicare & Medicaid Services (CMS) Hospital Compare. These prescored data have been fully processed into performance scores (or categories of performance), and CVEs generally cannot influence how these measures are specified or how performance is classified. The advantage of using prescored data is that a CVE can report these scores (or performance categories) without needing to process any data. However, using prescored data may limit a CVE's options for addressing performance misclassification, whether systematic or due to chance.

1. **Option 1: Obtain raw performance data ("aggregated data model").** Raw performance data include health plan claims, hospital discharge data, medical record abstracts, and patient survey responses.

   Advantages:

   - This approach maximizes a CVE's degree of freedom to decide how performance measures will be specified and reported. A CVE will be able to decide the organizational level of reporting (e.g., individual practitioner or provider group) and determine how to construct composite measures.
   - By handling raw performance data, a CVE will learn the limitations and flaws of these data. A CVE also may work with health plans and providers on data improvements to help facilitate future measurement and reporting efforts.
   - Because raw performance data can be processed on a patient-by-patient basis, this approach allows for a detailed data review and correction process.
   - This approach allows maximum flexibility and range of options in attributing data to providers, performing case mix adjustment, and dealing with the risk of performance misclassification due to chance.
   - This approach maximizes the potential for performance data to be used for research.

   Disadvantages:

   - Processing raw performance data may require substantial experience and can be difficult and expensive. A data management vendor will be needed.

[ii] A CVE using a distributed data model may be able to construct composite measures using a "weighted average" approach. Because a CVE using distributed data models may not receive patient-level data, it may not be possible to construct "all-or-none" composite measures. These types of composite measures are discussed in more detail in the section on Summary of Methodological Decisions Made by a Sample of CVE Stakeholders.

- Raw performance data may contain individually identifiable health data about patients. When such data are present, a CVE must take additional precautions to preserve the privacy, confidentiality, and security of these data. Depending on the type of data, there may be additional legal considerations (e.g., Health Insurance Portability and Accountability Act).

Example: Using raw health plan claims ("aggregated data model")

The **Oregon Health Care Quality Corporation** (q-corp.org) and **Puget Sound Health Alliance** (www.wacommunitycheckup.org) both receive raw claims data from commercial and Medicaid health plans. These CVEs share an experienced data contractor that processes the claims, working with the health plans and other CVE stakeholders to identify and address missing data, check data interpretation, and calculate provider performance scores.

2. **Option 2: Use a distributed data model.** Raw performance data can be processed by the original sources of these data, using measure specifications provided by the CVE. For example, health plans may process their own claims data and send provider-level performance measure numerators and denominators to a CVE, rather than patient-level data that would need to be aggregated up to the provider level.

   Advantages:

   - CVE avoids the cost and difficulty of processing raw performance data.
   - CVE retains some flexibility in specifying performance measures, specifying case mix adjustment methods, and addressing other analytic concerns.

   Disadvantages:

   - If the needs of a CVE change, it may be difficult for data sources to agree to reprocess the raw performance data and send new kinds of measure output to the CVE.
   - Potential exists for misleading reports if CVE partners who produce the data do not process their data in the same way or use exactly the same measure specifications.
   - When case mix adjustment is desired, using a distributed data model may limit the types of adjustment methods available to a CVE, as case mix adjustment often requires more granular information such as patient characteristics. The section on Task #5 discusses situations in which case mix adjustment may be warranted.
   - As with a vendor, a CVE may need to perform audits to determine whether data are being processed as specified by the CVE.

Examples: Using a distributed data model

- **Massachusetts Health Quality Partners** (www.mhqp.org), **Greater Detroit Area Health Council** (www.gdahc.org), and **Healthy Memphis Common Table** (www.healthymemphis.org) obtain HEDIS measure numerators and denominators from each of their health plans. Therefore, each health plan deals directly with its own raw administrative data.
- The **Quality Alliance Steering Committee (QASC) and America's Health Insurance Plans (AHIP)** are piloting a prototype distributed data model that includes a subset of Colorado and Florida health plans in an effort to generate HEDIS® (Healthcare Effectiveness Data and Information Set**)** measures (www.healthqualityalliance.org/hvhc-project).
- To generate performance data for reports by **Aligning Forces for Quality-South Central Pennsylvania** (www.aligning4healthpa.org), providers randomly sample their own medical records and abstract these records to generate numerators and denominators on diabetes quality of care measures.

3. **Option 3: Use "prescored" data.** Examples of fully processed performance scores include Leapfrog patient safety ratings and ratings from Medicare's Hospital Compare, patient experience measures (H-CAHPS [Hospital Consumer Assessment of Healthcare Providers and Systems]), hospital mortality rates, and readmission ratings.

   Advantages:

   - Data have already been completely processed into performance scores that may be ready for reporting (and may already have been reported).

   Disadvantages:

   - CVE has little or no control over measure specifications.
   - CVE has little or no control over providers (in the case of Leapfrog ratings) or payers (in the case of CMS Hospital Compare) that are represented in the prescored data.
   - CVE has limited options for performing case mix adjustment, addressing misclassification risk, and dealing with other analytic concerns.
   - Important measure details and data validity checks (e.g., specifications, attribution rules, case mix adjustment methods, and level of misclassification risk) may or may not be available, depending on the documentation available from the source of the prescored data.

B. How will data sources be combined?

If a CVE uses more than one source of performance data for a given measure, then these data sources will need to be combined to report for each provider a single level (or category) of performance on that measure. For example, a CVE may collect performance data on a diabetes quality of care measure from three commercial health plans, plus Medicare and Medicaid. A given provider may have patients with diabetes from each of these five payers. However,

reporting five separate performance scores for each provider on this measure (one for each data source) might confuse patients. Receiving multiple performance scores on the same measure may annoy providers, especially when the scores are very different across sources. These divergent scores may also be a sign of small denominators within each data source (as discussed in Appendix 2) or inadequate case mix adjustment (as discussed in the section on Task #5).

Combining data from multiple sources is not a trivial task, given variations in coding practices across public and private payers. For example, it is not unusual for different payers to have different provider identifiers, which creates challenges in generating a unified provider file. The degree of difficulty in aggregating data across multiple sources partly depends on the amount of data processing that has occurred before these data reach the CVE (see the section on Task #3) for more discussion of preprocessed data). In general, **the less preprocessed the data from multiple sources, the more work is necessary to combine these data.**

Scenario 1: Starting with raw performance data from multiple sources ("aggregated data model")

As discussed earlier, "raw" performance data are data to which measure specifications have not yet been applied. In other words, these raw data have had little or no processing. Claims for a health plan's members are a common example of raw performance data. To process raw data into performance scores, a CVE will need to work with a data vendor. The AHRQ decision guide *Selecting Quality and Resource Use Measures: A Decision Guide for Community Quality Collaboratives* contains guidance on selecting and working with a data vendor.[14]

Calculating the measures within each data source offers a chance to ensure that measure specifications are being correctly applied. For example, scores on a measure may change dramatically depending on whether the performance data are from source A or source B, with no reasonable explanation. (One data source may represent a higher risk population.) In this case, the measure specifications may have been incorrectly applied to one of the sources, or there could be problems with the data from one or more sources.

Because every source of raw data is different, it may be advisable for the CVE or data vendor to directly consult with each source to resolve any questions about how the data are coded (see the section on Task #4). For example, if a CVE is calculating a diabetes measure from a health plan's data, the CVE may want to review the measure specifications with health plan staff. This step can help ensure that specifications will identify the intended population of patients with diabetes.

Scenario 2: Using a distributed data model with multiple sources

In a distributed data model, raw performance data can be processed by the sources of these data before the data are shared with the CVE, using measure specifications provided by the CVE. For example, in a distributed data model, health plans can calculate a provider's numerator and denominator for each Healthcare Effectiveness Data and Information Set (HEDIS) measure and report these to the CVE.

In a distributed data model with multiple sources, the major challenge to combining data across sources is ensuring consistent provider identification. When data sources report performance to a CVE, the performance being reported must be linked to a provider identifier

(e.g., a numeric code or name representing the physician whose performance is being reported). The central problem that CVEs may commonly encounter is that each data source may use a different set of provider identifiers. In other words, Dr. Jones might have one identifier in Plan A and another identifier in Plan B. In addition, Dr. Jones' name may be represented differently across the different health plan files. To combine Dr. Jones' performance reported by Plan A with Dr. Jones' performance in Plan B, **a CVE will need a "crosswalk" that links the identifiers for each provider across the data sources to be combined.**

The following are two options that illustrate ways to create a provider crosswalk.

1. **Option 1 for provider crosswalk: Use readily available provider identifiers.** Some provider identifiers may be readily available to a CVE. These include provider taxpayer identifiers, national provider identifiers (NPIs), Drug Enforcement Administration (DEA) numbers, State medical license numbers, and Medicare billing identifiers. These identifiers may correspond to providers of different types, including individual physicians, medical groups, hospitals, and integrated health care delivery systems.

   Advantages:

   - Using these identifiers is relatively economical.
   - For hospitals and other large provider organizations, readily available identifiers may be highly accurate.

   Caveats:

   - For individual practitioners and small outpatient practices, a crosswalk based on readily available identifiers may have low accuracy. For example, a tax ID may include providers that actually have little to do with each other, aside from sharing a common billing system. In addition, it may be difficult to know which tax IDs represent individual practitioners and which represent larger groups, making comparisons more difficult.
   - The crosswalk may not be able to link a large number of providers across data sources. This may happen when one data source does not include the same "readily available" identifiers as another.
   - The crosswalk may not enable a CVE to identify provider attributes. For example, it may be impossible to know which tax IDs represent individual physicians and which represent small groups. It also may be impossible to determine the specialty of each provider.
   - The ability to change the level of reporting may be limited. If the crosswalk only contains individual physician identifiers, then it may not be possible to report performance at higher levels of provider organization (e.g., the medical group). Reporting for larger groups of providers can be an important option for limiting misclassification risk (see the section on Task #5).

Example: Building on readily available provider identifiers

The **Healthy Memphis Common Table** (www.healthymemphis.org) began its performance reporting efforts by using the Medicare GEM* dataset: provider identifiers were obtained from a single plan. When a group could not be identified (15% of the time), staff followed up directly with providers and practice managers to let them self-identify. When commercial health plan data were later used, checking health plan provider identifiers for accuracy and consistency (via telephone calls to providers) revealed that the tax IDs did not always match across health plans. In these cases, additional variables were used for matching, such as provider address. Solo practices also were examined specifically to ensure that the apparent "practice" was not part of a larger provider group; true solo practices were not reported. The CVE is now working to develop a master directory of providers in the Memphis area.

* GEM refers to the Generating Medicare Physician Quality Performance Measurement Results Project.

2. **Option 2 for provider crosswalk: Create a "master provider directory."** A master provider directory is an organizational mapping of all the known providers in a CVE's local geographic area. This mapping tells which individual practitioners are affiliated with which practice sites (or clinics), tells which practice sites are part of which larger medical groups, and may include affiliations with larger provider organizations. Other data that may be included in a master directory are individual provider specialties, certifications, and acceptance of new patients. Finally, to enable a CVE to combine performance data from multiple sources, the master directory must contain a crosswalk with the provider identifiers used by each data source.

   Having a master provider directory is especially important when a CVE is reporting the performance of individual providers or small groupings of providers. If a CVE is only reporting hospital performance, then a master directory may be less useful. Creating such a directory may require substantial time, effort, and resources since collecting new data (often by directly contacting providers) is almost certain to be necessary. Maintaining a master directory also requires ongoing investment, since providers often change their affiliations. However, once created, a master directory also has distinct advantages.

   Advantages:

   - The master provider directory can serve as a common reference point for all data sources to ensure valid aggregation of performance data.
   - The directory may help convince providers that performance is being accurately reported. By contacting providers as part of master directory maintenance, the CVE demonstrates a commitment to accurate reporting.
   - The directory offers flexibility in determining the best level of provider organization for performance reporting. Reporting for larger provider groups can be an important option for limiting misclassification risk (see the section on Task #5).

Disadvantages:

- Requires significant time and resources.
- Requires provider engagement. The accuracy of a CVE's master directory will only be as good as the information given by providers.

Examples: Creating a master provider directory

Massachusetts Health Quality Partners (MHQP; www.mhqp.org) **uses a "Master Physician Directory" to combine data from its five participating health plans.** This master directory also enables MHQP to report HEDIS performance at the medical group level and simultaneously report patient experience survey data at the practice site level (a lower level of provider organization). To create the directory, MHQP relied on readily available physician identifiers (e.g., license and DEA numbers) and provider addresses. Provider organizational mappings from local health plans conflicted with each other, so MHQP engaged in direct outreach to providers to learn their self-identified organizational relationships. The directory is updated annually, and this update is now facilitated by a computer interface that allows providers to correct their pieces of the directory. It took the MHQP directory roughly 10 years to reach a "steady state" in which the same percentage of providers (~5-10%) changes affiliations from one year to the next. At this point, MHQP leaders believe these changes of affiliation no longer represent corrections of past errors. Instead, these changes represent true changes in provider affiliation that occur when providers move or groups change their configurations.

The **Oregon Health Care Quality Corporation** (q-corp.org), created an Oregon practitioner directory listing primary care clinics with 4 or more physicians (including roughly 2,000 of 3,000 such physicians in the State). Creating an accurate directory required Internet sleuthing and direct outreach via telephone. When plans for public reporting were circulated, the clinics began to actively participate in correcting their directory entries. **Puget Sound Health Alliance** (www.wacommunitycheckup.org) similarly created a provider directory that included clinics with four or more physicians.

Minnesota Community Measurement (a constituent of the Minnesota CVE; www.mnhealthscores.org/) also created a master provider directory for ambulatory physician clinics. It took 3 years to create and verify this directory.

General approach to combining data once crosswalk is complete

Creating an accurate provider crosswalk may be the most difficult part of combining performance data from multiple sources. However, the best way to aggregate performance data for each provider can be unclear. From a methodological standpoint, aggregating provider performance across multiple data sources (on a single measure) is very similar to creating a performance composite from multiple individual measures.

Two key methodological concepts apply: validity and reliability. Greater validity means that a smaller share of providers will be *systematically* misclassified in a performance report. Greater reliability means that a smaller share of providers will be misclassified *due to chance alone* in a report. Both of these concepts are discussed in more detail in Appendixes 1 and 2, but their application to combining multiple-source data is briefly discussed here.

To maximize measurement reliability, performance data from each source can be weighted when they are combined. A general recommended strategy is to give performance scores that are based on fewer observations less weight than those that are based on more observations. In other words, this approach allows more reliably measured scores to have more influence than less reliably measured scores. This weighting strategy is straightforward for measures with numerators and denominators. By separately summing the numerators and denominators from all sources and then dividing the summed numerator by the summed denominator, a CVE will produce the most reliable performance estimate that is possible with the data available.

For example, a provider might deliver a HEDIS service to 40 out of 50 patients (80%) in health plan A and 5 out of 10 patients (50%) in plan B. Using the recommended strategy for combining data from these plans, the summed numerator is 45 (40 + 5) and the summed denominator is 60 (50 + 10). Therefore, the combined performance score is 75% (45 divided by 60). Note that 75% is much closer to 80% (the plan A score) than to 50% (the plan B score). The combined score is closer to the plan A score because this strategy of combining performance data *automatically weighted the data appropriately*, giving more weight to plan A, which had more observations and therefore a more reliable measured score.

Validity issues may arise when combining data sources because different data sources may contain data generated by dissimilar patient populations. For example, a CVE may want to combine data from Medicare with data from a commercial health plan. However, these two patient populations may differ in many important ways. It may be misleading to compare "Provider A," who mostly sees patients with Medicare, to "Provider B," who mostly sees patients with commercial insurance, on measures of mortality. This might be the case because patients with Medicare are probably older than those with commercial insurance and therefore have a higher baseline rate of mortality. Thus, even if Provider A gives care that is equal to Provider B's care for both patient populations, Provider A will appear to have worse performance (i.e., a higher mortality rate).

Problems combining data can be addressed through the case mix adjustment or stratification methods discussed in the section on Task #5. However, we note here—and explain in more detail in the section on Task #5—that case mix adjustment is not always a straightforward methodological decision.

C. How frequently will data be updated?

Provider performance is likely to change over time, so CVEs will want to periodically update the data in reports of provider performance. From a methodological perspective, there is no real downside to updating performance data as frequently as possible, using the most recent data available. After all, if the performance data contained in a report are too old, then they may no longer accurately represent provider performance.

One important caveat applies to updating performance data: As updates become more frequent, CVEs may be tempted to reduce the number of observations included in each update. An extreme example of this practice would be to send out weekly updates on a patient experience survey, sharing just the surveys that were returned in the preceding week. If only a few surveys are received each week, then week-to-week scores could fluctuate wildly due to chance alone (i.e., week-to-week scores would have low reliability).

To increase measurement reliability, frequent updates may need to be accompanied by a "rolling average" approach to calculating provider performance. In this approach (discussed in the section on Task #5), data from preceding periods are combined with data from the most recent period to increase the number of observations. In more complex versions of the "rolling average" approach (e.g., Bayesian methods), more recent performance data get more weight than older performance data.

Other than potentially incurring greater expense, there is no practical downside to updating performance data as frequently as possible. To help decide how much expense is worthwhile, CVE stakeholders may aim for matching the frequency of data updates to the minimum length of time necessary for changes in true performance to occur. For most performance measures, it is probably not plausible for true performance to change on a week-to-week or even month-to-month basis.

Example: Frequency of data updates

The **Greater Detroit Area Health Council** (GDAHC; www.gdahc.org) updates the performance data in its public reports on an annual basis, with a lag of at least 1 year between the time clinical care is delivered and the time performance data are reported. Devorah Rich, formerly Project Director of GDAHC, explains that in the future, "real time reporting" is desired: "The analogy is like trying to lose weight. When groups are working hard, they want to know whether these efforts are successful and they want to get recognized for this."

Decisions Encountered During Key Task #4: Checking Data Quality and Completeness

Whether a Chartered Value Exchange (CVE) receives raw data, uses a distributed data model, or reports "prescored" measures, the quality and completeness of performance data are key factors that determine whether a performance report can provide useful information to patients and providers. The delegation of "data auditing" tasks may depend on how a CVE is handling performance data:

- If a CVE receives and processes raw performance data (e.g., health plan claims), then the CVE itself may want to perform the "data auditing" tasks described in this section.
- If a CVE contracts with a vendor to process raw performance data, the CVE may request from its vendor a plan for data auditing and a report of what was done (once data auditing and preparation are finished). The data auditing plan may identify the processes the vendor will use to edit, clean, quality check, and amend the data. The auditing report may describe the results of these activities and provide a list of known data quality and completeness issues.
- If a CVE uses a distributed data model (see section on Task #3), the CVE may consider discussing these data auditing tasks with the sources of its performance data. A CVE may want to have each data source reviewed by an independent auditor.
- If a CVE reports "prescored" measures (see section on Task #3), then the CVE may want to consult the existing documentation for these measures to see what kinds of data auditing steps were performed.

A discussion of practical approaches to data auditing is available in a separate AHRQ decision guide titled *Selecting Quality and Resource Use Measures: A Decision Guide for Community Quality Collaboratives*.[14]

A. How will tests for missing data be performed?

To determine the extent to which data are missing, there are two main kinds of missing data to try to detect:

- **Missing data within a record.** A "record" refers to a unit of observation, such as a patient office visit. If a database contains a notation that an office visit occurred, but the diagnosis (or reason for visit) is absent, then this data element can be considered missing. Another example is in patient experience surveys. If a survey is returned but a question has been skipped, then this skipped question constitutes a missing data element within the record (the survey). It is generally easier to detect missing data within a record than to detect entire records that are missing.
- **Missing records.** Examples of missing records include entire surveys that are not returned and office visits that are not included in administrative data. For surveys, a list of patients to whom the survey was mailed will allow a CVE to know the extent of missing data. But for office visits, the situation is more difficult. How can a CVE tell that an office visit occurred when there is no record? After all, maybe the office visit never occurred in the first place.

Missing data are a difficult problem, even for experienced programmers and analysts. One way to check for missing data is to compare the performance data to an external standard, such as the documentation that accompanies the data. If performance data come with documentation that lists the number of records, a good first step is to check that the number of records in the data file is equal to the number listed in the documentation. In a related example, if the number of patients in the performance data from a given health plan is much smaller than the number of patients known to be enrolled in the health plan, then it is likely that many patient records are missing. Other examples include a complete lack of mental health data from a health plan (due to the plan's use of a "carve-out" subcontract for mental health services) and a complete lack of pharmacy data (due to use of a pharmacy benefit manager). The only way to fix these problems is to go back to the data source and figure out a way to obtain the missing data.

In general, missing records are detected in two situations. First, the number of records in a dataset may not match the number of records listed in the dataset documentation (i.e., a description of the dataset that gives the number of records). Second, the existing data may be implausible (e.g., it is extremely unlikely that an entire health plan's membership would consume no mental health services or no prescription drugs in a given year). Sometimes, however, there may not be any such red flags. It is much harder to detect missing records when the existing data still look plausible. **The best techniques for detecting missing data depend on the data source in question, and obtaining consultation from analysts who are experienced with each data source may be advisable.**

Examples: Missing data

The New York Quality Alliance (www.nyqa.org) is using adjudicated health plan claims data to calculate Healthcare Effectiveness Data and Information Set (HEDIS) performance measures. However, when patients are enrolled in capitated products, the health plans do not receive claims for every clinical service that is delivered. It is difficult, therefore, to know the extent to which HEDIS performance measure data for capitated patients are missing.

B. How will missing data be handled?

There are a number of options for handling missing data. Some options are based on statistical techniques, focusing on trying to make performance reports as complete and accurate as possible. Other options for handling missing data are intended to create incentives for data sources to report data that are more complete. These options may sacrifice some short-term accuracy in exchange for the longer term goal of getting more complete data in the future. The choice between statistical techniques and creating incentives can be guided by the reasons the data are missing.

For the purpose of creating reports of provider performance, there are two main reasons performance data might be missing[iii]:

- **Data can be missing in a way that is *not* related to true provider performance.** For example, a survey of patient experience could have had a printing flaw that led many patients to skip certain items. Or a computer problem could have deleted all data pertaining to clinical care over a 1-month period. Some types of clinical data (e.g., lab values) might be rarely recorded in administrative databases.
- **Data can be missing in a way that *is* related to true provider performance.** For example, if a CVE is getting performance data directly from providers, some providers might choose not to report data that are likely to show poor performance.

Missing data are problematic because they can cause performance misclassification. When data are missing in a way that is not related to true provider performance, then having more missing data will increase the risk of performance misclassification due to chance (i.e., lower the reliability of performance measurement, which is discussed in Appendix 2). Misclassification risk will rise because of fewer performance observations.

However, if data are missing in a way that is related to true provider performance, then having more missing data will increase the risk of systematic performance misclassification (i.e., introduce statistical bias, which is discussed in Appendix 1). If, for example, low-performing providers tend to selectively withhold data that would indicate poor performance, then they will be systematically misclassified as having performance that is higher than their true performance.

To determine the reasons for missing data, a CVE can query the suppliers of the performance data and perform data audits. In some cases, a statistician may be able to help distinguish between the reasons for missing data. The following options for handling missing data are intended to give an overview of the types of strategies a CVE can use. Additional discussion of options for handling missing data is available in an earlier RAND report.[22]

1. **Option 1: Imputation.** Imputation refers to a family of statistical methods that use the available data from a given provider (i.e., the data that are not missing) to "fill in" the missing data for that provider. In addition to providing estimated values for the missing data, these imputation methods can compute the amount of uncertainty associated with these estimated values. In other words, statistical imputation gives an educated guess for each missing data element as well as a sense of how good the guess is likely to be.

   Advantages:

   - Given the available data, enables calculation of performance estimates that are as accurate as possible. However, the successfulness of imputation techniques will depend on the reasons the data are missing: *Imputation will be most successful when data are missing in a way that is not related to true provider performance.*

[iii] Note that the *reasons*data might be missing are distinct from the *kinds* of missing data that were discussed in the preceding section.

- Maximizes the number of measures and providers that can be included in a performance report.

Disadvantages:

- Methodologically complex; may not be needed in many situations. Statistical imputation will require consultation with a statistician experienced in these techniques who can advise on the potential advantages of imputation.
- May produce inadequate results. Imputed data are only as good as the data that are not missing. If too many data are missing, then imputation may not produce performance estimates with enough certainty to be useful for reporting.
- May create a disincentive to report more complete data in the future.
- May be difficult to explain to stakeholders. Imputation may raise the likelihood that stakeholders will mistrust performance reports.

2. **Option 2: Report the average score for measures with missing data.** Suppose a CVE wants to include six performance measures in a report, but for a given provider ("Provider X"), performance data on two of these measures are missing. For these two measures, the CVE could report, for example, the average performance of all providers as the performance for Provider X.

   Advantages:

   - Methodologically simple; easy to explain to stakeholders.

   Disadvantages:

   - If data are missing in a way that is related to performance, then imputing the average score is likely to systematically misclassify performance.
   - May create a disincentive to report more complete data in the future. When performance data are obtained directly from providers, this approach *may create an incentive for providers to withhold data* for any measure on which performance is lower than average.

3. **Option 3: Report only the available data.** Suppose a CVE wants to include six performance measures in a report, but for a given provider, the data needed to generate scores on two of these measures are missing. In this case, the CVE could report performance for this provider on the four measures for which data exist, placing a "not reported" marker in its report for the other two measures.

   Advantages:

   - Methodologically simple; easy to explain to stakeholders.

Disadvantages:

- May confuse patients, who might not understand what the "not reported" marker means.
- May result in systematic performance misclassification if data are missing in a way that is related to performance and only the remaining measures are reported.
- May create a disincentive to report more complete data in the future.

4. **Option 4: Report performance only for providers that are not missing data on *any* measure.** Under this approach, if a provider were missing data on any measure reported by a CVE, then this provider would receive a "not reported" marker on *all* measures. This "not reported" marker would even apply to measures for which data are available.

   Advantages:

   - Methodologically simple; easy to explain to stakeholders.

   Disadvantages:

   - Many providers may have no reported performance data, limiting the usefulness of public reports to patients.
   - As the number of measures grows, the number of providers with reported performance may fall (since there are more chances to have missing data).
   - This approach carries unclear incentives for future reporting and may encourage nonreporting by low performers.
   - This approach may be unacceptable to providers and patients.

5. **Option 5: Report the lowest *possible* score when data are missing.** When a provider has missing data on a given measure, a CVE can report the provider's performance as the lowest possible score.

   Advantages:

   - Methodologically simple; easy to explain to stakeholders.
   - Creates an incentive for complete data reporting in the future.
   - May result in less systematic performance misclassification than Options 1 through 4 if data are more likely to be missing when performance is low..

   Disadvantages:

   - If data are missing in a way that is *not* related to performance (i.e., missing due to chance alone), likely to systematically misrepresent performance as being much lower than it really is.
   - May be unacceptable to providers.

6. **Option 6: Report the lowest *observed* score when data are missing.** When a provider has missing data on a given measure, a CVE can report the provider's performance as being equal to the lowest observed score among providers who are not missing data on the measure.

   Advantages:

   - Methodologically simple; easy to explain to stakeholders.
   - Creates an incentive for complete data reporting in the future.
   - May result in less systematic performance misclassification than Options 1 through 4 if data are more likely to be missing when performance is low..
   - May be more acceptable to providers than imputing the lowest *possible* score.

   Disadvantages:

   - If data are missing in a way that is *not* related to performance (i.e., missing due to chance alone), likely to systematically misrepresent performance as being much lower than it really is.

Table 1 summarizes the relationship between the reasons data are missing and the strengths and weaknesses of the options for handling missing data.

Table 1. Strengths and weaknesses of strategies for handling missing data

Reason data are missing	Data missing in a way that is *not* related to true performance (e.g., missing at random)	Data missing in a way that *is* related to true performance (e.g., low performers not reporting data)
Option 1: Imputation	*Stronger*: may reduce the risk of misclassification due to chance	*Weaker*: may introduce systematic performance misclassification
Option 2: Report the average score	*Stronger*: will not result in systematic performance misclassification	*Weaker*: high likelihood of resulting in systematic performance misclassification; *and* creates incentive not to report low performance
Option 3: Report only the available data	*Stronger*: will not result in systematic performance misclassification	*Weaker*: may create incentive not to report low performance
Option 4: Report only when providers are not missing any data	*Stronger*: will not result in systematic performance misclassification	*Weaker*: may create incentive not to report low performance
Option 5: Report the lowest possible score	*Weaker*: high likelihood of resulting in systematic performance misclassification*	*Stronger*: potentially less likely to result in systematic performance misclassification; creates incentive to report all performance data
Option 6: Report the lowest observed score	*Weaker*: high likelihood of resulting in systematic performance misclassification*	*Stronger*: potentially less likely to result in systematic performance misclassification; creates incentive to report all performance data

*For example, if providers with fewer patients are more likely to have missing data (regardless of their performance), then these providers will be systematically misclassified as low performers.

Different approaches to dealing with missing data can be used for different measures within the same performance report. For example, it is common to impute missing values in composite measures *without* imputing missing values when the individual measures are presented (i.e., in a drilldown screen). This combination of strategies is attractive because the cumulative risk of having one or more missing data elements increases with the number of individual measures included in a composite. In addition, the impact of any one missing element decreases as the number of indicators in a composite increases, so the misclassification risk associated with erroneous imputation is less for a composite measure than for an individual measure.

Examples: Approach to "missing" performance data

Each of the nine CVEs we interviewed described working with data sources to minimize the amount of missing data. After this step, CVEs report the available data (Option 3 above). When performance data for a provider cannot be reported due to concerns about the risk of misclassification due to chance (e.g., insufficient numbers of observations), a symbol is generally used to indicate that performance cannot be reported. For examples of using such symbols, see the reports of the **Healthy Memphis Common Table** (www.healthymemphis.org) and the **Puget Sound Health Alliance** (www.wacommunitycheckup.org).

C. How will accuracy of data interpretation be assessed?

Even when performance data are present, it is possible for these data to be misinterpreted during the computation of measure performance. Problems with data accuracy are likely to be greatest when a CVE plans to generate performance scores (on quality or cost measures) using health plan claims. For example, measure specifications may assume that patients with diabetes are identified using a particular set of diagnosis codes, when in fact such patients are identified using a different set of diagnosis codes. This is a particular problem when combining performance data from multiple data sources, because sources may not all have the same ways of coding conditions, health care services, and patient outcomes (as discussed in the section on Task #3.)

Generally speaking, the same kinds of approaches used to check for missing records can be used to assess the accuracy of data interpretation. Some of these approaches for detecting missing records are discussed in the section on Task #4. These approaches include:

- **Compute overall number of patients who qualify for a measure, and make sure this number makes sense.** For example, if a measure is supposed to apply to all patients with diabetes, does the number of diabetics identified seem realistic based on other known data? Comparison data may be available from local health departments or from the data source itself (e.g., a health plan that supplies data may also conduct disease management outreach and have a roster of its diabetic enrollees).

  This step also gives a CVE the opportunity to see how many patients are excluded by continuous enrollment criteria. In their specifications, some performance measures have "continuous enrollment criteria." This usually means that in order to contribute performance data, a patient must be a member of the same health plan (or a patient of the

same provider) for at least 1 or 2 years. However, a significant percentage of patients may switch health plans from year to year, becoming ineligible for inclusion in a measure. This may result in a *large* reduction in the percentage of patients who are contributing performance data to a report. In some cases, more than half of all patients in a CVE's area may be excluded by these criteria.

- **Recheck the number of patients who qualify for a measure on a provider-by-provider basis.** The question again is: "Do the numbers make sense, relative to some kind of external standard?" The advantage of calculating provider-by-provider patient counts for a measure is that these counts can, when providers have disease registries, be verified by the providers themselves. A measure's specifications may include a patient population that is somewhat different from what a provider would report (e.g., because of continuous enrollment requirements). Thus, some degree of disagreement between the measure-identified and provider-identified patient counts can be expected. But the figures should at least be in the same ballpark. A notable caveat to this approach is that some providers probably will not participate in data verification.
- **Check the range of numeric values within each type of data, and make sure these are consistent with what the data are supposed to represent.** For example, if a given variable is supposed to represent hospital length of stay, this variable should never be a negative number. The same is true of patient age, which should never be negative and will infrequently be greater than 100 years.
- **Compute population-level performance measure scores and compare to external benchmarks.** For example, a CVE may compute the overall performance rate within its geographic area on a given HEDIS measure. Then this overall rate can be compared to national performance data: does it seem to be in the right ballpark? The ability to make comparisons to national benchmarks is one advantage of using nationally endorsed measure specifications (also discussed in the section on Task #2.)).

When problems are discovered in these data checking steps, there may be inaccuracies in data interpretation. In other words, the measure specifications may assume that the raw performance data mean one thing when in fact they mean another. Just as when missing record problems are detected, **the best approach to correcting data interpretation problems may be to consult with analysts who are experienced with each data source.** Such analysts are likely to be in the best position to understand how data misinterpretations have occurred and to find solutions.

Decisions Encountered During Key Task #5: Computing Provider-Level Performance Scores

Once aggregated performance data have been audited and initial performance measures have been chosen, provider-level performance scores can be computed. Before a Chartered Value Exchange (CVE) computes these scores, however, performance data must be attributed to providers. Attribution is relatively simple when each patient receives the care that is being measured from only one provider. However, attribution becomes more complex when a performance measure encompasses care that patients receive from more than one physician or hospital (e.g., measuring care delivered over a time period, as in an episode of care).

The parties responsible for attribution and computation may vary depending on what kind of data a CVE receives. For example, if a CVE receives raw performance data, then the CVE will likely complete these tasks internally or hire a vendor to complete them. If a CVE uses a distributed data model in which performance scores are generated by the sources of performance data (see the section on Task #3), then these sources will perform these tasks with the CVE's guidance. If other organizations (e.g., health plans) will complete the attribution and computation, a CVE's role may be to ensure methodological consistency. In other words, each provider of performance scores in a distributed data model should be attributing performance data and computing performance scores in the exact same way.

A. How will performance data be attributed to providers?

To generate reports of provider performance, a CVE needs to ensure that the data used to calculate performance scores are attributed to providers. In other words, each piece of performance data that goes into a report must be associated with a provider. **There are many different ways to attribute performance data to providers. The best way to attribute performance data depends on the purpose of performance reporting and the type of measure being reported.** For example:

- If the purpose of performance reporting is **to foster a sense of teamwork and shared responsibility among a group of providers**, then performance data might be attributed to the group. Attribution to individual physicians (or other types of providers) might be reserved for confidential reports to each group of providers.
- If performance reporting is intended **to raise community awareness or foster cooperation among the providers** serving a community, it may be sufficient to attribute performance data at the community level (without attributing data to any particular provider). This strategy may be especially attractive when improving performance on a measure is likely to require the coordinated activities of many different providers. Public health measures such as infant mortality are one type of performance measure commonly attributed at the community level.
- If performance reporting is intended **to help patients choose a given type of provider**, then attribution of performance data to each provider of that type would be optimal. The goal is to make the types of providers in the report match the types of providers that patients are seeking. If patients are looking for individual physicians, then a report that attributes performance data to individual physicians might be the most useful to these patients.

In attributing data to a given provider, the goal is often to identify those patients for whom the provider is *responsible* for providing care and for whom the provider can *affect* the health services, patient experiences, costs of care, or clinical outcomes being measured. The same rules can be applied when attributing episodes of care (rather than patients) to providers. The goal is to identify episodes of care for which a provider is *responsible* and for which the provider can have an *impact* on the measure being applied to the episode.

On the surface, this sounds straightforward, and in some situations it is. For example, a hospital performance measure might indicate the rate of an immediate complication of surgery (e.g., intraoperative mortality). A straightforward rule would be to attribute performance data to the hospital in which the surgery took place. Another example of a straightforward attribution rule occurs when patients fill out surveys about their hospitalizations; the survey responses are generally attributed to the entire hospital, and this is the approach taken by Hospital Compare.[23]

However, attribution of performance data is often not straightforward. Attribution problems are especially pertinent when a CVE wants to attribute performance data to individual practitioners. Patients often receive care from many different providers, and the processes, outcomes, costs, and experiences of care may be influenced by all these providers. For example, it might seem logical to assign measures of care for hypertension to the primary care provider (PCP) for each patient. But how can a CVE tell who a patient's PCP is (especially when the patient is enrolled in a fee-for-service or preferred provider organization [PPO] product)? This can be especially difficult if only administrative data are available.

In an analysis of Medicare claims data that were supplemented by a physician survey, Pham and colleagues found that in a single year, patients saw a median of seven unique physicians.[24] Using some attribution rules discussed below, Pham and colleagues found that only 79 percent of patients could be assigned to a PCP. Moreover, PCP assignment changed for nearly half of patients over a 2-year period. Attribution to specialists was possible for fewer patients.

In another study of community health centers, Landon and colleagues found that many patients infrequently received care (less than once per year), and simulated performance on quality measures would depend on how these patients were attributed to each community health center.[25] An additional recent study by Mehrotra and colleagues found that 12 different attribution rules would lead to substantially different reports of the performance of individual physicians on episode-based measures of the costs of care.[26]

Held against the standards of *responsibility* and *impact* mentioned above, these are sobering results. With patients seeing so many different providers, deciding which provider is responsible for which performance measures can be a challenging task. This may be especially true of measures of health outcomes. Even if a PCP (who presumably would accept responsibility for providing certain services) can be reliably identified, can this PCP justifiably be held responsible for health outcomes that may have developed over decades? Assigning responsibility for certain health outcomes may not be justifiable unless patients can be consistently assigned to a provider over time.

There is no single, generally accepted "best way" to attribute performance data to providers. CVEs may choose from a variety of attribution strategies, and these strategies may

vary by measure and by provider type. We suggest that CVEs include all stakeholders in negotiations over attribution rules.

To help guide these negotiations, CVEs can refer to the guiding questions discussed above:

- What is the *purpose* of reporting?
- For a given patient or episode of care and the type of performance being measured, which providers are plausibly *responsible* for providing the associated care?
- For a given patient or episode of care and the type of performance being measured, which providers can plausibly have an *impact* on measured performance?
- *Later, once different attribution rules have been tried:* How much of a difference in reported performance do different attribution rules really make?

The options presented below illustrate some attribution strategies a CVE might consider, but many others are possible. Because there are so many possibilities, a CVE may want to revisit this decision at a later point and determine what effects a different choice of attribution strategy would have had on performance reports (see the section on Task #6).

1. **Option 1: Attribute performance data based on other sources of information.** In other words, a CVE can use information that is not derived directly from the performance data to determine attribution. For example, a health maintenance organization (HMO) may require all its members to choose a PCP soon after they enroll, regardless of whether these members generate any performance data. This list of chosen PCPs can be used to attribute measures of performance to the PCP identified for each patient. Surveys of patient experience may similarly ask patients to identify the provider on which they are reporting their experiences.

   Advantages:

   - High face validity, especially if corroborated by a service-based definition (see below).
   - Does not require patients to receive health care.
   - Easy to explain.

   Disadvantages:

   - Patient self-identification data may not be available for many patients (e.g., enrollees in fee-for-service or PPO health plans). If these data are unavailable, they can be difficult and expensive to obtain.
   - This strategy may only work for certain types of providers (e.g., PCPs) and certain types of performance measures.

2. **Option 2: Attribute performance data based on simple plurality of services (or visits).** This approach requires administrative data such as health plan claims or other kinds of data about services that have been delivered (e.g., records of the number of visits to a provider).[24] Patients can be assigned to the provider who has seen them the greatest

number of times during the measurement period. In the case of ties, patients can be assigned to the most recently seen provider (or assigned to both).

Advantages:

- Necessary data likely to be available for all patients who have received health care during the measurement period.
- Relatively inexpensive.
- Easy to explain.

Disadvantages:

- Patients who have received no health care cannot be assigned. This is an especially concerning problem in the case of quality measures that are based on underused health services (e.g., colorectal cancer screening).
- When patients see many providers, the "plurality" provider may actually only provide a small fraction of the total care received by the patient. Depending on the performance measure in question, responsibility and ability to affect care under these circumstances may be less clear.
- Level of face validity may vary, depending on the performance measure. Plurality attribution may make more sense for primary care performance measures than for measures intended to assess specialty care.

3. **Option 3: Attribute performance data based on "enhanced" plurality of services.** This option includes a family of strategies based on the plurality strategy discussed in Option 2. In addition to the most visits with a patient, "enhanced" plurality strategies include other requirements. For example, such a strategy may require that a patient have at least 50 percent of his or her visits with a provider before making an assignment.[24] Or a strategy may require that the duration of the relationship between a patient and provider be a certain length (measured as the time elapsed between the earliest and most recent services).

   For individual practitioners, such strategies can be devised on a specialty-by-specialty basis when practitioner specialty is known. For some specialties and some performance measures, even a single visit may be enough to allow credible attribution. However, practitioner specialty data may not be available in administrative sources. More complicated strategies, such as those used to determine patient-physician "connectedness," also may be used.[27]

   Advantages:

   - Improves face-validity, relative to simple plurality.
   - Relatively inexpensive.

Disadvantages:

- As more requirements are added before assignment of patients to providers, the number of patients who cannot be assigned will grow.
- Patients who have received no health care cannot be assigned.
- This approach may get methodologically complex and hard to explain.
- This approach may require data that are not commonly available in administrative sources. Such data may be difficult and costly to obtain.

4. **Option 4: Attribute performance data to multiple providers.** It may not be necessary to choose just one provider when assigning a patient (or assigning an episode of care). In fact, sometimes assigning a single patient's data to multiple providers makes sense (e.g., when calculating a measure of coordination of care among providers). An example of a multiple-provider attribution strategy is to assign a patient to every provider who accounts for at least 25 percent of services delivered to the patient during the measurement period.[24] Under this strategy, performance data from a single patient could be attributed to between one and four different providers. Similarly, it may make sense to attribute episodes of care to multiple providers when the actions of each provider affect the measure being applied to the episode (e.g., joint attribution of long-term hip replacement outcomes to the surgeon and to providers of rehabilitation care).

   Advantages:

   - May encourage cooperation between providers.
   - Relatively inexpensive.

   Disadvantages:

   - Providers may be attributed performance data for which they do not accept responsibility.
   - Lack of single-provider attribution may dilute the incentive to improve. Some providers may behave as "free riders," benefiting from the performance improvement efforts of others.

Examples: Attributing performance data to providers

The **Puget Sound Health Alliance** (www.wacommunitycheckup.org) uses different attribution rules for different types of measures. For measures of screening and first contact care, data are attributed to a single PCP for each patient based on a modified plurality algorithm that applies the following ordered rules: greatest number of "evaluation and management" (E&M) visits, highest sum of RVUs ("relative value units" associated with the E&M visits), and most recent service data. Each rule is applied only when the previous rule results in a tie between two or more providers who self-identify as PCPs. However, measures of chronic disease care can be attributed to multiple providers, including both PCPs and non-primary care specialists. For a given measure, all providers in certain specialties with any E&M visits in the past 24 months are attributed patients eligible for the measured service. For example, asthma measures are attributed to PCPs, allergists, and pulmonologists. More detail on this attribution strategy is available in the technical specifications at www.wacommunitycheckup.org/editable/files/July_2009/TechSpecs_CommunityCheckupJul09_final.pdf.

The New York Quality Alliance (www.nyqa.org**)** is using adjudicated health plan claims data to calculate Healthcare Effectiveness Data and Information Set (HEDIS) measures of primary care quality. For patients in an HMO or point of service (POS) product, the CVE attributes patients to PCPs based on the identification supplied by each health plan (since HMO and POS enrollees are required to choose a PCP). However, for patients not enrolled in these products, attribution is based on an "enhanced plurality" strategy: Patients must have a plurality of visits with the PCP during the time period specified in the HEDIS measure, including at least one preventive visit or two E&M visits.

For each provider, **Aligning Forces for Quality-South Central Pennsylvania** (www.aligning4healthpa.org) allows any patient seen at least once in the past year for any purpose (including urgent care) to be sampled for performance score calculation. Therefore, this CVE allows patients to be attributed to multiple providers.

B. What are the options for handling outlier observations?

Outlier observations are performance measure values that are far outside the usual range (e.g., a patient or episode of care for which costs are 20 times the average, or a hospital stay that costs $1). These outlier observations are a critical concern when measuring the costs of care, because within-provider (patient-to-patient) variation in costs can greatly exceed between-provider variation in average costs of care. In other words, when outliers are present, they can greatly increase the average error per observation (discussed in Appendix 2), which can reduce measurement reliability and raise misclassification risk. In addition, outlier observations often reflect data values that are erroneous or that are being incorrectly interpreted.

Options for handling outliers include:

- ***Exclude* the outlier data from calculations of performance scores.** When there are few outliers, this may be a reasonable option.
- ***Change* the values of outlier data so that they are within the range that is usually seen.** Values can be truncated (or "Winsorized") so that the outlier values are replaced by a more commonly-seen value (e.g., the 5th percentile value for low outliers and the 95th percentile value for high outliers).[28-29] To address outliers, it may be advisable to consult a statistician with experience in performance reporting.

C. Will case mix adjustment be performed? (If so, how?)

Case mix adjustment[iv] (also known as "risk adjustment") refers to statistical techniques that "adjust" performance scores to compensate for the characteristics of providers' patients and other factors felt to be beyond providers' control. To see why case mix adjustment might be desirable, suppose Provider A cares for patients who tend to be older than average. If Provider A is rated on patient mortality, Provider A's performance will probably be worse than average. This may be entirely due to Provider A's older patient population. After all, mortality rates increase dramatically as patients grow older. If these differences in patient age are not taken into account, performance reports might systematically mislead patients who are trying to assess the performance of Provider A. **This kind of systematic, predictable misleading information is due to a problem called "statistical bias."**

There is an important difference between statistical bias and misclassification due to chance.[v] If there is a large amount of statistical bias, a provider will have measured performance that is either *consistently* worse than "true performance" (e.g., Provider A) or *consistently* better than true performance (e.g., a provider with younger, healthier patients). On the other hand, if the risk of misclassification due to chance is high, a provider's measured performance may deviate from true performance in a manner that is unpredictable and inconsistent in direction.

This distinction between statistical bias and misclassification risk has important real-world methodological implications. **Getting more observations may reduce the risk of misclassification due to chance, but more observations will not address statistical bias. Only case mix adjustment methods (or a related method called "stratification") can address statistical bias. On the other hand, case mix adjustment is unlikely to improve the risk of misclassification due to chance.**

CVE stakeholders may benefit from understanding a very important point about case mix adjustment: Choosing which characteristics (if any) to include in case mix adjustment is a value judgment. **There is a strong case for adjusting for a patient characteristic when two criteria are met: (1) The patient characteristic is beyond the control of a provider, *and* (2) CVE**

[iv] Case mix adjustment techniques broadly fall into two categories: adjustment based on regression models, and adjustment based on reweighting the data that are used to calculate performance scores. The technical differences between these two approaches are beyond the scope of this guide. The discussion in this section pertains to both case mix adjustment approaches.

[v] See the section on Task #1 for more on misclassification due to chance.

stakeholders agree that the existing overall relationship between the patient characteristic and measured performance is acceptable.

The first criterion seems straightforward for patient characteristics such as age and socioeconomic status (i.e., patient income and education level); a provider will have little or no influence over these characteristics. But what about patient adherence with recommended medical treatments and counseling? Some stakeholders might argue that providers should not be held accountable for patient adherence since it is entirely up to patients to adhere to their recommended health care. Such a point of view seems to be a value judgment and a potentially controversial one. Other stakeholders might argue that providers in fact exert a significant influence over patient adherence, since providers can give patients enhanced self-management support, education, appointment reminders, and other services that improve adherence.

If performance reports are adjusted for patient adherence, this adjustment will prevent the report from incentivizing providers to improve patient adherence. After all, if a provider focuses on improving performance by improving adherence and succeeds, the provider's adherence-adjusted performance will not budge (because the newly adherent patients are effectively held to a higher standard, assuming adherence is reassessed). *On the other hand, not adjusting for patient adherence may incentivize providers to avoid treating nonadherent patients*, especially when providers doubt their ability to improve adherence (or believe that improving adherence will require unrealistic levels of effort and expense).

The second criterion for adjustment—whether the relationship between the patient characteristic and measured performance is acceptable—is fundamentally a value judgment as well. CVE stakeholders may generally accept that older patients will have higher mortality rates than younger patients and believe this will always be the case. By adjusting mortality rates for patient age, a CVE would implicitly accept the mortality "disparity" between older and younger patients. This does not mean that there would be no incentive to reduce mortality for older patients; it just means that all other things being equal, publicly reporting age-adjusted mortality rates will not incentivize providers to reduce the age-related mortality "disparity."

However, when CVE stakeholders find performance disparities unacceptable (e.g., racial or socioeconomic disparities), there are important drawbacks to adjusting for patient characteristics. This is true even though the first criterion for case mix adjustment may be satisfied (as it is for patient race and socioeconomic status, or SES). *For example, adjusting performance for patient SES implicitly accepts the continuation of performance disparities between providers that serve greater and lesser shares of low-SES patients.*

Providers would still be incentivized to improve performance for all patients. But there would not be a systematically different *degree* of improvement incentive for providers serving greater shares of low-SES patients (i.e., a greater incentive that could result in greater improvement of care for low-SES patients and a reduction in the overall performance disparity in a CVE's area). Whether such a disparity-reducing incentive would in fact be created by the performance report would depend on exactly how performance is reported (see the section on Task #6) and on how providers respond to the report. *On the other hand, not adjusting for patient characteristics such as SES could demoralize providers serving low-SES populations and create an unintended incentive to avoid serving vulnerable patient groups.*

One option for balancing the advantages and disadvantages of case mix adjustment is to pursue an alternative approach: stratification of performance results (discussed below). Table 2 presents a framework for thinking through the "value judgments" inherent in deciding which patient characteristics to include in case mix adjustment.

Table 2. Framework for considering which patient characteristics to include in case mix adjustment

Guiding questions	Examples of patient characteristics		
	Age	**Socioeconomic status**	**Adherence**
Is the patient characteristic considered to be beyond control of providers?*	Yes	Yes	Controversial
Is the relationship between the characteristic and performance considered acceptable?*	Yes, for some performance measures (e.g., mortality rates)	Controversial	Controversial
Include characteristic in case mix adjustment?	Yes, for some performance measures	Controversial	Controversial
As an alternative to case mix adjustment, present stratified results?	A reasonable option, but no advantage over case mix adjustment	Yes, may be a good alternative†	Stratification still controversial‡

*The answers to these questions are "value judgments" on which CVE stakeholders can attempt to achieve consensus.

† If stratified performance reports display scores on the same scale in each stratum, then performance disparities will be reported. However, if stratified reports display scores on different scales (e.g., a ranking within each stratum), then stratification will have no advantage over case mix adjustment. Both techniques will make performance disparities appear to vanish.

‡ Like case mix adjustment, stratifying by adherence (i.e., separately reporting performance for "adherent" and "nonadherent" patients) eliminates the incentive to improve adherence.

A complementary discussion of case mix adjustment is available in a separate AHRQ decision guide, *Selecting Quality and Resource Use Measures: A Decision Guide for Community Quality Collaboratives*.[14] Here, we repeat an important limitation about case mix adjustment: Case mix adjustment works by accounting for observable differences in the characteristics of patients. However, not all differences between patients are observable in the data available to a CVE. Therefore, case mix adjustment cannot guarantee that a provider's low (or high) performance is not due to some unobserved patient characteristic.

In general, case mix adjustment is considered appropriate for measures of health outcomes, patient experience, and costs of care. However, case mix adjustment is generally not performed for measures of processes of care (such as checking cholesterol levels in patients with diabetes). This is because process measures generally have restrictive criteria for patient inclusion, and these criteria enforce a kind of uniformity among these patients (at least regarding the need for the measured service). In other words, all patients who qualify for a process measure should be receiving the measured service according to guidelines.

CVEs should consider an important caveat to the "do not adjust process measures" rule. **When aggregating performance data across different data sources, a CVE may want to consider**

performing case mix adjustment based on the *source* of each observation (e.g., **the health plan reporting each observation).** This adjustment accounts for the different ways data sources may collect and report performance data.

Without accounting for data source, a CVE may find that some providers have performance that is higher or lower than others' simply because their data came predominantly from a different source. Adjusting for data source when aggregating multisource performance data is analogous to "standardizing" individual performance measures in the creation of performance composites (see the section on Task #6).

1. **Option 1: Perform case mix adjustment using predetermined methods.** As mentioned in the section on Task #2), case mix adjustment instructions may already be available for performance measures with nationally endorsed specifications.

   Advantages:

   - A CVE can perform case mix adjustment without having to derive a custom case mix adjustment methodology.
   - Use of predetermined methods improves the likelihood that performance will be reported in an accurate category.

   Caveat:

   - If a measure's specifications have been altered by the CVE, then the predetermined case mix adjustment methods will probably not be valid.

2. **Option 2: Perform case mix adjustment using locally derived "custom" methods.** With statistical consultation, a CVE may be able to use the performance data it has aggregated to generate risk-adjusted performance scores.

   Advantages:

   - Allows flexibility in measure specification and in choosing the reference value to which case mix adjusted performance scores can be compared (e.g., the local average rather than a national average).

   Disadvantages:

   - Deriving case mix methodologies can be a complex undertaking, and significant time, resources, and expertise may be required.
   - If insufficient numbers of observations are present, deriving valid case mix adjustment methods may not be possible.

3. **Option 3: Report "stratified" performance.** Stratification involves calculating multiple performance scores for each provider on a given measure. For example, a report might separately display providers' mortality rates for younger and older patients. Or a report might separately display performance for Medicare, Medicaid, and commercial health plan enrollees. Stratification can be an alternative to case mix adjustment because within

each "stratum" (or subset of patients), patients are similar to each other, which reduces the amount of statistical bias. However, like case mix adjustment, stratification can only account for observable patient characteristics.

Advantages:

- Methodologically simpler than case mix adjustment; easier to explain and understand (one can see that "apples are compared to apples, and oranges compared to oranges").
- Enables "fair" comparisons between providers without hiding performance disparities that a CVE would like to reduce.

Disadvantages:

- Sample sizes can become small within each stratum, increasing the risk of misclassification due to chance. This is especially true when trying to account for more than one or two patient characteristics.
- This approach increases the number of scores included in a report, which may make the report more difficult to understand.

Examples: Using stratification instead of case mix adjustment

For clinic and medical group performance on ambulatory quality measures, the **Puget Sound Health Alliance** (www.wacommunitycheckup.org) displays provider performance stratified by insurance (commercial health plan vs. Medicaid enrollees). Natasha Rosenblatt, Data Projects Manager of the Alliance, explained that the stratified reports were added because some clinics that predominantly served Medicaid enrollees "were doing terrific work with a difficult population, but this performance wasn't showing up in the overall results." By comparing the overall performance report with the stratified reports, one can see how stratification reduces the number of observations within each stratum. For example, there are more clinics with no reported performance information in the stratified reports than in the overall reports.

The **Oregon Health Care Quality Corporation** (q-corp.org) is also planning to stratify performance for commercial health plan and Medicaid enrollees. Nancy Clarke, formerly Executive Director of Q-Corp, explains: "Trying to make disparities disappear by adjustment won't help anybody with anything. Stratification shines a light on disparities."

D. What strategies will be used to limit the risk of misclassification due to chance?

Even though some amount of misclassification risk will be present whenever performance reports have more than one provider (or more than one performance category), there are ways to minimize the risk of performance misclassification to a level acceptable to Chartered Value Exchange (CVE) stakeholders. CVEs and community stakeholders will need to determine a rate of performance misclassification that is reasonably acceptable to all parties. Ideally, this rate will

balance the consequences of misclassification with the purposes and expected benefits of performance measurement and reporting.

There is no mathematically or scientifically "best" rate of misclassification due to chance. A 2006 survey found that patients vary widely in their tolerance for misclassification in physician performance reports. Roughly a third of patients thought misclassification risk needed to be less than 5 percent, another third thought misclassification risks between 6 and 20 percent were acceptable, and the remaining third would tolerate levels of misclassification risk between 21 and 50 percent.[12]

There are also important tradeoffs associated with limiting the risk of misclassification due to chance. These tradeoffs include potentially:

1. Reducing the number of providers for which performance can be reported.
2. Reducing the number of measures that can be reported.
3. Reducing the precision with which performance can be reported (i.e., reducing the number of performance categories).

In addition, there may be tradeoffs between *types* of misclassification due to chance. For example, a CVE may want to reduce the probability that providers are reported as low performing when they are actually high performing. One way to accomplish this goal is to report performance using a "zone of uncertainty" (a "buffer zone" that may give the benefit of the doubt to providers just below a performance threshold, which is discussed later in this section). However, this way of reducing the risk of reporting performance in too low a category will *raise* the risk of reporting a provider's performance in too high a category. How a CVE chooses to weigh the risks involved in this tradeoff is a value judgment.

In this section, we present some of the most commonly used options for limiting the risk of misclassification due to chance. These options work by influencing the factors that determine misclassification risk (i.e., the factors listed in Figure 5 in Appendix 2). Many combinations of these and other options may be used by CVEs. But first, we offer some general guidance on approaching misclassification risk.

General guidance to proceeding through the options.

- If possible, negotiate a maximum risk of misclassification due to chance that is acceptable to stakeholders. This may require making some general decisions about the classification system to be used in performance reports. For example, how many performance categories will there be? How will performance thresholds be determined?
- In addition, discuss the *magnitude* of misclassification. For example, if there are four reported performance categories (e.g., a report of stars on a 4-star scale), it might be more acceptable to be off by just one star than to be off by two stars.
- Once performance data have been collected and the reporting format has been decided, calculate the risk of misclassification for providers in the report. This step will require consultation with a statistician who has expertise in performance measurement and reporting.
- If the calculated risk of misclassification is higher than CVE stakeholders can accept, then the following options for limiting misclassification risk may be considered.

1. **Option 1: Exclude *providers* for which the risk of misclassification due to chance is too high from reporting.** Once the risk of misclassification has been calculated for each provider, it is possible to exclude providers with high risks of misclassification.

   Advantages:

   - Limits the risk of misclassification due to chance.

   Disadvantages:

   - May result in the exclusion of many providers, limiting the usefulness of performance reports. See the section titled "Missing data" for options on including providers without reportable performance data in performance reports. In particular, it may be important to understand how patients, providers, and other users of the report will interpret the absence of performance information for certain providers. Will such providers be presumed to have good performance? Poor performance? How will these interpretations affect the goals that the CVE wants to achieve?
   - May result in the exclusion of entire categories of providers (e.g., providers of certain specialized health care services who only manage a small number of patients with a measured clinical condition) from reporting.

2. **Option 2: Exclude *measures* for which the risk of misclassification due to chance is too high for too many providers.** Just as misclassification risk varies from provider to provider, ***misclassification risk may vary from measure to measure***. Those measures for which the risk of misclassification is too high for too many providers can be excluded from public reporting.

   Advantages:

   - Limits the risk of misclassification due to chance.

   Disadvantages:

   - Performance measures that are important to stakeholders may not be available for reporting. This may limit the usefulness of performance reports.

3. **Option 3: Modify the classification system used in the performance report.** We present five general types of options for modifying the classification system. All of these options may require the assistance of a statistical consultant.

   - **Option 3a: Report performance using fewer categories.** For example, move from reporting provider rankings (in which the number of categories equals the number of providers) to broader provider categories such as quartiles.

     Note that at the extreme option of reporting just one performance category, there is zero risk of misclassification due to chance. Provider rankings, which represent the opposite extreme, maximize the risk of performance misclassification due to chance.

Advantages:

- Limits the risk of misclassification due to chance.

Disadvantages:

- Moving to a "coarser" scale of reporting may cause small but possibly important differences in performance to be missed. In other words, when performance categories get big, each category may actually contain many distinct levels of performance.

- **Option 3b: Change the thresholds used for deciding categories.** Without changing the number of reported categories, change the performance thresholds used to decide the performance category in which a given provider will be placed. For example, the definition of a 4-star provider may change from "performance above 75%" to "performance above 90%" on a given measure. Because threshold changes may either decrease or increase the risk of misclassification, recalculating misclassification risk after making these changes is recommended. The *type* of threshold can also be altered. Thresholds can be based on absolute observed performance (relative to some predetermined standard) or on relative performance (e.g., a percentile- or ranking-based approach).

  Changing performance thresholds can have complex effects on the risk of misclassification due to chance. It is possible to simultaneously lower one kind of misclassification risk while raising another. For example, moving from the 75 percent to 90 percent performance threshold may *decrease* the risk that a true 3-star provider is misclassified as a 4-star provider and *increase* the risk that a true 4-star provider is misclassified as a 3-star provider.

  In addition, **thresholds can be based on tests of statistical significance**.[30] Tests of statistical significance compare observed performance to some reference value. This reference value is often, but does not have to be, the average performance of the entire provider population.

  Performance thresholds based on statistical significance have a special property. They automatically limit the risk of one kind of misclassification due to chance: Type I statistical error, or the probability that a provider whose true performance is *equal* to the reference value is misclassified as having performance that is *different* from the reference value. **Statistical significance-based thresholds commonly limit this kind of misclassification risk to 5 percent, but there is nothing special about the 5 percent figure.** Other levels of misclassification risk may be acceptable to CVE stakeholders. A recent survey of patients found that only a minority think that a risk of misclassification below 5 percent is necessary for reporting provider performance.[12]

  One notable drawback of significance-based thresholds is that while they limit one type of misclassification risk (Type I statistical error), they may *increase* Type II

statistical error. Type II statistical error is the probability that a provider whose true performance is *different* from the reference value is misclassified as having performance that is indistinguishable from the reference value. So all else being equal, efforts to reduce Type I statistical error may misclassify more providers as having average performance (i.e., the rate of Type II statistical error will be higher).

Advantages:

- May limit misclassification risk of one type (such as Type I statistical error).

Disadvantages:

- May increase other types of misclassification risk (such as Type II statistical error). Rechecking misclassification risk is advisable.

Example: Basing performance thresholds on tests of statistical significance

The **Puget Sound Health Alliance** (www.wacommunitycheckup.org) reports the performance of clinics, medical groups, and hospitals in three categories: above regional average, at regional average, and below regional average. Providers are classified as "above regional average" or "below regional average" only if tests of statistical significance show that there is a less than 5 percent chance that their true performance is at the regional average. In other words, the probability of misclassifying an average provider as above or below average is limited to no more than 5 percent).

The **Healthy Memphis Common Table** (www.healthymemphis.org) calculates a 95 percent confidence interval around providers' scores on performance measures. For each provider, the *upper limit* of the confidence interval determines which category of performance is reported (i.e., how many stars are reported).

- **Option 3c: Introduce a "zone of uncertainty" around performance cutpoints.**[31] As a provider's performance gets closer to a classification threshold, the risk of misclassification due to chance becomes greater. Using a "zone of uncertainty" (also known as a "buffer zone") typically means giving providers the benefit of the doubt when they are just below a performance threshold by reporting them in the performance category that is above this threshold. This option decreases one kind of misclassification risk: the risk of reporting providers in too low a class. Adjusting the width of the "zone of uncertainty" can limit the risk of this type of misclassification to any value that is desired. However, this approach simultaneously increases the risk of reporting providers in too high a class.

Advantages:

- Reduces one kind of misclassification risk: the risk of reporting providers in too low a performance class.
- May address provider concerns about being misclassified into a category that is lower than their true performance.

Disadvantages:

- Increases another kind of misclassification risk: the risk of reporting providers in too high a performance class.

Example: Using a "zone of uncertainty"

The California Chartered Value Exchange uses a "buffer zone" in determining the performance categories of medical groups (http://opa.ca.gov/report_card/doctors.aspx) in its Doctors and Medical Groups Quality Report Card. This CVE reports aggregated composite scores on technical quality and patient experience using 4-star scales, so there are four categories of performance on each composite measure. Any group whose overall performance is less than 0.5 percent below the next highest performance category is reported in the higher category. This 0.5 percent zone is the "buffer zone."

- **Option 3d: Report "shrunken" performance rather than observed performance.** "Shrunken" performance refers to performance estimates that are produced by special statistical techniques. These techniques are used to adjust the individual provider's observed "raw" scores by borrowing information from the entire population of providers to reduce the likelihood of misclassifying a provider. Names for these techniques include "smoothed estimates," "random intercepts," "hierarchical model estimates," and "empirical Bayes estimates."[32]

  These shrunken estimates work by taking within-provider error into account. When within-provider error is high (e.g., because of low N), shrunken estimates "shrink" performance estimates back toward the mean of the entire provider distribution. When within-provider error is lower (e.g., large N), the shrunken estimates still pull performance back toward the mean, but the amount of this pulling is lower. In other words, higher reliability estimates borrow less from the mean performance of all providers, while lower reliability estimates borrow more (i.e., the mean performance of all providers receives greater weighting in the construction of the shrunken performance estimate). To generate these "shrunken" performance estimates, consultation by a statistician will be needed.

  Advantages:

  - Limits the risk of misclassification.

Disadvantages:

- Providers' own, independent performance is no longer the only thing that determines their performance category. Instead, the performance of the entire provider population plays a role, which may be counterintuitive. Stakeholders who prefer the exclusive use of observed (rather than shrunken) performance may object to this approach.
- "Shrunken" performance can be hard to explain to stakeholders.

- **Option 3e: Use a "mixed" performance classification system that accounts for both reliability and observed performance.** This "mixed" option refers to a family of classification systems that allow the category of *reported* performance to differ from the category of *observed (raw)* performance, depending on the reliability of measurement. First, within-provider measurement error is calculated for each provider and used to create a "margin of error" (just like the "range of uncertainty" shown in Figure). Then, the margin of error is combined with observed performance to see whether, for each provider, this margin of error overlaps a performance category threshold (or potentially more than one threshold when the margin is large). When these overlaps occur, performance can be reported in a different category than observed performance.

  Generally, reported performance will be somewhere close to the middle of the margin of error, even though observed performance may be closer to one end of the margin than the other. For example, a provider with very low observed performance on a measure might still be reported as having average performance. This reporting would occur when the margin of error for this provider is very wide, overlapping the average level of performance. This situation is especially likely when reliability is low (e.g., because of low N). In a sense, the "shrunken" performance option is just one member of this family of "mixed" classification systems. It is advisable to consult a statistician when constructing a "mixed" classification system.

  Advantages:

  - Limits the risk of misclassification due to chance.
  - Simplifies performance reports, potentially making them more patient friendly, because measurement reliability and observed performance are combined into a single reported performance category for each provider.

  Disadvantages:

  - Mixed performance classification systems can be complex to design and hard to explain to stakeholders.

Example: Using a mixed performance classification system

California Hospital Compare uses a "mixed" performance classification system in reporting the performance categories of hospitals. A table describing how this system works is available at the following link: www.calhospitalcompare.org/resources-and-tools/choosing-a-hospital/about-the-ratings.aspx.

The system has five performance categories based on three performance cutoffs. The category of performance reported for each hospital depends on the upper and lower bounds of each hospital's margin of error (rather than just depending on average performance for each hospital).

4. **Option 4: Set a minimum reliability for reporting on each measure.** Setting a minimum reliability is an approach to limiting misclassification risk currently used by some CVEs. Frequently, a minimum reliability of 0.7 is used; but again, the decision on where to set this minimum reliability depends on a value judgment about how much risk of misclassification CVE stakeholders can tolerate. Providers whose reliability is below the minimum on a given measure are excluded from reports on that measure. Because the classification system used in performance reports also determines misclassification risk, *on its own, setting a minimum reliability may not guarantee any particular limit on the risk of misclassification due to chance.* But once a classification system is decided, setting a minimum reliability will limit the risk of misclassification. The amount of risk will depend on the classification system that is decided; *if the classification system is changed, the minimum reliability level may not guarantee the same limits on misclassification risk.*

   Advantages:

   - Coupled with a classification system, limits the risk of misclassification due to chance.

   Disadvantages:

   - If a CVE does not know the classification system that will be used, the range of possible misclassification risks is unknown. Although a minimum reliability of 0.7 is frequently used, this may not limit the risk of misclassification to a level that is acceptable to CVE stakeholders.
   - Reliability is not an intuitively interpretable number, which may make stakeholder consensus difficult to achieve. Misclassification risk is more intuitively meaningful (it is analogous to the risk of convicting an innocent person or acquitting a guilty one; see Appendix 2).
   - Many providers and measures may be excluded.

Examples: Using a minimum reliability criterion

Massachusetts Health Quality Partners (MHQP; www.mhqp.org) and the **California CVE** (http://opa.ca.gov/report_card/doctors.aspx) **both use a minimum reliability criterion** for reporting performance on patient experience surveys. For survey results to be reported, both CVEs require the reliability to be more than 0.7. Because of this criterion, some practices (in Massachusetts) and medical groups (in California) have no reported results on some survey domains.

The Pacific Business Group on Health is combining a minimum reliability criterion with a "zone of uncertainty" reporting approach. Ted von Glahn, Director of Performance Information and Consumer Engagement, says that the reporting effort is aiming to achieve a less than 5 percent rate of provider misclassification due to chance.

Implications of using a minimum reliability criterion

Massachusetts Health Quality Partners (MHQP; www.mhqp.org) **only reports provider performance on a measure when at least 50 percent of all providers meet MHQP's minimum reliability criterion (discussed in the section on Task #3).** This 50 percent requirement means that performance reports contain performance scores on most providers. However, this requirement also means that certain measures generated by MHQP are not publicly reported (even though they may be confidentially reported to the providers).

Similarly, the California Physician Performance Initiative found that of the 17 measures initially tested for public reporting, some measures *did not meet the minimum reliability criterion for virtually any physician in the State*. Therefore, only 10 measures will be included in the performance report being developed by one of the health plan stakeholders (for use by its members).

5. **Option 5: Set a minimum N (number of observations).** Setting a minimum number of observations is also a popular approach to addressing misclassification risk. This approach generally applies the same lower limit on N to all providers and all measures of performance. However, N is not the only thing that determines reliability, and reliability is not the only thing that determines misclassification risk. *Thus, on its own, setting a minimum N does not guarantee that the risk to misclassification due to chance will be acceptable.* In other words, the relationship between N and misclassification risk depends on (1) the properties of the measure, (2) the population of providers, and (3) the performance classification system being used. All other things being equal, a greater N will reduce misclassification risk. But simply specifying a minimum N without calculating the risk of misclassification in the provider population being reported can result in a very high (and unappreciated) misclassification risk.

Advantages:

- Limits the risk of misclassification due to chance. However, just setting a minimum N does not, on its own, determine what this limit is. The amount of misclassification is knowable only when: (1) providers' average error per observation is known, (2) between-provider variation in performance is known, *and* (3) the classification system is decided.
- Intuitive, computationally simple.

Disadvantages:

- Without information on providers' average error per observation, between-provider variation in performance, or the classification system that will be used, the amount of misclassification due to chance is unknown.
- If a CVE sets the *same* minimum N for all measures, this may actually produce *different* levels of misclassification risk for each measure.
- A minimum N may provide false reassurance about the risk of misclassification.
- Many providers and measures may be excluded.

Cautionary Note on Using 25 or 30 Observations as a Minimum N:

Many CVEs and other performance reporting entities have gravitated toward the numbers 25 or 30 as the minimum numbers of observations needed to report a provider's performance score (see below). While these numbers have been widely adopted, they may not be high enough to ensure a level of reliability ($\geq$0.7) that is considered to be adequate to prevent excessive misclassification due to chance. There is no "right" amount of misclassification risk (see the section on Task #1), but CVE stakeholders may benefit from knowing the implications of choosing each minimum N.

A recent paper by Sequist and colleagues presents a helpful set of tables that demonstrate the relationship between reliability and minimum N for ambulatory quality measures in a large sample of Massachusetts primary care practice sites.[33] At a minimum N of 30, only 4 of the 14 measures investigated by Sequist and colleagues achieved a reliability of $\geq$0.7. For some measures, more than 200 observations from each site would be needed to achieve this level of reliability.

Examples: Setting a minimum N

In creating performance reports, most report sponsors use a minimum N.

Organization	Minimum N
CMS Hospital Compare	25 observations (process measures of the technical quality of care)
Oregon Health Care Quality Corporation (q-corp.org)	25 observations (claims-based measures of ambulatory care quality)
Wisconsin Healthcare Value Exchange (www.wchq.org and www.wicheckpoint.org)*	25 observations (claims- and chart review-based measures of hospital care quality); 50 observations for measures of ambulatory care quality
New York Quality Alliance (www.nyqa.org)	30 observations (claims-based measures of ambulatory care quality)
Healthy Memphis Common Table† (www.healthymemphis.org)	30 observations (claims-based measures of ambulatory care quality)
Aligning Forces for Quality-South Central Pennsylvania (www.aligning4healthpa.org)	30 observations (chart review-based measures of ambulatory diabetes care quality)
Leading organizations of the Minnesota Healthcare Value Exchange‡ (http://www.mnhealthscores.org/ and http://www.mnhospitalquality.org/)	30-60 observations (claims- and chart review-based measures of ambulatory quality); 25 observations (chart review-based measures of hospital quality)
Greater Detroit Area Health Council (www.gdahc.org)	50-60 observations (claims-based measures of ambulatory care quality)
Puget Sound Health Alliance§ (www.wacommunitycheckup.org)	160 observations (claims-based measures of ambulatory care quality); 25 observations (measures of hospital care quality)

* For hospital measures in Wisconsin, performance scores are generally still available even when there are fewer than 25 observations (after a mouse click). When these small-denominator scores are displayed, there is a disclaimer that the scores may have low reliability.

† In Memphis, measures of cardiovascular care were excluded from reporting because very few providers had more than 30 observations.

‡ Minnesota Healthcare Value Exchange stakeholders arrived at minimum denominators for ambulatory measures by "statistically eyeballing" the performance data; no formal assessment of misclassification risk was performed.

§ The Puget Sound Health Alliance originally had a requirement of 250 denominator observations for each of its ambulatory quality measures. This minimum N of 250 was based on analyses of reliability. However, there was pushback from clinics whose performance was not being reported due to this minimum N (and also pushback from health plans and employers). After analyses found little difference between a denominator of 250 and 160 in terms of reliability, the minimum N was changed to 160.

6. **Option 6: Report composite performance measures.**[34] Composite performance measures mathematically combine provider performance data across multiple measures. While this approach increases N, the construction of composites has some important caveats. To make the best possible use of composites (especially when making a new composite that has not been previously developed by a national body), consultation with a statistician will be needed. This consultation is particularly important if the composite combines different types of measures with different types of statistical distributions.

Refer to the section on Task #6 for further discussion of how composites can be constructed.

Advantages:

- May limit the risk of misclassification due to chance, depending on the type of composite.

Disadvantages:

- May increase the risk of misclassification due to chance. This paradox can occur with certain combinations of measures. How can this happen? While N may increase when combining individual measures, it is possible for average error per observation to *also* increase because the nature of the observation changes when creating composites. The amount of between-provider variation on composite measures is also likely to differ from the amount of between-provider variation on individual measures.

  Figure 5 (Appendix 2) shows factors related to misclassification due to chance. Based on the figure, if creating a composite measure results in enough of an increase in within-provider error and decrease in between-provider variation, then this composite will have lower reliability than the individual measures. The composite will therefore carry a higher risk of misclassification due to chance. This paradoxical increase in misclassification risk is most likely to occur when combining measures that are negatively correlated with each other (i.e., when performance on some measures is high, performance on the others tends to be low).
- May limit the interpretability of reported performance. For example, it may be harder to know what is meant by a low score on a composite measure of diabetes performance. Does this mean performance is poor on all of the individual measures of diabetes care, or does it mean that performance is good on some but especially poor on others? Reporting composite performance makes it impossible to tell.
- May reduce the usability of performance data to guide provider performance improvement efforts.
- May unintentionally overemphasize certain individual measures and underemphasize others. See the section on Task #6 for further discussion of how this may occur.

7. **Option 7: In the case of physician ratings, report performance for larger provider groupings.** For example, a CVE may "roll up" the performance scores of individual physicians into practice sites or larger physician groups, making these larger organizations the units of reporting. This option is especially important when a CVE is thinking about reporting performance on measures that could be attributed to individual practitioners. When the risk of misclassification on a measure is higher than acceptable for a large proportion of practitioners, reporting performance at higher organizational levels (e.g., practice sites, groups, hospitals) is another way to increase N for each provider reporting unit.

Advantages:

- May limit the risk of misclassification due to chance, relative to reporting the performance of lower levels of organization.

Disadvantages:

- May increase the risk of misclassification due to chance, relative to reporting the performance of lower levels of organization. As with reporting composites, this paradox can occur because the nature of the observation changes when reporting on aggregations of providers. Although N may increase, the average error per observation may also increase, and between-provider variation may decrease. This paradoxical result can occur, for example, when the performance scores of individual practitioners are negatively correlated (i.e., when some practitioners do well, the others tend to do poorly).
- May limit the usefulness of performance reports to patients and other stakeholders who want performance data on individual practitioners.
- May unintentionally mask good (or poor) performance by individual practitioners or other subunits of provider organizations.
- May dilute individual practitioners' accountability for performance.

Examples: Reporting performance at higher levels of provider organization

Massachusetts Health Quality Partners (MHQP; www.mhqp.org) **reports performance on HEDIS measures at the physician group level.** The minimum group size is three physicians. Because of this criterion, solo and two-physician practices have no reported HEDIS results unless they are reported within a larger group. On the other hand, the maximum risk of performance misclassification is reduced.

The **Greater Detroit Area Health Council** (www.gdahc.org) reports performance on measures of ambulatory care quality (mostly HEDIS measures) at the physician organization level. There are 16 total physician organizations in the GDAHC's reports, some with thousands of physicians. Reporting at this level results in measure denominators far in excess of GDAHC's minimum N (50-60 observations).

8. **Option 8: Report performance over a longer time period ("rolling average" performance).** When the risk of misclassification on a given measure is higher than acceptable for a large proportion of physicians, reporting performance data accumulated over a longer period is another way to increase N for each provider reporting unit. For example, instead of reporting performance data from just the most recent available year, a CVE may report performance data aggregated over the most recent 3 years. In doing so, a CVE may decide to weight the most recent year's performance more heavily in the "rolling average" or weight each year's performance equally in the calculation. Whether this approach reduces misclassification risk depends on whether "true" performance is

stable over time. If "true" performance is changing (maybe because a provider is implementing an improvement strategy), then this approach may paradoxically increase the risk of performance misclassification (because the provider's older performance does not accurately reflect current performance).

Advantages:

- Reduces the risk of misclassification due to chance if "true" performance does not change over time.

Disadvantages:

- May increase the risk of misclassification if "true" performance changes over time. To determine whether true performance is changing and to address this issue, consultation with a statistician may be needed.
- May limit the usefulness of performance reports to patients and other stakeholders who want only the most recent performance data.

9. **Option 9: Include more data sources for a measure.** Including performance data by aggregating data from a greater number of sources—such as multiple commercial plans, Medicare, and Medicaid—is another way to increase N for each provider reporting unit (as discussed in the section on Task #3). Aggregated multipayer data not only reduce the risk of performance misclassification due to chance, but also may reflect the care delivered to a broader patient population (compared with data from just one payer). However, the issues mentioned in the section on Task #3 should be addressed when aggregating multipayer data. Care must be taken to avoid combining data in ways that are not valid (i.e., combining data without ensuring that the data have the same interpretation across all sources).

   Advantages:

   - May reduce the risk of misclassification due to chance.
   - May produce a fuller picture of provider performance across a broader patient population.

   Disadvantages:

   - May increase the risk of misclassification due to chance. As with reporting composites, this paradox can occur because the nature of the observation changes when aggregating performance data generated by different patient populations. Although N may increase, the average error per observation may also increase, and between-provider variation may decrease.
   - Data aggregation across sources creates the possibility of introducing statistical bias when data do not have the same interpretation across all sources (i.e., when data are not combined in a valid way). See the section on Task #3 for guidance on data aggregation. In addition, case mix adjustment methodologies can be used to guard against increasing statistical bias, discussed in the section on Task #5.

10. **Warning about a pitfall: the "finite population correction"** (also known as the "finite population sampling model"). The finite population correction refers to the practice of reporting a lower amount of uncertainty in a performance estimate by incorporating information about the overall size of the patient population from which a data sample is drawn. The practical argument in favor of this technique is: "If I know a provider took care of 10 patients in a year and I sample all 10 patients for a performance measure, then I know the provider's score in that year with complete certainty. So I don't have to worry about misclassification risk, even though the sample size is only 10 patients." A similar argument can be made to reduce the amount of reported uncertainty in performance estimates based on samples that are less than 100 percent of a provider's overall patient population (e.g., 80%, or 50% of a provider's patients).

    For the purposes of public reporting, the finite sample correction should not be used. When patients use performance reports to choose a provider, past performance matters *only* because it gives some indication of what kind of care a patient will receive in the future (with some degree of uncertainty, of course). Past performance would matter on its own only if a patient had a time machine that allowed him or her to actually receive care that happened in the past (i.e., the care that generated the data used to calculate performance scores).

    Because every provider has a theoretically infinite population of future patients, the finite sample correction is likely to mislead patients who use public reports of provider performance. Small sample sizes, no matter how completely they capture a provider's past patient population, are likely to produce performance estimates that have probabilities of misclassification due to chance. For example, a performance report based on 100 of a provider's patients out of a total population of 1,000 will have *lower* misclassification risk than a report based on all 10 of a provider's patients. A more technical explanation of what the finite sample correction is and why it should not be used in performance reporting is available in a paper by Elliott, Zaslavsky, and Cleary.[35]

    Advantages:

    - There are no real advantages. The finite sample correction *appears* to reduce the risk of misclassification due to chance. But this is a mirage: Misclassification risk regarding future performance is not reduced.

    Disadvantages:

    - High likelihood of covering up true misclassification risk. Reporting past performance that is not a good predictor of future performance is likely to mislead patients and may alienate providers.

Decisions Encountered During Key Task #6: Creating Performance Reports

Creating a performance report involves decisions about methods for calculating and categorizing performance scores as well as other, equally important decisions that affect the usability (or evaluability) of the report. A way of distinguishing these two sets of considerations follows:

Usability (or evaluability) considerations focus on the information a user can extract from a performance report. The way performance data are displayed may confuse patients. When this happens, patients may be unable to extract any information whatsoever. Or they may misinterpret data and therefore be misled by the report. Examples of usability decisions include:

- Should a report contain many performance categories or only a few?
- Should providers be displayed in order of performance ranking or in some other kind of order (e.g., alphabetical order)?
- Should numbers indicate performance, or should some other kind of symbol be used (e.g., star ratings)?
- How should the concept of statistical uncertainty be displayed in order to maximize public understanding?
- How many measures should be reported?

These usability decisions can draw guidance from studies that have investigated which kinds of data displays are most understandable to patients. Separate AHRQ reports by Hibbard and Sof7aer provide guidance on usability and evaluability decisions.[1-2]

The methodological considerations covered here focus on whether the information in a performance report might be misleading, *even when this information is understood perfectly by the patient.* In other words, even if a performance report is so clear that patients and providers can extract and understand all the information it contains, this performance information may contain fundamental problems. Providers who are truly higher performing may be reported as lower performing, and vice versa. The degree to which these problems might be present depends on the available data and the methodological decisions a Chartered Value Exchange (CVE) makes.

In creating performance reports, decisions about usability and other methodological issues are linked to each other. The desirability of each methodological option may change, depending on decisions about usability, and vice versa. For example, a CVE may initially make a usability-based decision to create a performance report that ranks providers in five performance categories on two composite measures of performance. However, once the performance data are computed into performance scores for providers, the CVE may find that the number of misclassified providers is unacceptably high. There also may be methodological problems with the composites (e.g., the individual measures may not "agree" with each other—and therefore cannot create a "coherent" composite measure).

Faced with methodological issues, a CVE may want to compromise between usability and methodological considerations. Providers might be reported in only three categories instead of five, and eight individual measures might be reported instead of two composites. But these

changes may make the report more confusing to patients. **Striking the right balance between usability and other methodological considerations is a "value judgment" that may be negotiated among CVE stakeholders.**

A. Will performance be reported at single points in time, or as trends?

Once provider performance scores are calculated, these scores can be reported in many different formats. This section and the following two sections illustrate the methodological tradeoffs a CVE may encounter when choosing among some commonly discussed ways of reporting performance. These tradeoffs can apply regardless of which performance measure is being reported.

Here, we discuss reporting performance at a single point in time or reporting trends. These two options are not mutually exclusive, and they can be combined in a performance report.

1. **Option 1: Report performance at a single point in time ("achieved performance").** For each provider included in a report, a CVE may report a performance level representing care delivered over a single period. The period is usually chosen to be as recent as possible. If provider performance is not changing, reported performance may predict the kind of future performance a patient is likely to receive.

   Advantages:

   - The reporting period can be lengthened to help deal with misclassification risk (the "rolling average" approach, discussed in the section on Task #5).

   Disadvantages:

   - If provider performance is changing, "achieved performance" may be misleading due to the lag between the time health care is delivered and the time performance is reported. In other words, past performance may not accurately predict the future performance a patient is likely to receive.

2. **Option 2: Report performance change over time ("performance trends").** Performance change is rarely reported on its own, but performance change can be added to reports of "achieved performance."

   Advantages:

   - If provider performance is changing, then reporting which providers are improving (or not) may enable patients to better predict the kind of performance they are likely to receive from a provider.

Disadvantages:

- Reporting performance change may increase the complexity of methods for dealing with misclassification risk (see Introduction) and performing case mix adjustment (see section on Task #5). Consultation with a statistician will be necessary.
- It may not be possible to adequately separate changes in "true" performance from random variation in measured performance (caused by chance alone). This may result in unacceptably high risk of misclassification due to chance.

B. How will numeric performance scores be reported?

As in the preceding section, the following options are not mutually exclusive. A performance report can simultaneously use combinations of these strategies.

1. **Option 1: Report numeric performance scores.** An example of a numeric performance score is the actual percentage of patients receiving a measured service. This numeric score is inherently meaningful. Other numeric scores, such as average ratings on a patient satisfaction survey, may not be inherently meaningful.

   Advantages:

   - Provides detailed performance score data to patients. For example, a patient will be able to see that one provider delivers a measured service to 80 percent of patients while another only delivers the service to 79 percent.

   Disadvantages:

   - The performance classes implied by numeric scores may lead to unacceptably high rates of misclassification due to chance (see Introduction). In the 80 percent versus 79 percent example, this 1 percentage point difference may be almost entirely due to chance. But reporting numeric scores may mislead patients to believe that the provider with the 80 percent observed score has a higher "true" score than the provider measured at 79 percent.

2. **Option 2: Report performance scores with a representation of measurement error.** Measurement error can be represented as a numeric range of uncertainty (often a 95% confidence interval). Alternatively, measurement error can be represented in a categorical fashion. For example, performance scores with high measurement error might be marked with a special symbol.

   Advantages:

   - Still provides detailed performance score data to patients, but may also convey a sense of the range over which "true" performance is likely to be located.
   - Even if no range of uncertainty is included, gives "fair warning" to patients through the use of special symbols to mark scores with high measurement error.

Disadvantages:

- Numeric ranges of uncertainty can be very difficult to understand, even for experts.[vi]
- Ranges of uncertainty may not communicate the right information to patients who are trying to compare providers. Numeric ranges of uncertainty based on statistical significance are only valid for comparing one provider's performance to a fixed (nonrandom) benchmark. However, a comparison between two providers is a comparison between two random variables (because measurement error is present for both providers). Uncertainty about the performance difference between every pair of providers is what determines the chances that these providers are misclassified relative to each other. Without a series of charts showing the range of uncertainty about performance differences between all possible combinations of providers, users cannot know the likelihood of misclassification (even if patients understand the report and use it exactly as instructed). For further explanation on this point, it may be advisable to consult a statistician.
- Despite the instruction, "Do not use this report to make comparisons between providers," patients may still compare providers. If comparisons are still made, the effective risk of misclassification may be unacceptably high.
- Patients may not understand what is meant by a special symbol of measurement error. They may ignore this special symbol and be misled about relative provider performance.
- There is no "best" range of numeric uncertainty to display. The 95 percent confidence interval, although conventional, is essentially the result of a value judgment.

3. **Option 3: Report "shrunken" performance scores.** "Shrunken" performance refers to performance estimates produced by special statistical techniques that incorporate measurement error into the performance score itself. When within-provider error is high (e.g., because of low numbers of observations), shrunken estimates "shrink" performance scores back toward the average of the entire provider distribution. When within-provider error is lower, the shrunken estimates still pull performance back toward the mean, but the amount of this pulling is lower.[vii]

   Put another way, each shrunken performance score is a weighted average of each provider's performance and the average performance of all providers. When there is high uncertainty about an individual provider's score, the average performance of all providers is more heavily weighted. When there is less uncertainty about a provider's score, the average performance of all providers is less heavily weighted (so the provider's shrunken score is close to the raw score). Shrunken performance scores and a related strategy

[vi] See reports by Hibbard and Sofaer for guidance on whether patients can generally understand such ranges of uncertainty.[1-2]

[vii] Technical note: There is increasing interest in shrinking performance scores not to the overall mean, but to a stratified mean based on a relevant stratifying variable. For example, Dimick and colleagues have shrunken mortality rates for selected procedures to the corresponding volume-stratified mean, given that the best *a priori* estimate of a hospital's performance (in the absence of actual data) is based on its procedure-specific volume.[36]

called “mixed performance classification” are discussed in more detail in the section on Task #5.

Advantages:

- Relative to reports of raw performance scores, shrunken scores may be less likely to mislead patients about relative provider performance. Because shrunken scores incorporate uncertainty about provider performance and about the entire provider population, shrunken scores provide better predictions of future provider performance.

Disadvantages:

- Generating shrunken performance estimates is methodologically complex and can be difficult to explain. Stakeholders may not understand why the performance reported for a given provider incorporates information about the entire population of providers in a report.

4. **Option 4: Report provider rankings.** Ranking can be done by ordering providers from highest to lowest based on their performance scores.

   Advantages:

   - Easy to understand; facilitates comparison between providers.
   - High degree of detail. For example, patients can see which provider was ranked seventh and which was ranked eighth.

   Disadvantages:

   - The rate of provider misclassification due to chance is likely to be unacceptably high because ranking maximizes the number of reporting categories: each rank defines a category. As a rule, the more reporting categories are included in a report, the higher the misclassification risk (see “classification system” in Appendix 2).

C. How will performance be categorized?

While numeric performance scores and performance rankings implicitly categorize provider performance (i.e., by allowing comparisons between providers, with each score or ranking constituting a “category), strategies for explicitly categorizing performance are also common.

1. **Option 1: Report categories of performance based on national benchmarks.** For example, a CVE might report local providers as having performance in categories defined by the 25^{th}, 50^{th}, and 75^{th} percentiles of national performance.

   Advantages:

   - Using benchmarks enables comparison of local provider performance relative to national performance.

- Reporting performance in a small number of categories may reduce the risk of misclassification due to chance (see "classification system" in Appendix 2).

Disadvantages:

- If nearly all local providers are in the same category relative to national performance (e.g., all are above the 75th percentile), then the report will not be useful to patients in choosing among local providers. On the other hand, it may be reassuring to patients that all local providers are "good enough" on a given measure (assuming providers are indistinguishable because they are all high performers).
- National benchmarks may not have the intended meaning if measure specifications have been locally modified (see section on Task #2).
- If categories are too wide (i.e., include a broad range of scores), then meaningful performance variation may be hidden.

2. **Option 2: Report categories of performance based on local benchmarks.** For example, a CVE might report local providers as having performance in categories defined by the 25th, 50th, and 75th percentiles of *local* performance. Under this system, there will always be some providers reported in the highest category of performance and some reported in the lowest.

   Advantages:

   - By always including some providers in each performance category, increases the likelihood of providing useful information to patients choosing among local providers.
   - May motivate performance competition between local providers (*however*, if this harms professional relationships that benefit patients, it may not be desirable).

   Disadvantages:

   - May make it difficult to compare local provider performance to national benchmarks.

3. **Option 3: Report categories of performance based on tests of statistical significance.** Tests of statistical significance compare each provider's observed performance to some reference value. This reference value is often, but does not have to be, the average performance of the entire provider population (local or national). Statistical significance-based thresholds commonly use a 5 percent "level of significance" (or "95% confidence"), but there is nothing special about the 5 percent figure. The level of statistical significance that is acceptable to CVE stakeholders is a value judgment that can be negotiated among the stakeholders of each CVE. The section on Task #5 explains how statistical significance relates to misclassification risk.

   Advantages:

   - May limit the number of providers who, due to chance alone, are misclassified as having performance that is different from the reference value (usually the mean).

Disadvantages:

- When providers' true performance is different from the reference value, this approach may *increase* the number of such providers who are misclassified as having performance that is the same as the reference value. In other words, more truly high- or low-performing providers will be reported as having average performance.
- This approach may result in categories that are too wide, especially if only three categories are reported (e.g., above average, average, and below average). Meaningful performance variation may be hidden.

D. Will composite measures be used?

Composite measures (also known as "summary measures") combine the performance data from two or more individual performance measures into a single performance score. For example, a provider's performance on four individual measures of diabetes care might be combined into a composite measure of "overall diabetes care."

A separate AHRQ decision guide titled *Selecting Quality and Resource Use Measures: A Decision Guide for Community Quality Collaboratives* provides a complementary discussion that defines composite measures in more detail and describes their possible uses more broadly.[14] Here, the discussion focuses on key methodological decision points regarding composite measures for public reporting:

- Will composite measures be used?
- If composites will be used, which individual measures will be combined?
- For a given collection of individual measures, exactly how will these measures be combined? In other words, how will the composite measure be constructed?

The following options illustrate the tradeoffs involved in making these decisions. These options are not mutually exclusive and may be chosen in various combinations. The first question is considered below, and the second and third questions are presented in the following two sections.

To avoid redundancy, we present only the advantages and disadvantages of the "yes" answer to the question, "Will composite measures be used?"

1. **Option 1: Report composite measures of provider performance.** The alternative to this option is to only report performance on individual measures. Relative to reporting performance on individual measures, creating and reporting composites has advantages and disadvantages.

   Advantages:

   - Compared to reports of a large number of individual performance measures, reports of a small number of composite measures may be less overwhelming to patients who are trying to choose a provider.[viii]

[viii] See reports by Hibbard and Sofaer for guidance on which kinds of reporting formats might be preferable for purposes such as helping patients choose providers.[1-2]

- Use of composites may reduce the risk of performance misclassification due to chance (see section on Task #5).

Disadvantages:

- The inherent meaning of individual performance measures may be lost. For example, an individual measure score such as "75% of diabetic patients receive lipid screening" has clear clinical meaning. But the meaning of a composite score such as "75% on overall diabetes quality" is unclear. This concern is less important when the individual measures themselves have no clear inherent meaning (e.g., measures of patient experience).[ix] This concern may also be less important when categories of performance will be reported, rather than numeric performance scores.
- By reducing the amount of data detail, performance reports may be less useful to providers who are trying to improve.
- Patients with particular health conditions may care about specific individual measures. Presenting these measures as composites means that these patients will not be able to see the individual measures.
- Composite scores may be very sensitive to exactly which measures are included and how they are combined.[37] There is no single "right" way to make composites. The best choice for many decisions about composite construction will be uncertain, even when CVE stakeholders have agreed on the value judgments of performance reporting. If sensitivity analysis (redoing the performance report using different but justifiable methods; see item G, below) reveals that scores change dramatically when alternative composite construction strategies are used, then reporting composite measures may mislead patients. Patients may not know that other possible composite constructions would produce different results.
- Some types of composites may unintentionally overemphasize certain individual measures and underemphasize others. This can happen when one measure has too much weight (see discussion of weighting in item F, below).

E. If composite measures will be used, which individual measures will be combined?

Once a CVE has decided to use composite measures, the next methodological question is which individual performance measures will be combined into a composite measure.

1. **Option 1: Choose individual measures for inclusion in a composite based on whether they statistically "belong together."** Each composite measure contains two or more individual measures. But which measures should be included in a composite? One way to decide is to use special statistical techniques to let the data decide which measures to include. These techniques work by looking for sets of measures that are "correlated" or associated with each other: when a provider does well on one of these measures, the provider tends to also do well on the others. Composites that are constructed in this way are called "reflective" or "latent" composites.[38] **A statistical technique known as**

[ix] Note that many patient experience (or patient satisfaction) survey results are reported as composite measures.

"factor analysis" is a common approach used to identify the measures included in these composites.

Advantages:

- Using statistical procedures to select the measures that will go into each composite is a relatively automatic process. However, consultation with a statistician may be necessary.
- There is extensive precedent for this methodology: Such composites are the most common way to present data from patient experience surveys such as CAHPS® (Consumer Assessment of Healthcare Providers and Systems).[39]
- Individual measures within a composite will be "correlated." When a provider does well on one of these measures, the provider will also tend to do well on the others.

Disadvantages:

- This methodology may result in composite measures that do not make intuitive clinical sense. For example, four individual diabetes measures may be available to a CVE. It might make clinical sense to expect these four measures to form a composite. However, if statistical techniques are used to determine which measures will be included in which composite, these four measures could end up in two or more composites (where they might be combined with measures of depression and cancer screening).
- This methodology may not identify a composite for every individual measure. Some individual performance measures may not be correlated with the others. These "orphan" measures can be reported individually or excluded from performance reports.
- This approach relies on complex statistical methodology that may be difficult to explain.

2. **Option 2: Choose individual measures for inclusion in a composite based on nonstatistical judgment.** With this option, a CVE uses its own judgment (clinical or otherwise) to choose which measures to include in a composite. For example, a CVE may decide to make composites that include all measures for a health condition (e.g., all measures for heart disease or all measures for cancer screening or preventive care). The Apgar score for newborns is a commonly used example of a clinical composite that combines a variety of vital signs and physical findings.[40] In the development of such composites, statistical correlation between the constituent individual measures is a secondary concern. Composites that are constructed in this way are sometimes called "formative" composites.[38]

   Advantages:

   - When they are based on clinical judgment, these composites make intuitive clinical sense.

Disadvantages:

- CVE stakeholders may not agree on which individual performance measures belong in which composite.
- The individual measures within a composite may not be correlated, which is especially likely when trying to create a single "global" composite that combines all measures.[41-42] Worse, individual measures may be *inversely* correlated, so that high performance on one measure tends to predict low performance on another. When the measures within a composite are not correlated, patients will not be able to see when a provider is truly better on one kind of measure than on another. For example, a provider may be very good at delivering colorectal cancer screening but not as good at delivering cervical cancer screening. If a clinical "cancer screening" composite combines these measures, this difference in performance will be masked. A CVE can check composites for "internal consistency" to see how well the constituent measures are correlated with each other.[x]
- If the individual measures within a composite are not statistically correlated with each other, the risk of performance misclassification due to chance may increase.

3. **Option 3: Choose composite measures that have been endorsed by a national body.** The National Quality Forum (NQF) has endorsed a small number of composite measures. These include composites for measuring inpatient quality of care that were developed by AHRQ.[43] Patient experience surveys also generally include instructions on how data from individual survey items should be combined into specific composite measures.[39]

   Advantages:

   - Documented rationales and usage advice for these measures may be available.

   Disadvantages:

   - Endorsed composites may not cover all the measures a CVE would like to report.
   - A CVE may not have access to all the measures included in an endorsed composite.
   - Even though endorsed composites may be internally consistent in national data, these composites may not be internally consistent within the performance data being reported by a CVE. In other words, individual measures may behave differently in a CVE's local area. In this case, nationally endorsed composites may mask differences in performance within a composite (as discussed in Option 2 earlier in this section).

F. How will each composite measure be constructed from a given set of individual measures?

Once a CVE has determined which individual performance measures will be combined into a composite measure, the next methodological question is exactly how the composite measure will

[x] A special statistic called "Cronbach's alpha" is a common way of checking internal consistency. This technique will be familiar to statisticians.

be constructed from the individual measures. There are many options for calculating composite performance from performance data on a given set of individual measures.

1. **Option 1: Combine performance data from individual measures using a weighted average approach.** Once a CVE has identified the measures that will go into a composite, data from these measures can be combined in many different ways. "Weighted averaging approaches" refers to a large family of specific strategies that multiply scores on individual measures by a weight and then take the average of these weighted scores. **In creating such a composite, a CVE will need to specify the following:**

    - **The weights that will be used** (i.e., how much each individual measure will matter within a composite). There is no single "best" weighting strategy. When the measures included in a composite are identified based on whether they statistically correlate with each other, the same kinds of statistical techniques can also determine how much weight to give each measure. When measures are included in a composite based on nonstatistical judgment, a common strategy is to give more weight to the measures for which more observations are available. However, many other weighting strategies can be used: equal weighting, weighting based on local health priorities, etc.

      The important thing is to purposefully choose and understand the effects of the weights that are used. Composite measures *always* weight their constituent measures, implicitly or explicitly. If a CVE does not explicitly consider its weighting strategy, then unintended results may occur. For example, if measures are weighted by their numbers of observations and one measure has many more observations than the others, then this one measure will dominate the entire composite. If a CVE does not intend for one measure to dominate the composite, then the composite may mislead patients who believe the composite reflects *all* its constituent measures (i.e., the composite measure will have low validity).

    - **The way measures will be standardized.** Some measures of performance have higher "degrees of difficulty" than others. A measure's degree of difficulty is generally felt to be higher when its average performance score is lower. The reasoning is that if *all* providers have low performance on a measure, the measure must be difficult. To avoid penalizing providers for having more observations on measures with high degrees of difficulty, the individual measures can be "standardized" so that they have equal average scores. Other standardization techniques are also possible, such as using the exact same measure weighting scheme for every provider in a report.[43] From a mathematical perspective, standardizing the measures within a composite is no different from performing case mix adjustment; the kinds of techniques that can be used are the same. However, unlike case mix adjustment, standardizing the measures within a composite is unlikely to cause controversy. (For weighted average composites, standardization is *necessary* to avoid unintentional systematic performance misclassification.)
    - **The way missing data will be handled** in computing composite scores. When a provider has no data on an individual measure, this can affect the calculation of a composite that contains this measure. The section on Task #4 discusses ways a CVE

can deal with missing data. If a strategy for handling missing data is not identified, a report can unintentionally report misleading composite scores. Consultation with a statistician is advisable.

Examples of weighted average approaches are available in published papers,[34, 37, 42, 44-46] and several are summarized in the AHRQ decision guide by Romano, et al.[14]

Advantages of weighted average approaches:

- The composite score takes all its constituent measures into account.
- Weighted averages are good conceptual fit for health conditions such as screening and chronic disease (where imperfect care is probably better than no care at all).

Disadvantages of weighted average approaches:

- Not a good conceptual fit for sets of measures in which one failure is clinically equivalent to multiple failures (e.g., a breach in operating room sterility).

Example: Reporting weighted average composite measures

For acute myocardial infarction, the **Wisconsin Healthcare Value Exchange** (www.wchq.org and www.wicheckpoint.org) reports hospital performance on a composite measure that combines seven individual process measures with one measure of patient survival. The composite quality measure is created in two steps. First, the seven process measures are averaged, weighting each measure by its denominator. Second, the "process composite" created in the first step is averaged with the patient survival measure, with the process composite having seven times the weight of the survival measure. Of note, the process and survival components of the overall composite are calculated differently. The process component is a performance rate between 0 and 1, but survival is expressed as a ratio of observed-to-expected survival events (a ratio that may exceed 1). Combining measures based on different units (and with different scales of measurement) complicates the interpretation of the weighting scheme.

Similar composite measures are also reported for hospital quality of care for pneumonia and heart failure.

2. **Option 2: Combine performance data from individual measures using an "all-or-none" approach.**[47] "All-or-none" performance composites start by giving a score of one for each patient who receives satisfactory performance on *every* measured service included in a composite. But if a single service was not delivered, the all-or-none composite score is zero for that patient. The all-or-none performance score for a provider is the number of ones divided by the number of patients. For example, suppose a composite includes four measures. Only a patient for whom performance is satisfactory on all four measures will count in the numerator of an all-or-none composite. In other words, the all-or-none composite measures the percentage of patients who receive

"perfect" care. This means that all-or-none composites treat "almost perfect" performance for a given patient the same as the lowest possible performance. Somewhat less stringent variations of all-or-none composites are also possible.[37]

Advantages:

- Easy to explain.
- In situations where one failure produces the same result as multiple failures (such as a breach in sterility in an operating room), may be clinically meaningful.[48]
- May encourage providers to design system-level strategies for delivering all necessary care.
- When performance on individual measures is already high, will result in lower performance scores, potentially motivating further improvement efforts by providers.[47]

Disadvantages:

- Interpretation of performance scores may be unclear for all-or-none composites in most clinical situations (e.g., composites for diabetes or screening). If a provider has an all-or-none score of 40 percent, it is impossible to tell whether the remaining 60 percent of the provider's patients are receiving almost perfect care or very poor care. Therefore, important differences in provider performance may be masked by all-or-none composites.
- All-or-none composites may unintentionally encourage providers to "give up" on a patient for whom there is a failure on just one measure within the composite.

Examples: Reporting all-or-none composite measures

Leading organizations of the **Minnesota Healthcare Value Exchange** (www.mnhealthscores.org/ and www.mnhospitalquality.org/) report "all-or-none" composite measures for both ambulatory and hospital quality of care. The "all-or-none" approach was selected because it was felt to be easy to explain to patients and clinicians and to represent a more comprehensive view of a condition or episode. In addition, this approach was chosen because it enabled more providers with smaller volumes of patients to be included in public reporting. Also, there was a larger amount of between-provider performance variation on these composites (relative to individual performance measures). The measures included in each composite were chosen on the basis of clinical, nonstatistical judgment (see Option 2 earlier in this section). These composite measures enjoy stakeholder buy-in.

The **Wisconsin Healthcare Value Exchange** (www.wchq.org) also reports an "all-or-none" composite measure for "diabetes optimal testing" in the ambulatory setting.

G. What final validity checks might improve the accuracy and acceptance of performance reports?

By checking the validity of performance reports before they are made public, CVEs may improve their acceptance by key stakeholders. Here are some final checks a CVE may consider performing.

1. **Assess and report the risk of misclassification due to chance.** Misclassification is unavoidable in performance measurement and reporting. As discussed in the Introduction, the risk of misclassification due to chance can be assessed for each provider included in a performance report. For some types of measures, each provider in a report can, in theory, have a different probability of performance misclassification.

   Advantages:

   - The degree to which a performance report could misrepresent provider performance and mislead patients will be known.

   Disadvantages:

   - Will require consultation with a statistician.

   Implication:

   - The overall amount of misclassification due to chance that is actually found in the report may be higher than the allowable amount negotiated by CVE stakeholders (if a level was negotiated earlier, as suggested in the section on Task #1). In this case, a CVE may consider renegotiating the maximum acceptable level of misclassification or choose one of the options discussed in the section on Task #5.

2. **Gather feedback from the providers in the report and make corrections.** Before releasing a performance report to the public, a CVE may give providers a confidential preview of how their performance will be reported. These providers can also be given a mechanism for responding to the CVE with their questions and concerns, and the CVE may use provider feedback to make corrections to the performance report.

   Advantages:

   - May uncover previously unknown problems with data quality. A CVE may be able to address these problems prior to publication of the final performance report.
   - May enhance provider buy-in.
   - May create an incentive for providers to create more accurate data for performance measurement (e.g., an incentive to submit more accurate billing codes to health plans).
   - Depending on the level of data detail available to a CVE, can include information that might be useful to providers seeking to improve their performance. For example, if a CVE has access to patient-level performance data, the CVE may be able to give each

provider a list of patients who have not received a measured service (e.g., a list of patients overdue for cervical cancer screening).

Limitations and caveats:

- A CVE may not have access to all the data a provider might want to see (e.g., a list of patients who were included in a performance measure).
- Some providers may not respond to requests for feedback.
- If data problems are uncovered, addressing these problems can consume time and resources.

Examples: Gathering feedback from providers

- The **New York Quality Alliance** (NYQA; www.nyqa.org) has not yet produced a public report of provider performance, but the NYQA has produced confidential performance reports in preparation for public reporting. In order to be compliant with New York's Patient Charter, the NYQA has instituted a correction loop that allows physicians, through a secure Web portal, to correct the patient-by-patient claims data used to calculate Healthcare Effectiveness Data and Information Set (HEDIS) quality measures. This correction process involves uploading clinical notes to support the requested corrections. The Web portal also allows physicians to confirm whether they provide primary care.
- **Massachusetts Health Quality Partners** (MHQP; www.mhqp.org) **gives each physician group a preview of the group's scores on HEDIS measures.** If the physician groups think these scores are inaccurate, they can make an appeal. However, because MHQP uses a distributed data model in which health plans calculate the measures, MHQP does not know which patients are included in the HEDIS measure scores. Therefore, if a physician group requests a HEDIS score correction, the group must communicate directly with the health plans. MHQP staff report that over time, these communications with health plans have encouraged physician groups to submit more accurate billing codes (since these are the basis for the HEDIS measures).
- When including a provider in a report for the first time (or including a new performance measure), the **Oregon Health Care Quality Corporation** (q-corp.org) performs a confidential round of reporting to providers before starting public reporting. This step "lets off steam" and allows an accuracy check. For subsequent public reports, providers can still check their performance data on a patient-by-patient basis using a secure Web site. Finally, each provider is given a one-time chance to opt out of public reporting, if needed, to sort out why the reporting is not working in that setting. This opt-out also is intended to give low-performing providers a chance to improve.
- In preparation for publicly reporting the performance of individual physicians on measures of ambulatory care quality, the **California Physician Performance Initiative** (CPPI) has produced confidential performance reports. CPPI has requested that physicians review their performance scores, affirm their results, identify patient exclusions, and supply missing information. The CVE has received feedback that checking these data on a patient-by-patient basis can be quite onerous, especially for

cancer screening measures with large denominators (i.e., large numbers of patient events for review). The CVE is therefore trying to find a sampling strategy for this data accuracy check (i.e., a strategy in which physicians will only need to check a sample of their patients, and reported performance scores will be based on this checked sample).

3. **Assess the sensitivity of performance reports to earlier decisions.** In producing reports of provider performance, CVEs must make choices at the decision points discussed in this paper. In general, these decisions do not have "right answers," and CVEs may justifiably select from among many options. Sometimes, however, the tradeoffs involved at a decision point may not be entirely clear. The impact of each decision on the final published performance report may depend on the particular combination of decisions a CVE makes. CVE stakeholders may therefore be justified in asking, "What would have happened to the final report if we had made decision X differently?" The answer to this question is especially important in areas where a methodological decision was the result of contentious negotiation.

   One way to see how methodological decisions have affected performance reports is to conduct "sensitivity analysis." This process includes going back to a certain decision point (or combination of decision points), choosing another option, and recreating the performance report. **Sensitivity analyses can also be incorporated at each step in the report-generating process. Early identification of methodological decisions that have dramatic effects on performance scores (or categories) can be addressed by stakeholders before a full report is created.**

   CVEs may want to start with the following sensitivity analyses. These analyses include some of the decision points where (1) the "best" choice of methods is least certain, and (2) the impact of methodological choices on performance reports is likely to be greatest.

   - **Assess sensitivity to choice of attribution strategy** (discussed in the section on Task # 5). Many different attribution strategies can be considered, and research suggests that the choice of attribution strategy may affect which providers and measures can be included in a report.[26, 49] Choice of attribution strategy may also affect the reported performance of providers.
   - **Assess sensitivity to choice of strategy for creating composite measures, if composites are used.** There are many different strategies for creating composite measures (as discussed above), and research suggests that the choice of strategy can have a substantial impact on the reported performance of providers.[37] Both absolute performance (i.e., the composite performance score itself) and relative performance (i.e., how providers compare with each other in a report) can be affected.
   - **Assess sensitivity to choice of strategy for limiting the risk of misclassification due to chance.** Many different strategies for limiting misclassification risk can be used, alone and in combination (see the section on Task #5). Each of these strategies has strengths and weaknesses, and choice of strategy can affect the reported performance of providers.

- **Assess sensitivity to choice of strategy for handling outliers.** As mentioned in the section on Task #5, outliers are an especially important concern when reporting measures of the cost of care (but there may also be outliers on other performance measures). Because of multiple options for handling outliers, sensitivity analyses that try different approaches can provide valuable guidance (and possibly reassurance) to stakeholders when reporting performance on cost measures.
- **Assess sensitivity to case mix adjustment.** As discussed in the section on Task #5, whether to adjust (or stratify) performance data for patient characteristics can be a controversial decision. When such controversy is present, producing performance reports with and without case mix adjustment can help CVE stakeholders get a sense of whether case mix adjustment meaningfully changes the report.
- **Assess sensitivity to type of performance data.** As discussed in the section on Task #3, many different types of performance data can be used to generate the scores included in performance reports. Research suggests that the type of data might have a substantial impact on performance reports.[21] Obtaining some types of data (e.g., hybrid or medical record data) may require significant time and resources, but a CVE could consider performing sensitivity analysis on just a subset of the providers included in a report.

Advantages of sensitivity analysis:

- Sensitivity analysis provides a sense of how sensitive a report is to methodological decisions where the best answer is unclear. **If the performance report is essentially the same regardless of the methodological decisions (i.e., the same providers are categorized as higher and lower performers), then acceptance of the report may improve.**
- This analysis may improve buy-in from CVE stakeholders who dissented on a key methodological decision, because they get to see what would have happened with their way of doing things.

Limitations of sensitivity analysis:

- In a distributed data model (discussed in the section on Task #3), some sensitivity analyses will require the cooperation of each data source. For example, a CVE might obtain HEDIS measure numerators and denominators from a health plan. Performing sensitivity analyses on the attribution strategy will require the health plan to recalculate these numerators and denominators for each new attribution rule.
- If "prescored" performance data are used (discussed in the section on Task #3), it may not be possible to conduct many important sensitivity analyses.

Summary of Methodological Decisions Made by a Sample of CVE Stakeholders

During the spring of 2010, we interviewed the leaders of nine Chartered Value Exchanges (CVEs) and their stakeholder organizations to get a sense of how their collaboratives were approaching the decision points discussed in this paper. In this section, we present a synthesis of these responses. Because the interviews were qualitative in nature and the sample of CVEs was small, we refrain from presenting counts of each type of response. The aim of this section is to convey the range of methodological choices and, in illustrative cases, provide examples of the rationale behind some of these choices.

What are the purposes of publicly reporting provider performance?

In general, while public reporting efforts may have begun with the primary goal of helping patients choose providers, providers themselves turned out to be the primary audience for public performance reports. For example, access logs for Internet-based reports revealed that the overwhelming majority of individuals reviewing the reports were located in physician offices or hospitals.

The reports were felt to motivate providers to improve, and some reporting organizations also engaged in efforts to assist providers' improvement efforts. In addition, reports were often intended to help patients become partners in producing high-quality health care. For example, a report of diabetes performance could be used to educate patients about their care.

What will be the general format of performance reports?

CVEs and their stakeholder organizations adopted a variety of reporting formats, ranging from simplified reports that used symbols to indicate categories of provider performance to more complex numeric displays of performance data. In some cases, Internet-based reports offered both simple and complex formats, with numeric results available by selecting a provider's performance symbol.

Often the distinction between reports of relative performance and absolute performance was blurred. For example, a report might continue absolute performance percentages (e.g., the percentage of patients who received a necessary medical service) and arrange the providers in rank order based on these percentages. Thus, the report had a table of relative performance in which some providers were, by necessity, at the bottom and others at the top.

What will be the acceptable level of performance misclassification due to chance?

CVE leaders tended to view the acceptable risk of performance misclassification as being subject to ongoing conversation and negotiation, partly because misclassification risk is a relatively new concept in the field of health care provider performance reporting. However, there was consensus that providers, patients, and CVE leaders all wanted valid and reliable performance reports (i.e., reports that displayed performance data that were close to providers' "true" performance).

Because "misclassification risk" can be a foreign concept to many CVE stakeholders, it was suggested that to engage stakeholders in fruitful discussion, CVEs could improve participation by debating more concrete questions, such as "What constitutes a fair minimum sample size?" These concrete questions are important, fundamentally, because they influence the risk of performance misclassification.

Which measures will be included in a performance report?

In general, CVEs reported measures of the technical quality of care. Less commonly, CVEs reported measures of patients' health care experiences, and very few CVEs planned to report measures of cost or efficiency of care in the near future.

Leaders of CVEs or CVE stakeholder organizations with several years of performance reporting experience tended to describe formalized measure selection processes. These frequently involved committees of providers, purchasers, patient advocates, academics, and other interested parties. Measure selection processes were designed to identify measures that were aligned with local and national priorities, that would be plausibly valid and reliable in the provider population to be measured, and that had already been developed. In some cases, some of the older CVEs or CVE stakeholder organizations developed their own performance measures when no existing measures were available to address key local priorities for performance improvement.

Newer CVEs gravitated toward performance measures that were in common use across the country and that were already familiar to local stakeholders, often because individual health plans already had begun to give providers feedback on these measures. Measures from the Healthcare Effectiveness Data and Information Set (HEDIS) of the National Committee for Quality Assurance (NCQA)—or measures that are designed to capture similar aspects of provider performance—were commonly identified as initial priorities for public reporting.

How will performance measures be specified?

The leaders of CVEs and CVE stakeholder organizations reported a strong desire to use nationally endorsed measure specifications whenever these were available. Leaders found nationally endorsed specifications advantageous because they allowed comparison with national performance benchmarks and because national endorsement facilitated stakeholder buy-in. CVE leaders generally were not eager to try to "improve" nationally endorsed measure specifications. When improvements were deemed necessary, some leaders indicated that their preferred strategy would be to present their suggestions for improvement to national bodies, with the aim of changing the nationally endorsed measure.

However, it was not always possible to use nationally endorsed specifications when constructing performance measures. For example, nationally endorsed specifications might be designed for application to health plan claims data, but the performance data available to a CVE might come from provider registries. These types of "data source mismatch" often necessitated modifications to the nationally endorsed specifications so that similar performance measures could be constructed from locally available data. When modifications were made, CVE leaders emphasized the need to explain to stakeholders that comparisons with national benchmarks probably would not be valid.

What patient populations will be included?

CVE leaders generally sought to include as many patients as possible within their communities, subject to the limitations imposed by the sources of performance data. When health care providers supplied the performance data, such as data from registries or medical records, all patient populations receiving care from the providers could be included. However, when health plan claims were used, only the patients enrolled in the participating plans could be included. In some cases, this meant that only patients with commercial health insurance (and in some cases Medicaid) were included in performance reports. CVE stakeholders almost unanimously expressed a strong desire for Medicare fee-for-service claims data so that performance reports would reflect the care delivered to the Medicare population.

What kinds of data sources will be included?

CVEs and their stakeholder organizations reported using a variety of data sources as the basis for constructing performance measures. These sources included health plan claims, data from provider registries or medical records, patient survey data, and "prescored" data such as Leapfrog measures of hospital safety. When using health plan claims, CVEs usually contracted with an experienced claims analysis firm (when "raw" claims data were obtained directly from health plans) or used a distributed data model in which each health plan processed its own raw claims according to the measure specifications the CVE supplied.

How will data sources be combined?

CVEs and stakeholder organizations with more extensive public reporting experience heavily emphasized the importance of having a complete and accurate provider directory when combining sources of performance data (especially for ambulatory care, since even a relatively small locality can have many ambulatory providers). Such a directory was felt to be the best way to create a "crosswalk" between data sources, since each source might have its own identifier for the same provider. However, the leaders of these experienced CVEs noted that creating an accurate provider directory required the investment of substantial staff time and financial resources over multiple years. In addition, once the directory was accurate, maintaining its accuracy required significant ongoing investment.

Establishing good relationships with the provider community was mentioned as a key ingredient for successfully building such a directory. But even the CVEs that had accumulated greater reporting experience using directories of ambulatory providers had reporting limitations. For example, these directories tended not to include providers in practices below a certain size threshold (e.g., below two to four physicians in a single clinic).

How frequently will data be updated?

CVE leaders generally described updating the performance data in their public reports every 1 to 2 years. However, nearly all expressed a desire to both increase the frequency of data updates and decrease the lag between the time clinical care is delivered and the time of performance reporting based on that care. Some expressed the hope that electronic health records would enable "real-time" data collection that would enable these goals to be achieved.

How will tests for missing data be performed?

CVE leaders' approach to missing data depended on the data source. For data obtained directly from providers, some CVEs used auditing procedures that examined a sample of provider records to ensure completeness. For health plan claims data, some CVEs contracted with experienced claims analysis firms and verified that these firms performed tests for missing data. In addition, some CVE leaders emphasized the importance of knowing the data source. For example, if a certain health plan is known to have capitation products for which no fee-for-service claims are generated, claims data for patients enrolled in these capitation products will be "missing" from the standpoint of performance measure construction.

How will missing data be handled?

The approach of CVEs and stakeholder organizations to missing data was generally to first attempt to recover as much missing data as possible by working with data sources. After this step, CVEs reported provider performance based on the available data. Performance data were not imputed (i.e., statistically estimated), primarily because imputation was thought to be unacceptable to CVE stakeholders, especially providers. When a provider was known to be providing patient care in a CVE's community but no performance data for that provider could be reported, CVE reports generally displayed a symbol indicating that performance could not be reported due to a lack of sufficient performance data.

How will accuracy of data interpretation be assessed?

CVEs' general approach to ensuring accuracy of data interpretation was similar to the approach to identifying and handling missing data. However, in addition to working with experienced data analysts and knowing their data suppliers, some CVE leaders pointed out the importance of calculating community-level performance scores for a "reality check" (as an initial way to assess the accuracy of data interpretation).

How will performance data be attributed to providers?

Attribution rules varied from CVE to CVE and from performance measure to performance measure (even within the same CVE). For example, the attribution rules applied to screening measures might differ from the rules applied to measures of chronic disease care. Attribution to organizations (e.g., hospitals) followed national guidelines when these were available, but there was more heterogeneity in attribution strategies for ambulatory providers (including individual practitioners). In general, CVEs and CVE stakeholder organizations used plurality-based algorithms (e.g., majority of visits) or minimum-visit thresholds (e.g., at least one visit for a certain condition within the measurement year) to attribute ambulatory care measures to providers.

Will case mix adjustment be performed? (If so, how?)

When nationally endorsed measure specifications incorporate methods for case mix adjustment (e.g., measures of mortality rates for hospitals), CVEs and CVE stakeholder organizations generally applied these nationally endorsed case mix adjustment methods. However, when nationally endorsed case mix adjustment methods were not available (which was the case for most measures reported by CVEs), the leaders of CVEs and CVE stakeholder organizations reported that they did *not* apply new case mix adjustment methods in creating performance

reports. For example, no CVE leader that we interviewed performed case mix adjustment of process measures of the quality of care. When there was concern that certain patient populations might be more "challenging" than others, some CVEs reported stratified results instead of results that were case mix adjusted. The rationale for using stratification rather than case mix adjustment was that adjustment would "hide" undesirable disparities in care, while stratification would allow fair comparisons between providers without hiding disparities.

What strategies will be used to limit the risk of misclassification due to chance?

CVEs and CVE stakeholder organizations used a wide variety of strategies to limit the risk of performance misclassification due to chance. These included:

- *Basing performance thresholds on tests of statistical significance*. When this option was chosen, CVEs generally used statistical significance thresholds to err on the side of classifying provider performance as "average." This approach limited the probability of misclassifying a truly average provider as above or below average to no more than 5 percent. But in some cases, statistical confidence intervals were used to always give providers the benefit of the doubt. Each provider's performance was classified in the highest category that overlapped the provider's 95 percent confidence interval for the measure in question.
- *Using a "zone of uncertainty."* Because the risk of performance misclassification rises as provider performance gets close to a classification threshold, some CVEs gave providers the benefit of the doubt whenever their performance was within a "zone of uncertainty" around each threshold. In other words, performance was reported as being above threshold for all providers whose performance was within the zone of uncertainty.
- *Using a minimum reliability criterion*. Some CVEs limited the risk of misclassification due to chance by setting a minimum reliability for performance reporting (generally using a minimum reliability of 0.7 when this strategy was chosen). For this strategy, CVEs calculated reliability on a measure-by-measure and provider-by-provider basis, excluding from public reporting measures and providers that did not meet the minimum reliability standard.
- *Using a minimum number of observations (a "minimum N")*. Instead of using a minimum reliability criterion, some CVEs used a minimum N criterion. In deciding the right minimum number of observations, some CVEs looked for guidance from other reporting collaboratives and negotiated with their stakeholders. Other CVEs took a more mathematical approach, calculating for each performance measure the number of observations needed to achieve a minimum level of reliability (or limit the risk of misclassification to a certain level). CVEs taking the more mathematical approach found that (1) the minimum necessary number of observations could vary by measure, and (2) the minimum number of observations could be far greater than the minimum numbers used by other performance reporting collaboratives that had not taken a mathematical approach.
- *Reporting performance at higher levels of provider organization*. Most CVEs reported the performance of provider organizations (including ambulatory clinics) rather than individual practitioners. But some did express the goal of eventually finding ways to

report individual practitioners' performance in ways that would not introduce too much risk of performance misclassification due to chance.

Will composite measures be used?

Some CVEs and CVE stakeholder organizations reported provider performance on composite measures. When they were reported, composite measures were generally based on single health conditions (e.g., diabetes care, heart attack care) and used an approach based on taking the weighted average of the individual measures when calculating the composite measure score. However, some CVEs used an "all-or-none" approach to combining individual measure scores. CVE leaders observed two advantages of all-or-none composites:

- The range of between-provider variation on "all-or-none" composites was higher than the range on individual measures.
- The "all-or-none" approach was thought to be relatively easy to explain to stakeholders, including both patients and providers.

What final validity checks might improve the accuracy and acceptance of performance reports?

In general, CVEs and CVE stakeholder organizations described making final validity checks with the assistance of health care providers. These validity checks included giving providers a confidential preview of their performance results, which, in some cases, included patient-by-patient performance data. With these previews, providers could correct or appeal their performance results in some cases or, in other cases, opt out of a single round of public reporting to either correct problems with their data or improve their performance.

Appendix 1: Validity and Systematic Performance Misclassification

A. What is validity?

The validity of a provider performance report is the extent to which the performance information contained in the report means what it is supposed to mean (rather than meaning something else).[50-51] What a provider performance report is "supposed to mean" may depend on the purpose of reporting.[xi] But generally speaking, a report of provider performance is supposed to indicate something about providers (or something under providers' control), rather than something not inherently about providers or not under providers' control.

Put another way, a performance report will have *high* validity when its quality results truly represent the quality of care that a provider delivers. Similarly, a report with high validity will show efficiency results that truly represent provider efficiency, and so on. Reports would have *low* validity if they claimed to represent the quality of care delivered by a provider but instead truly represented the availability of parking in the provider's vicinity.

As a first step toward creating valid performance reports, CVEs should select performance measures that have "construct validity," which means that under ideal circumstances, the *measures should actually represent what they are supposed to represent*. For example, consider a hypothetical quality measure that counts the number of times drug X is given to patients with diabetes. For this measure to truly represent the quality of care, drug X should produce some kind of health benefit for patients with diabetes. If drug X helps diabetics' health, then the measure has construct validity. On the other hand, if drug X actually produces no health benefit (or even causes harm) for patients with diabetes, then the measure does not have construct validity.

Having performance measures with construct validity should be considered a bare minimum requirement for performance reporting. However, even when *measures* have construct validity, they can still be used to produce performance *reports* that have low validity. The following sections on systematic performance misclassification explain how even "valid measures" can lead to invalid reports of provider performance.

B. Systematic performance misclassification: a threat to validity

One way for a report of provider performance to have low validity is for the report to *systematically misclassify provider performance*. Systematic performance misclassification happens when the performance being reported is actually determined, to a significant degree, by something other than the performance that the report is supposed to present. To see how this can happen, consider the following scenario. Imagine that a CVE is reporting the performance of two hospitals on a measure of patient mortality and that the hospitals are identical in every way, except for one thing. One hospital serves a much older population than the other. Such a report of mortality is *supposed* to indicate which hospital is truly doing a better job at keeping its patients alive. However, because one hospital has an older patient population and older patients have higher average mortality than younger patients, the report will instead indicate which

[xi] Purposes of performance reporting are discussed in the section on Task #1.

hospital has a younger patient population (rather than which hospital is truly better). The report will therefore have low validity, even if there is no random measurement error.[xii]

In the real world, it is unlikely that two hospitals will be alike in every way, and measurement error will be present. It would be possible for a hospital serving older patients to still outperform a hospital serving younger patients on a mortality measure, either due to extraordinary efforts or due to chance. But on average, we would expect a group of hospitals serving younger patients to outperform those serving older patients on mortality measures. Therefore, on average, the mortality measure will still represent the age of the patient population rather than measuring how good each hospital is at keeping its patients alive. In other words, the hospitals will be systematically misclassified on the mortality measure.

C. Causes of systematic performance misclassification

In this section, we present three major causes of systematic performance misclassification that are addressed in this report. These causes are statistical bias, selection bias, and information bias.

1. **Statistical bias. When systematic performance misclassification is present because of differences in the patient populations served by different providers, the performance report contains "statistical bias" (also known as "omitted variable bias").** Two major techniques to address the problem of statistical bias in performance reports are case mix adjustment and stratification. Case mix adjustment uses statistical models to remove associations between patient characteristics and reported performance. For example, in a report that is case mix adjusted for patient age, there will be no association between patient age and reported provider performance (in other words, no providers will be "penalized" for having younger or older patients). Case mix adjustment is especially desirable when stakeholders feel that the patient characteristic in question is a *cause* of lower or higher measured performance.

   Stratification, which means reporting separate results for different groups of patients (e.g., younger and older patients), can accomplish the same goals as case mix adjustment in some cases. A more detailed overview of these techniques is presented in the section on Task #5. We emphasize case mix adjustment here to make three key points:

   - Statistical bias that causes systematic performance misclassification cannot be solved by adding more observations or specifying minimum sample sizes. The problem of systematic performance misclassification is methodologically distinct from the problem of performance misclassification due to chance (which is discussed in Appendix 2).
   - Statistical bias can only be detected when the factor that is causing the bias (e.g., different patient age distributions) can be identified and measured by the CVE. In the example of hospital mortality rates, there would be no way to know whether statistical bias and systematic performance misclassification are present without first knowing the ages of the patients who receive care from each hospital.

[xii] Measurement error is discussed in the section on Task #1.

- Even when statistical bias may be present, whether and how to account for it in performance reports is a value judgment. It depends on the nature of the performance measure in question, the story behind the statistical bias (i.e., the most likely reasons the bias is present), the purposes of public reporting, and the results of negotiations between CVE stakeholders.

2. **Selection bias. When the patients for whom performance data are available are not representative of the patients who will using a performance report, the validity of the report may be threatened by "selection bias."** For example, a performance report may be based only on the care provided to patients in commercial health plans. If providers' care for commercial enrollees systematically differs from the care for other patient populations (e.g., Medicare or Medicaid), then from the perspective of a noncommercial enrollee, the performance report may systematically misclassify provider performance. In other words, a patient enrolled in Medicaid may believe that a given provider has average performance when the provider actually has low (or high) performance for Medicaid enrollees. CVEs can address this threat to validity by gathering performance data from a wide variety of patients (discussed in the section on Task #2) and by creating "stratified" performance reports that show different performance scores for different patient populations (discussed in the section on Task #5).

3. **Information bias. When certain providers underreport performance data (a particular concern when these data would indicate low performance), the validity of a performance report is threatened by "information bias."** If providers with low performance tend to have more missing data than other providers, the report may systematically misclassify low-performing providers as having "observed" performance that is higher than their "true" performance. This threat to validity, and potential ways of addressing it, is discussed in the section on Task #4.

Appendix 2: Performance Misclassification Due to Chance

This section contains background information on the concept of performance misclassification due to chance. We recommend this section to Chartered Value Exchange (CVE) stakeholders who are interested in learning why this concept is important and understanding how the following more commonly discussed topics relate to each other:

- Reliability.
- Measurement error.
- Sample sizes.

For practical guidance on options for limiting the risk of performance misclassification, see the section on Task #5. For a more detailed discussion of reliability and performance misclassification due to chance, we refer interested readers to two technical reports: *The Reliability of Provider Profiling: A Tutorial* by Adams,[51] and *Estimating Reliability and Misclassification in Physician Profiling* by Adams, Mehrotra, and McGlynn.[52]

A. What is misclassification due to chance?

Any time performance is measured, there will always be some amount of **random measurement error**. The unavoidable presence of measurement error means that for every provider in a report, there is a certain probability that due to chance alone, performance is reported in the wrong class or category. In other words, any performance report that contains more than one category (i.e., a report that enables any kind of comparison between providers) will have some degree of misclassification due to chance.

However, patients, providers, and other CVE stakeholders may want to limit the amount of misclassification due to chance. Reports with too much misclassification due to chance may mislead large numbers of patients, and providers may also be concerned about the impact of misclassification, as shown in Figure 3.

Unfortunately, it is impossible to know exactly which providers are misclassified due to chance alone. On the other hand, it generally *is* possible to know, for each provider, the *risk* (i.e., the probability) that performance is misclassified.

B. Why focus on the risk of misclassification due to chance?

Some CVE stakeholders may be familiar with the statistical concept of "reliability," which is related to misclassification. Reliability, which is more formally defined later in this appendix, can be conceptualized as the "signal-to-noise ratio" in measuring performance. With larger amounts of measurement "noise" (i.e., greater quantities of random measurement error), it becomes hard to discern the "signal" in performance data (i.e., which providers are truly higher performing and which are truly lower performing).

The reason that this paper focuses on the "risk of misclassification due to chance" rather than solely focusing on reliability is that *on its own, reliability does not have a direct, easily understood interpretation in performance reporting.* For performance reports, the significance of reliability depends on the system for classifying performance.

The relationship between reliability and performance misclassification due to chance was originally highlighted in the 2006 work of Safran and colleagues.[31] As Dr. Safran recalls:

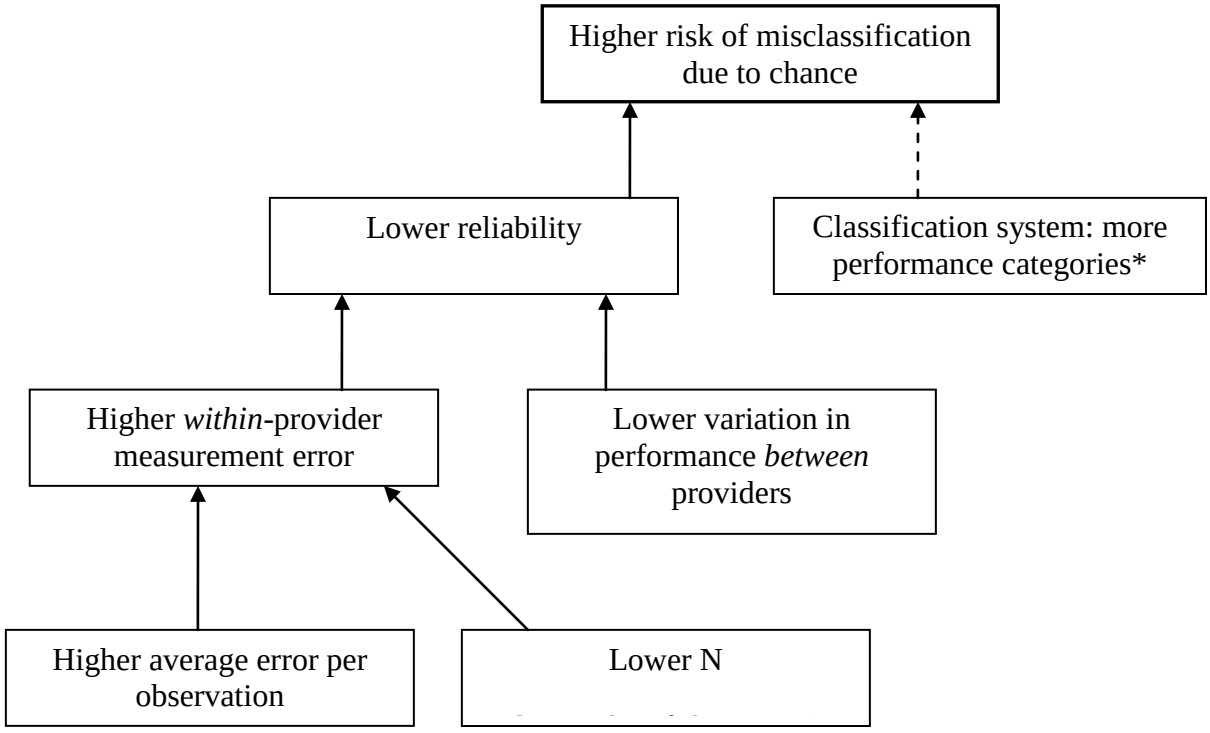
> The idea of evaluating risk of misclassification came out of a wonderful question that I was asked by a clinician, who was troubled by our reliability criterion of 70 percent. The clinician asked: "Does that mean there is a 30 percent chance that you have my score wrong? Because if that's what it means, then maybe statisticians think 70 percent is a good standard, but clinicians will find it unacceptable." This question inspired our work to elucidate the "risk of misclassification" construct and to have a methodology that allowed us to operationalize it.

In a performance report that is constructed with misclassification in mind, reliability of 70 percent might translate into a risk of misclassification that is quite low (less than 2.5 percent in the classification system for reporting patient experience results that was presented in Safran's 2006 paper).[31] Recent work by Adams and colleagues also provides an example of the relationship between reliability and performance misclassification in reports of physicians' performance on cost measures.[29, 51-52]

C. What determines the risk of misclassification due to chance?

The risk, or probability, of misclassification due to chance is directly determined by the statistical "reliability" of the provider's measured performance and the classification system that is used in the performance report. The reliability of a measure is affected by the number of observations, the average level of "error" per observation, and the amount of provider-to-provider variation in performance.[51] The relationship between these factors and the risk of misclassification is illustrated in Figure 5.

Figure 5. Factors that determine the risk of misclassification due to chance



* Note: having more performance categories generally raises the risk of misclassification, but this is not *always* true. In addition to depending on the *number* of categories, misclassification risk depends on *where* the performance thresholds between categories are drawn.

To help explain Figure 5, we define the terms below.

Classification system refers to the way provider performance is presented in reports. The kind of classification system used is a value judgment, and there is no single best classification system for all purposes and users. Examples of classification systems include categorizing providers as "below average," "average," and "above average"; giving providers star ratings based on designated performance thresholds; and ranking providers according to their relative performance.

Deciding on the method of classifying provider performance will be influenced by how the results will be used. The classification system is the result of decisions about (1) whether to use performance thresholds, (2) how many thresholds to use, (3) where to set thresholds, and (4) what kind of performance scores to report ("shrunken" or "observed" performance; discussed in section on Task #5). Because the classification system used in a performance report is one of the key determinants of misclassification risk, ***it is impossible to calculate the misclassification risk for providers included in a report without first deciding upon a classification system.***

Reliability is a property of the performance measure, the individual provider, *and* the provider population being measured. Therefore, if a CVE truly wants to know the magnitude of the risk of misclassification in its reports, the CVE will need to compute reliabilities for the measures it applies within its own provider community.

Reliability is a statistical concept that describes how well one can confidently distinguish the performance of one provider from another.[xiii] Put another way, reliability is determined by the relative amounts of "signal" and "noise" in performance data. In Figure 5, *within*-provider measurement error (i.e., random measurement error) is the "noise" and *between*-provider variation in performance is the "signal" a CVE may want to detect. Reliability is very important to determining misclassification risk: For any given classification system, the higher the reliability, the lower the misclassification risk.

For some types of performance measures, each individual provider in a report may have a different level of reliability. In general, when providers can have different numbers of observations (e.g., measures of diabetes quality) or different amounts of error per observation (e.g., cost profiles), reliability can only be calculated on a provider-by-provider basis. An example of how reliability can vary by provider is presented in a recent paper by Adams and colleagues that investigates the reliability of physician cost profiles.[29] The technical appendix accompanying Adams' paper contains a more detailed statistical explanation of reliability and how it can vary from provider to provider, even on the same performance measure.[53]

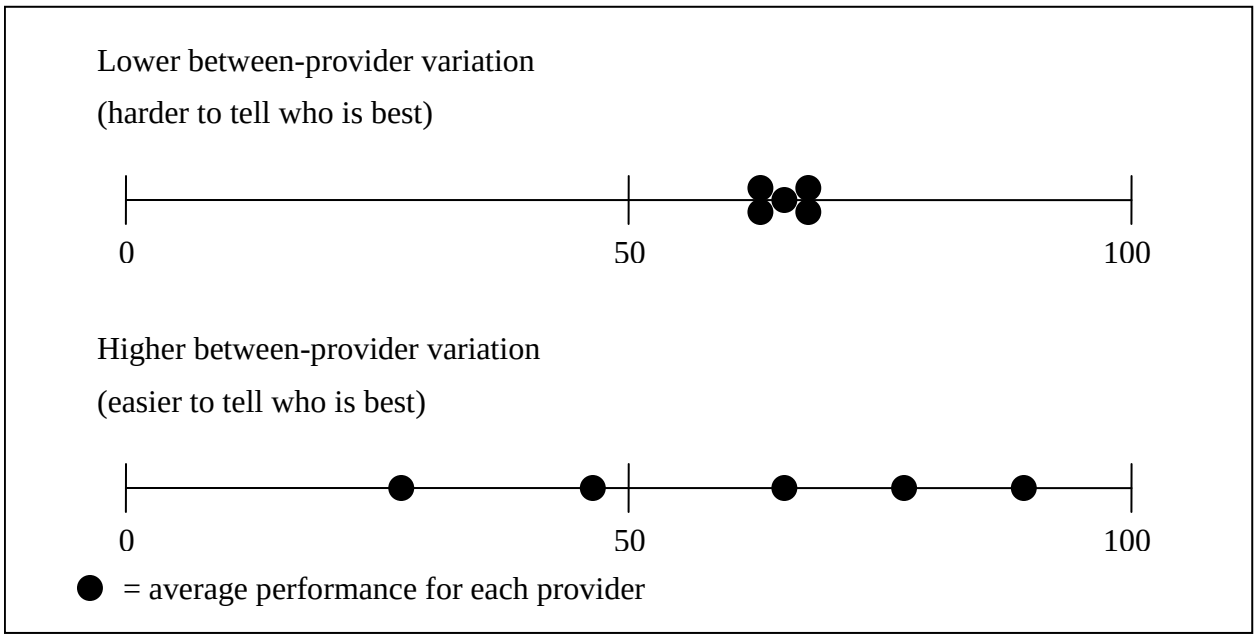
On the other hand, for measures such as patient experience ratings, reliability may be the same for all providers in a report. This can occur because the amount of error per observation is generally a property of the survey instrument (rather than the provider), and the number of observations (i.e., the number of survey responses) can be equalized across providers.

[xiii] Reliability also describes how close the *measured* performance of a provider is to the *true* performance of that provider. Mathematically, these two definitions are identical.

Variation in performance *between* providers. Performance variation matters for many reasons, including the usefulness of reports to patients choosing a provider. If all providers have the exact same performance, or zero variation, patients cannot use the report to choose among them.[xiv] For the sake of misclassification risk, performance variation matters because it affects reliability. All other things being equal, the higher the performance variation between providers, the higher the reliability (or ability to discriminate performance between providers) and the lower the risk of misclassifying providers due to chance.

- *Illustration:* Imagine that you are trying to report the performance of providers on a measure that goes from 0 (bad performance) to 100 (good performance). Figure 6 gives an example of how performance variation might look for two populations, each containing five providers.

Figure 6. Performance variation in two populations of providers



***Within*-provider measurement error** is a statistical term that describes the amount of uncertainty in the performance that is measured for a single provider, taking account of all the available observations for that provider. Although the word "error" is used, it does not mean a mistake is being made in performance measurement and reporting. Instead, measurement error is a natural phenomenon that occurs in all measurement processes, from taking a patient's weight and blood pressure to evaluating a provider's performance. There is a hypothetical (and unobservable) "true" value for all the things we might try to measure. Measurement allows us to determine the range in which this "true" value probably exists.

Provider performance is no different. The lower the within-provider measurement error, the more precise the estimate of "true" performance becomes. Statistical confidence intervals are one

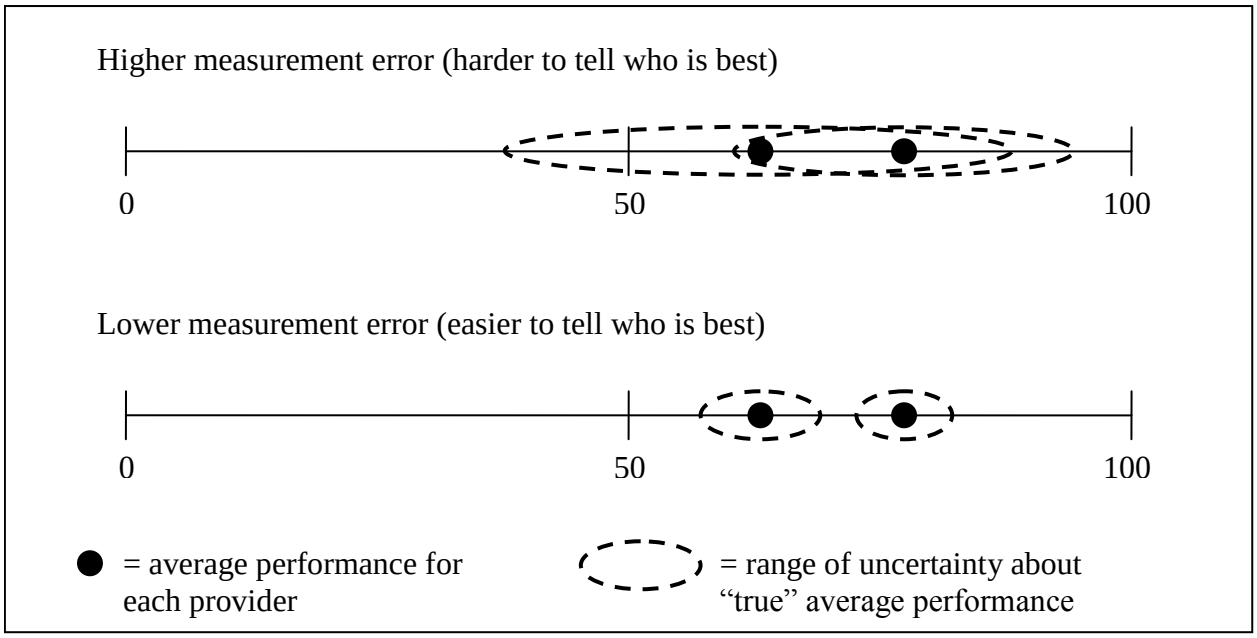
[xiv] This uniformity of performance would not necessarily be a bad thing. If providers had uniformly high performance, then patients could choose providers based on factors such as out-of-pocket cost and convenience, resting assured that no matter what provider they chose, performance would be above an acceptable threshold.

example of a technique to calculate the range in which "true" performance probably exists (for a 95% confidence interval, there is a 95% chance that "true" performance is within the interval).

Within-provider measurement error matters to misclassification risk because, all other things being equal, the higher the measurement error, the lower the reliability and the higher the risk of misclassification due to chance.

- *Illustration:* Figure 7 shows an example of two providers, showing both the observed average performance on a single measure and the amount of uncertainty around about "true" average performance. Even though the observed average performance levels are identical in both examples (so the variation in performance between providers is the same), the within-provider measurement error is different.

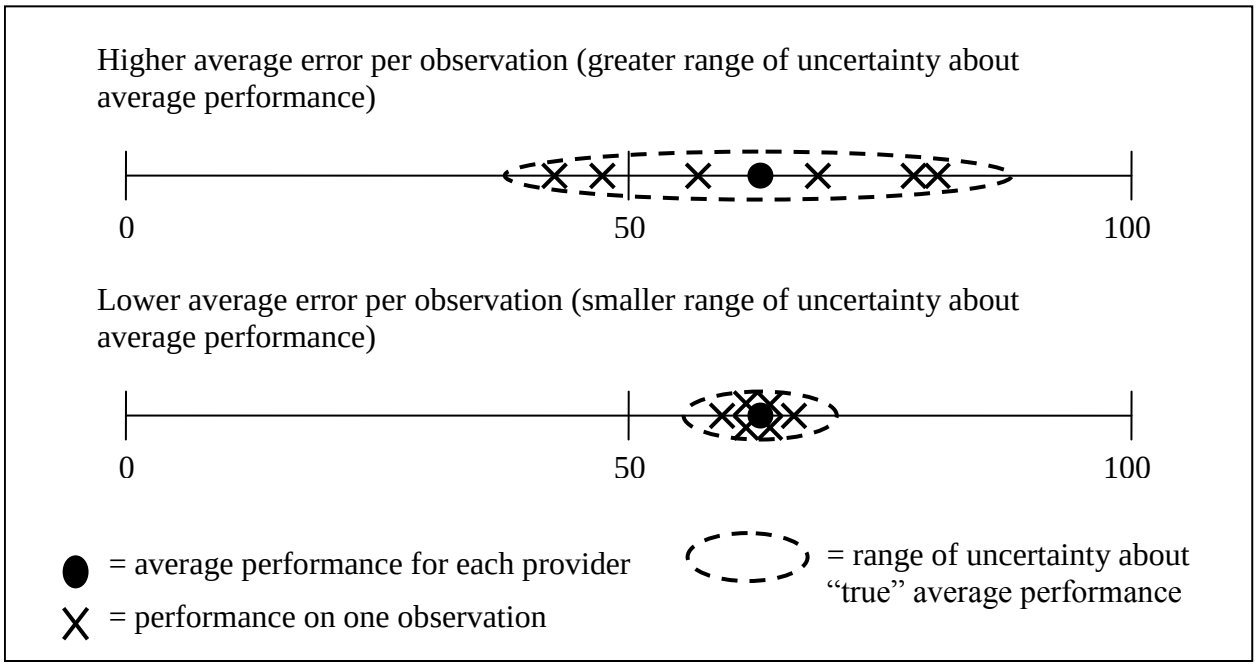
Figure 7. Different levels of measurement error (uncertainty about "true" average performance)



Average error per observation is a statistical term that describes how much variation there is in the observation-to-observation performance of a single provider. Due to chance alone, a given provider's performance on a measure may vary from patient to patient, from day to day, from week to week, etc. The more this performance varies, the harder it is to distinguish one provider from other providers. Average error per observation matters to misclassification risk because, all other things being equal, the higher the average error, the higher the within-provider measurement error, the lower the reliability, and the higher the risk of misclassification due to chance.

- *Illustration:* Figure 8 shows an example of observations for two providers on a single performance measure (e.g., a measure of costs). Each provider has six observations. The average score is the same for both providers, but one has higher average error per observation than the other.

Figure 8. How average error per observation affects uncertainty about the "true" average performance

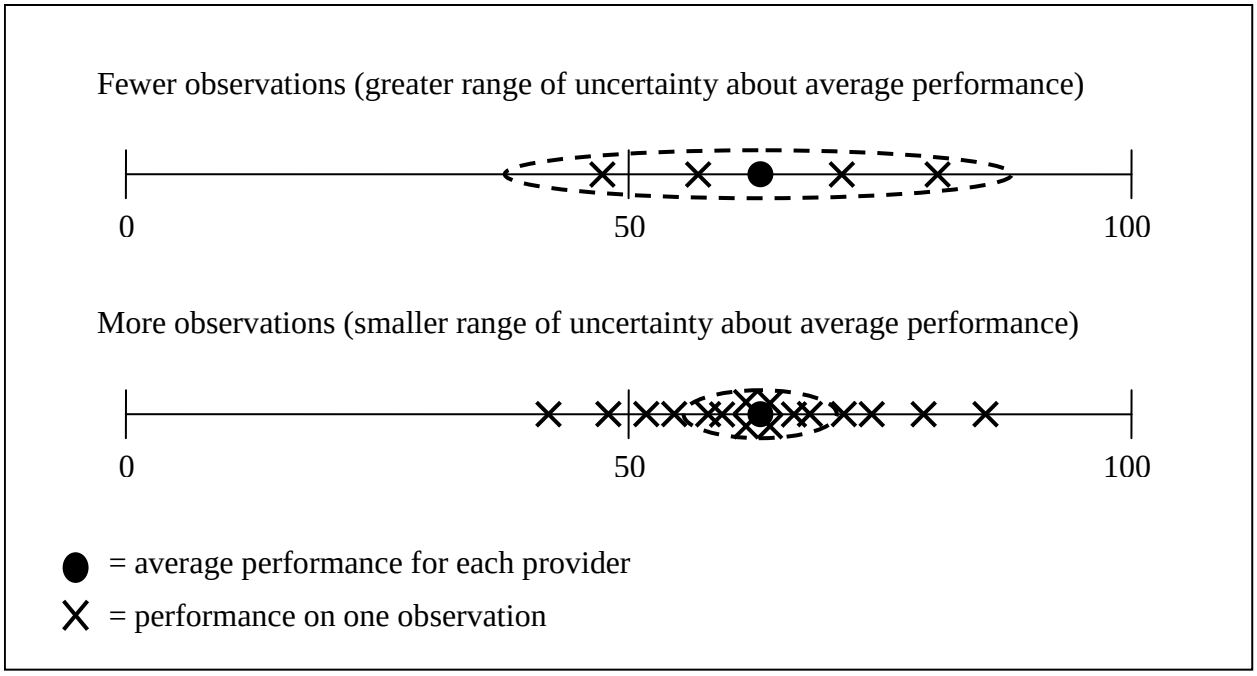


N (number of observations) refers to the number of observations a given provider has on a performance measure. For example, if the performance measure assesses hemoglobin A1c control in patients with diabetes, the number of observations for a provider will be the number of that provider's patients who have diabetes and who qualify for inclusion in the measure. The number of observations matters to performance misclassification because, all other things being equal, the higher the number of observations, the lower the within-provider measurement error, the higher the reliability, and the lower the risk of performance misclassification.

Misclassification risk is affected by factors other than the number of observations (i.e., the average error per observation and the classification system). ***Therefore, it is impossible to specify a minimum number of observations that will limit the risk of misclassification across all providers, all classification systems, or all measures.*** In fact, because the average error per observation can vary from provider to provider, different providers may need different numbers of observations to reach the same risk of misclassification.

- *Illustration:* Figure 9 shows an example of how differing numbers of observations affect the amount of uncertainty around the average performance for two providers. The average error per observation is the same for both providers, and both have the same observed average performance. But because one provider has more observations than the other, the range of uncertainty about the "true" average performance is smaller.

Figure 9. How the number of observations affects uncertainty about the "true" average performance



References

1. Hibbard J, Sofaer S. Best practices in public reporting no. 1: how to effectively present health care performance data to consumers. Rockville, MD: Agency for Healthcare Research and Quality; June 2010. Available at: www.ahrq.gov/qual/pubrptguide1.htm. Accessed September 27, 2010.
2. Sofaer S, Hibbard J. Best practices in public reporting no. 2: maximizing consumer understanding of public comparative quality reports: effective use of explanatory information. Rockville, MD: Agency for Healthcare Research and Quality; June 2010. Available at: www.ahrq.gov/qual/pubrptguide2.htm. Accessed September 27, 2010.
3. Berwick DM, James B, Coye MJ. Connections between quality measurement and improvement. Med Care 2003 Jan;41(1 Suppl):I30-38.
4. Hibbard JH. Engaging health care consumers to improve the quality of care. Med Care 2003 Jan;41(1 Suppl):I61-70.
5. Hibbard JH, Stockard J, Tusler M. Does publicizing hospital performance stimulate quality improvement efforts? Health Aff (Millwood) 2003 Mar-Apr;22(2):84-94.
6. Marshall MN, Shekelle PG, Leatherman S, et al. The public release of performance data: what do we expect to gain? A review of the evidence. JAMA 2000 Apr 12;283(14):1866-74.
7. Fung CH, Lim YW, Mattke S, et al. Systematic review: the evidence that publishing patient care performance data improves quality of care. Ann Intern Med 2008 Jan 15;148(2):111-23.
8. Faber M, Bosch M, Wollersheim H, et al. Public reporting in health care: how do consumers use quality-of-care information? A systematic review. Med Care 2009 Jan;47(1):1-8.
9. Hibbard JH. What can we say about the impact of public reporting? Inconsistent execution yields variable results. Ann Intern Med 2008 Jan 15;148(2):160-1.
10. Romano PS, Rainwater JA, Antonius D. Grading the graders: how hospitals in California and New York perceive and interpret their report cards. Med Care 1999 Mar;37(3):295-305.
11. Tu JV, Donovan LR, Lee DS, et al. Effectiveness of public report cards for improving the quality of cardiac care: the EFFECT study: a randomized trial. JAMA 2009 Dec 2;302(21):2330-7.
12. Davis MM, Hibbard JH, Milstein A. Issue Brief: Consumer tolerance for inaccuracy in physician performance ratings: one size fits none. Washington, DC: Center for Studying Health System Change; March 2007.
13. Krumholz HM, Normand SLT. Public reporting of 30-day mortality for patients hospitalized with acute myocardial infarction and heart failure. Circulation 2008 Sep;118(13):1394-7.
14. Romano PS, Hussey P, Ritley D. Selecting quality and resource use measures: a decision guide for community quality collaboratives. Rockville, MD: Agency for Healthcare Research and Quality; November 2009. AHRQ Pub. No. 09(10)-0073. Available at: www.ahrq.gov/qual/perfmeasguide/. Accessed September 27, 2010.
15. National Quality Forum. Measure Evaluation Criteria. Washington, DC: National Quality Forum; August 2008. Available at: www.qualityforum.org/docs/measure_evaluation_criteria.aspx. Accessed September 27, 2010.
16. National Priorities Partnership. National priorities and goals: aligning our efforts to transform America's healthcare. Washington, DC: National Quality Forum; 2008. Available at: www.nationalprioritiespartnership.org/uploadedFiles/NPP/08-253-NQF%20ReportLo%5B6%5D.pdf. Accessed September 27, 2010.
17. McGlynn EA. Identifying, categorizing, and evaluating health care efficiency measures (prepared by the Southern California Evidence-based Practice Center—RAND Corporation, under Contract No. 282-00-0005-21). Rockville, MD: Agency for Healthcare Research and Quality; April 2008. AHRQ Pub. No. 08-0030. Available at: www.ahrq.gov/qual/efficiency/efficiency.pdf. Accessed September 27, 2010.
18. AQA parameters for selecting measures for physician performance. AQA Alliance; June 2009. Available at: www.aqaalliance.org/files/AQAParametersforSelectingAmbulatoryCare.pdf. Accessed September 27, 2010.
19. Remus D, Fraser I. Guidance for using the AHRQ quality indicators for hospital-level public reporting or payment. Rockville, MD: Department of Health and Human Services, Agency for Healthcare Research and Quality; 2004. AHRQ Pub. No. 04-0086-EF. Available at: www.qualityindicators.ahrq.gov/archives/documents/qi_guidance.pdf. Accessed September 27, 2010.
20. Guidance for using the AHRQ quality indicators for public reporting or payment, Appendix B: Public reporting evaluation framework—comparison of recommended evaluation criteria in five existing national frameworks. Rockville, MD: Agency for Healthcare Research and Quality; 2004. Available at: qualityindicators.ahrq.gov/downloads/technical/qi_guidance_appendix_B.pdf. Accessed September 27, 2010.
21. Pawlson LG, Scholle SH, Powers A. Comparison of administrative-only versus administrative plus chart review data for reporting HEDIS hybrid measures. Am J Manag Care 2007 Oct;13(10):553-8.

22. McGlynn EA, Adams J, Hicks J, et al. Creating a coordinated autos/UAW reporting system (CARS) for evaluating health plan performance. Santa Monica, CA: RAND; September 1999.
23. U.S. Department of Health and Human Services. Hospital Compare. Available at: www.hospitalcompare.hhs.gov. Accessed September 27, 2010.
24. Pham HH, Schrag D, O'Malley AS, et al. Care patterns in Medicare and their implications for pay for performance. N Engl J Med 2007 Mar 15;356(11):1130-9.
25. Landon BE, O'Malley AJ, Keegan T. Can choice of the sample population affect perceived performance: implications for performance assessment. J Gen Intern Med 2010 Feb;25(2):104-9.
26. Mehrotra A, Adams JL, Thomas JW, et al. The effect of different attribution rules on individual physician cost profiles. Ann Intern Med 2010 May 18;152(10):649-54.
27. Atlas SJ, Grant RW, Ferris TG, et al. Patient-physician connectedness and quality of primary care. Ann Intern Med 2009 Mar 3;150(5):325-35.
28. Thomas JW, Ward K. Economic profiling of physician specialists: use of outlier treatment and episode attribution rules. Inquiry 2006 Fall;43(3):271-82.
29. Adams JL, Mehrotra A, Thomas JW, et al. Physician cost profiling -- reliability and risk of misclassification. N Engl J Med 2010 March 18;362(11):1014-21.
30. Adams JL, McGlynn EA, Thomas JW, et al. Incorporating statistical uncertainty in the use of physician cost profiles. BMC Health Serv Res 2010 March 5;10:57.
31. Safran DG, Karp M, Coltin K, et al. Measuring patients' experiences with individual primary care physicians. Results of a statewide demonstration project. J Gen Intern Med 2006 Jan;21(1):13-21.
32. Zaslavsky AM. Statistical issues in reporting quality data: small samples and casemix variation. Int J Qual Health Care 2001 Dec;13(6):481-8.
33. Sequist TD, Schneider EC, Li A, et al. Reliability of medical group and physician performance measurement in the primary care setting. Med Care 2011 Feb;49(2):126-31.
34. Scholle SH, Roski J, Adams JL, et al. Benchmarking physician performance: reliability of individual and composite measures. Am J Manag Care 2008 Dec;14(12):833-8.
35. Elliott M, Zaslavsky A, Cleary P. Are finite population corrections appropriate when profiling institutions? Health Serv Outcomes Res Methodol 2006;6(3):153-6.
36. Dimick JB, Staiger DO, Baser O, et al. Composite measures for predicting surgical mortality in the hospital. Health Aff (Millwood) 2009 Jul-Aug;28(4):1189-98.
37. Reeves D, Campbell SM, Adams J, et al. Combining multiple indicators of clinical quality: an evaluation of different analytic approaches. Med Care 2007 Jun ;45(6):489-96.
38. Shwartz M, Ash AS. Composite measures: matching the method to the purpose. November 2008. Available at: www.qualitymeasures.ahrq.gov/expert/expert-commentary.aspx?id=16464. Accessed September 27, 2010.
39. Reporting Measures for the CAHPS Clinician & Group Survey. CAHPS Clinician & Group Survey and Reporting Kit. Rockville, MD: Agency for Healthcare Research and Quality; 2007. Available at: www.cahps.ahrq.gov/CAHPSkit/files/309_CG_Reporting_Measures.htm. Accessed September 27, 2010.
40. Feinstein AR. Multi-item “instruments” vs Virginia Apgar’s principles of clinimetrics. Arch Intern Med 1999 Jan 25;159(2):125-8.
41. Gandhi TK, Francis EC, Puopolo AL, et al. Inconsistent report cards: assessing the comparability of various measures of the quality of ambulatory care. Med Care 2002 Feb;40(2):155-65.
42. Jha AK, Li Z, Orav EJ, et al. Care in U.S. hospitals—the Hospital Quality Alliance program. N Engl J Med 2005 Jul 21;353(3):265-74.
43. Quality Indicators (AHRQ QI). Inpatient Quality Indicators Composite Measure Workgroup final report. Rockville, MD: Agency for Healthcare Research and Quality; March 2008. Available at: www.qualityindicators.ahrq.gov/downloads/iqi/AHRQ_IQI_Workgroup_Final.pdf. Accesssed September 27, 2010.
44. Lied TR, Malsbary R, Eisenberg C, et al. Combining HEDIS indicators: a new approach to measuring plan performance. Health Care Financ Rev 2002 Summer;23(4):117-29.
45. Zaslavsky AM, Shaul JA, Zaborski LB, et al. Combining health plan performance indicators into simpler composite measures. Health Care Financ Rev 2002 Summer;23(4):101-15.
46. Timbie JW, Shahian DM, Newhouse JP, et al. Composite measures for hospital quality using quality-adjusted life years. Stat Med 2009 Apr 15;28(8):1238-54.
47. Nolan T, Berwick DM. All-or-none measurement raises the bar on performance. JAMA 2006 Mar 8;295(10):1168-70.

48. Van Matre JG. All-or-none measurement of health care quality. JAMA 2006 Jul 26;296(4):392; author reply 393.
49. Scholle SH, Roski J, Dunn DL, et al. Availability of data for measuring physician quality performance. Am J Manag Care 2009 Jan;15(1):67-72.
50. Hays RD, Fayers P. Reliability and validity (including responsiveness). In: Fayers P, Hays RD, eds. Assessing quality of life in clinical trials: methods and practice, 2nd ed. New York: Oxford University Press; 2005. pp. 25-39.
51. Adams JL. The reliability of provider profiling: a tutorial. Santa Monica, CA: RAND Corporation; 2009. Available at: www.rand.org/pubs/technical_reports/TR653/. Accessed September 27, 2010.
52. Adams JL, Mehrotra A, McGlynn EA. Estimating reliability and misclassification in physician profiling. Santa Monica, CA: RAND Corporation; 2010.
53. Adams JL, Mehrotra A, Thomas JW, et al. Physician cost profiling: reliability and risk of misclassification: detailed methodology and sensitivity analyses. Santa Monica, CA: RAND; 2010. Available at: www.rand.org/pubs/technical_reports/TR799/. Accessed September 27, 2010.

Index