RAND HEALTH

THE ARTS
CHILD POLICY
CIVIL JUSTICE
EDUCATION
ENERGY AND ENVIRONMENT
HEALTH AND HEALTH CARE
INTERNATIONAL AFFAIRS
NATIONAL SECURITY
POPULATION AND AGING
PUBLIC SAFETY
SCIENCE AND TECHNOLOGY
SUBSTANCE ABUSE
TERRORISM AND HOMELAND SECURITY
TRANSPORTATION AND INFRASTRUCTURE
WORKFORCE AND WORKPLACE

This PDF document was made available from www.rand.org as a public service of the RAND Corporation.

Jump down to document

The RAND Corporation is a nonprofit institution that helps improve policy and decisionmaking through research and analysis.

Support RAND

Browse Books & Publications

Make a charitable contribution

For More Information

Visit RAND at www.rand.org

Explore RAND Health

View document details

Limited Electronic Distribution Rights

This document and trademark(s) contained herein are protected by law as indicated in a notice appearing later in this work. This electronic representation of RAND intellectual property is provided for non-commercial use only. Unauthorized posting of RAND PDFs to a non-RAND Web site is prohibited. RAND PDFs are protected under copyright law. Permission is required from RAND to reproduce, or reuse in another form, any of our research documents for commercial use. For information on reprint and linking permissions, please see RAND Permissions.

This product is part of the RAND Corporation technical report series. Reports may include research findings on a specific topic that is limited in scope; present discussions of the methodology employed in research; provide literature reviews, survey instruments, modeling exercises, guidelines for practitioners and research professionals, and supporting documentation; or deliver preliminary findings. All RAND reports undergo rigorous peer review to ensure that they meet high standards for research quality and objectivity.

TECHNICAL
REPORT

Estimating Reliability and Misclassification in Physician Profiling

John L. Adams, Ateev Mehrotra, Elizabeth A. McGlynn

Sponsored by the American Medical Association and the Massachusetts Medical Society

RAND HEALTH

This work was sponsored by the American Medical Association and the Massachusetts Medical Society. The research was conducted in RAND Health, a division of the RAND Corporation.

The RAND Corporation is a nonprofit institution that helps improve policy and decisionmaking through research and analysis. RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

RAND® is a registered trademark.

© Copyright 2010 RAND Corporation

Permission is given to duplicate this document for personal use only, as long as it is unaltered and complete. Copies may not be duplicated for commercial purposes. Unauthorized posting of RAND documents to a non-RAND website is prohibited. RAND documents are protected under copyright law. For information on reprint and linking permissions, please visit the RAND permissions page (http://www.rand.org/publications/permissions.html).

Published 2010 by the RAND Corporation
1776 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138
1200 South Hayes Street, Arlington, VA 22202-5050
4570 Fifth Avenue, Suite 600, Pittsburgh, PA 15213-2665
RAND URL: http://www.rand.org
To order RAND documents or to obtain additional information, contact
Distribution Services: Telephone: (310) 451-7002;
Fax: (310) 451-6915; Email: order@rand.org

Preface

This technical report explains the relationship between reliability measurement and misclassification for physician quality and cost measures in health care. It provides details and a practical method to calculate reliability and misclassification from the data typically available to health plans. This report will be of interest to national and state policymakers, health care organizations and clinical practitioners, patient and provider advocacy organizations, health researchers, and others with responsibilities for ensuring that patients receive cost-effective, high-quality health care.

This work was sponsored by the American Medical Association (AMA) and the Massachusetts Medical Society (MMS). The research was conducted in RAND Health, a division of the RAND Corporation. A profile of RAND Health, abstracts of its publications, and ordering information can be found at www.rand.org/health.

Summary

Public and private purchasers and health plans are demanding more information about the quality and relative cost of U.S. physicians to increase physician accountability and to aid in value-based purchasing.[1,2] Although performance measurement has been in place for some time in hospitals and managed care organizations (MCOs), the focus on physicians is a relatively new development. The inherent limitations of the data available at the physician level have brought to the fore technical issues that were less important at higher levels of aggregation in hospitals and MCOs.[3,4]

One of these technical issues is the statistical reliability of a physician's performance measure and how it may lead to misclassification of physicians as high or low performers. While the use of more reliable measures is obviously important to all stakeholders, the meanings of reliability and misclassification in this context are sometimes unclear, and the methods for measuring both reliability and misclassification in practice may seem daunting to those designing and implementing performance measurement systems. Addressing these needs is the focus of this report. This report builds on other RAND reports and publications on reliability and misclassification and has two major goals.[5,6,7] First, it can serve as a tutorial for measuring reliability and misclassification. Second, it goes beyond our previous work to describe the potential for misclassification when physicians are categorized using statistical testing.

Fundamentally, reliability is a quantitative measure indicating the confidence one can have that a physician is different from his or her peers. One concern is that for most readers, interpreting

[1] McKethan A, Gitterman D, Feezor A, and Enthoven A, "New Directions for Public Health Care Purchasers? Responses to Looming Challenges," *Health Affairs*, Vol. 25, No. 6, 2006, pp. 1518–1528.
[2] Milstein A and Lee TH, "Comparing Physicians on Efficiency," *New England Journal of Medicine*, Vol. 357, No. 26, 2007, pp. 2649–2652.
[3] Associated Press, "Regence BlueShield, WA Doctors Group Settle Lawsuit," *Seattle Post-Intelligencer*, August 8, 2007.
[4] Kazel R, "Tiered Physician Network Pits Organized Medicine vs. United," *American Medical News*. March 7, 2005.
[5] Adams JL, Mehrotra A, Thomas JW, and McGlynn EA, *Physician Cost Profiling—Reliability and Risk of Misclassification: Detailed Methodology and Sensitivity Analyses*, Santa Monica, Calif.: RAND Corporation, TR-799-DOL, 2010a.
[6] Adams JL, *The Reliability of Provider Profiling: A Tutorial*, Santa Monica, Calif.: RAND Corporation, TR-653-NCQA, 2009.
[7] Adams JL, Mehrotra A, Thomas JW, and McGlynn EA, "Physician Cost Profiling—Reliability and Risk of Misclassification," *The New England Journal of Medicine*, Vol. 362, No. 11, March 18, 2010b, pp. 1014–1021.

reliability is not intuitive. Additionally, there is no agreement on a level of reliability that is considered acceptable in the context of provider profiling. We describe how we used reliability to estimate a more intuitive concept: the rate at which physicians are misclassified for a particular application of cost profiling.

The most commonly used applications of cost profiles (e.g., public reporting, pay for performance, tiering) typically require classifying physicians into categories. Reliability can be used to calculate the probability that a physician will be correctly or incorrectly classified in a particular application.[8] The reliability-misclassification relationship can be estimated for most common reporting systems that include cut points based on percentiles and statistical testing. In this report, we explore a number of categorization systems and the misclassification rates associated with each system.

Reliability is a topic of increasing importance in profiling applications. We hope this tutorial will reduce some of the confusion about what reliability is and how it relates to misclassification. Ultimately, whether reliability is "good enough" for any given application will be judged by the probability that the system misclassifies physicians. For any newly proposed system, the methods presented here should enable an evaluator to calculate the reliabilities and, consequently, the misclassification probabilities. It is our hope that knowing these misclassification probabilities will increase transparency about profiling methods and stimulate an informed debate about the costs and benefits of alternative profiling systems.

Introduction

Public and private purchasers and health plans are demanding more information about the quality and relative cost of U.S. physicians to increase physician accountability and aid in value-based purchasing.[9,10] Although performance measurement has been in place for some time in hospitals and MCOs, the focus on physician profiling is a relatively new development. The inherent limitations of the clinical and other data available at the physician level have brought to the fore

[8] If assignment to categories is based on a fixed external standard (e.g., cost profile less than 0.5), reliability can be used to estimate misclassification probabilities after the fixed external standard is transformed to a percentile of the scale.
[9] McKethan et al., 2006.
[10] Milstein and Lee, 2007.

technical issues that were less important at higher levels of aggregation in hospitals and MCOs.[11,12]

One of these technical issues is the statistical reliability of a physician's performance measurement and how it affects the misclassification of physicians. While the use of more reliable measures is obviously important to all stakeholders, the meanings of reliability and misclassification in this context are sometimes unclear, and the methods for measuring both reliability and misclassification in practice may seem daunting to those who design and operate performance measurement systems. Addressing these needs is the focus of this report. This report builds on other RAND work on reliability and misclassification and has two main goals. First, it can serve as a tutorial for measuring reliability and misclassification. Second, it will describe the likelihood of misclassification in a situation not addressed in our prior work in which physicians are categorized using statistical testing. A 2010 paper published in the *New England Journal of Medicine* and its accompanying technical report described the use of reliability and misclassification analysis to characterize physician costs.[13,14] The 2009 technical report addresses the calculation of reliability primarily for dichotomous process quality measures and composites built from multiple measures of this type.[15]

At the outset, it is important to clarify terminology. We will use the term *performance measurement* to refer generically to measures of utilization, process, or outcomes aggregated across patients and reported at the physician level for the purpose of comparing performance. Within this broad label, there are many possible domains of measurement, including quality of care, patient satisfaction, and costs. Each of these domains can be further divided into subdomains. For example, quality of care may be divided into outcomes and process of care. We use the term *cost profiles* to refer to measures that examine the costliness of a physician relative to his or her peers. These types of measures have also been called efficiency, cost-efficiency, or relative resource use measures.[16] Note that this use of *efficiency* is not the classical meaning from

[11] Associated Press, 2007.
[12] Kazel, 2005.
[13] Adams et al., 2010b.
[14] Adams et al., 2010a.
[15] Adams, 2009.
[16] Hussey PS, de Vries H, Romley J, et al., "A Systematic Review of Health Care Efficiency Measures," *Health Services Research*, Vol. 44, No. 3, June 2009, pp. 784–805.

economics. Here, *efficiency* is just defined as costs relative to average costs. Examples in this technical report are drawn primarily from costs and process quality measures, but the findings can be expanded to other domains using the same methods or minor variations.

Many important decisions must be made in developing performance measures for physician profiles. Among them are attribution of patient data to physicians, case mix adjustment methods, and assessment of the validity of the measurement approach. Even though each of these decisions can have a significant effect on reliability and misclassification, these issues will only be addressed in passing here. The focus of this report is on how to quantify reliability and misclassification after these decisions have been made.

The report is structured as follows. First, we define reliability and describe how reliability relates to misclassification. We then describe how to calculate misclassification in various categorization scenarios (e.g., categorize the top 25 percent of physicians as high-performance versus categorize those physicians that are statistically different from the average physician). In the appendixes we provide more technical detail on how to measure reliability with related program code as well as a set of lookup tables that can be used to obtain the rate of misclassification associated with a reliability estimate under various scenarios.

A Statistical Model for Performance Measures

What Is Reliability?

We have discussed the definition of reliability in some detail in a recent paper and accompanying technical appendix.[17] We review that material here.

What is reliability? It is a key characterization of the suitability of a measure for profiling. It describes how confidently we can distinguish the performance of one physician from another. Conceptually, it is a ratio of signal to noise. The signal in this case is the proportion of the variability in measured performance that can be explained by real differences in performance. A

[17] Adams et al., 2010a.

reliability of zero implies that all the variability in a measure is attributable to measurement error. A reliability of one implies that all the variability is attributable to real differences in performance.

There are three main drivers of reliability: sample size, differences between physicians, and measurement error. At the physician level, sample size can be increased by increasing the number of patients in the physician's data as well as increasing the number of measures per patient. Differences between physicians are largely outside the performance measurement system's control because they are determined by the heterogeneity of the selected population. However, inclusion of measures that are homogeneously high or low will decrease differences between physicians. Some level of measurement error may be unavoidable, but work on standardizing data elements or collection methods can provide more-precise measurement through reductions in this component of variability.

Several mathematical identities or derived formulas follow from the definition of reliability. They can provide insight into the meaning and consequences of reliability and provide methods of calculating reliability measures. These include the intraclass correlation coefficient and the squared correlation between the measure and the true value. Unfortunately, the use of very different methods to estimate the same underlying concept can be a source of some confusion among practitioners trying to use these methods.

Reliability is analogous to the R-squared statistic as a summary of the predictive quality of a regression analysis. High reliability does not mean that performance on a given measure is good, but rather that one can confidently distinguish the performance of one physician from another. Although most directly useful for relative comparisons, reliability can also be useful to understand for absolute comparisons to fixed standards.

The reliability and misclassification calculations in this technical report are motivated by a statistical model of physician performance measurement that suggests there is a "true" score for a physician, but we see an imprecise version of this score because it is corrupted by measurement error (or "noise"). To understand what reliability means, it is important to understand how to think about the nature of this measurement error.

We should first mention several types of measure design errors that are sometimes thought of as causes of noise but are not the type of noise we are considering here. First, the measures could be poorly conceived and could not be measuring what we want to measure. This is a form of validity and is not the noise that is used in thinking about reliability. Second, there could be data errors (e.g., bad record keeping) that differentially affect some physicians. This is a potential source of bias but is not the type of noise we are considering in reliability. Third, variation in the severity of patients' underlying health status may vary from physician to physician. This is another potential source of bias that may be addressed with case-mix adjustment methods. None of these or other sources of systematic bias or lack of validity are part of the noise that we incorporate in reliability estimation. In fact, reliability calculations assume that the bias and validity issues have been resolved before moving on to reliability estimation.

The noise considered in reliability calculations can be appreciated by thinking about what could have occurred instead of the observed data in light of an underlying "true answer" that the performance measure is attempting to estimate. For a performance measure that is calculated on a year's data (e.g., the fraction of a physician's patients that had their blood pressure checked) this would be the variation in potential scores that we would have expected if the year were repeated again. This way of thinking about the noise, or variation, is called the superpopulation model in statistical sampling theory.[18] Although the notion that we are conceptually "rerunning" the same year repeatedly can seem strange at first, this is a standard approach that provides a logic for estimating measures of variability that do not require multiple years of data. Measures of physician clinical quality, patient experience, peer review, medical errors, and utilization have been evaluated for their reliability.[19,20,21,22]

Now we can be more precise and formal about the definition of the reliability of a physician profile. A common basic definition is:

---

[18] Hartley HO and Sielken RL, Jr., "A 'Super-Population Viewpoint' for Finite Population Sampling," *Biometrics*, Vol. 31, No. 2, June 1975, pp. 411–422.
[19] Safran DG, Karp M, Coltin K, et al., "Measuring Patients' Experiences with Individual Primary Care Physicians: Results of a Statewide Demonstration Project," *Journal of General Internal Medicine*, Vol. 21, No. 1, 2006, pp. 13–21.
[20] Hofer TP, Bernstein SJ, DeMonner S, and Hayward RA, "Discussion Between Reviewers Does Not Improve Reliability of Peer Review of Hospital Quality," *Medical Care*, Vol. 38, No. 2, 2000, pp. 152–161.
[21] Hofer et al., 1999.
[22] Hayward RA and Hofer TP, "Estimating Hospital Deaths Due to Medical Errors: Preventability Is in the Eye of the Reviewer," *Journal of the American Medical Association*, Vol. 286, No. 4, 2001, pp. 415–420.

Reliability is the squared correlation between a measurement and the true value.

Or, in mathematical notation:

$$reliability = \rho^2(measurement, truevalue)$$

This would be easy to calculate if only we knew the true value! Most of the complications of reliability calculations come from various workarounds for not knowing the true value. Conceptually, the true value is the score the physician would receive if there were unlimited data available. The measures assigned to a provider can be considered a sample of the entire population of measures that we do not have available but which could in theory be assigned to the provider over repeated samples as different patients or indicators appear in the samples.

The most common way to make reliability an implementable quantity is to characterize it as a function of the components of a simple hierarchical linear model (HLM).[23] A simple two-level HLM separates the observed variability in physician scores into two components: variability between physicians and variability within a physician. The equivalent definition of reliability from this framework is:

$$reliability = \frac{\sigma^2_{between}}{\sigma^2_{between} + \sigma^2_{within}}.$$

Or, with a more intuitive labeling:

$$reliability = \frac{\sigma^2_{Signal}}{\sigma^2_{Signal} + \sigma^2_{Noise}}.$$

Or, made more specific to our setting:

$$reliability = \frac{\sigma^2_{provider-to-provider}}{\sigma^2_{provider-to-provider} + \sigma^2_{provider-specific-error}}.$$

At first glance, it is not obvious to most analysts why this version of reliability is equivalent to the basic definition. These ratios of variance components are equivalent to correlations in the same way that regression R-squareds are calculated from variance components but are equivalent to the correlation between the predicted and actual values. Correlations can affect the variance explained, and these relationships can be used to estimate correlations from variance components. The equivalence between the two formulas for reliability can be established by a simple mathematical proof, but the variance components formula allows for an easy calculation

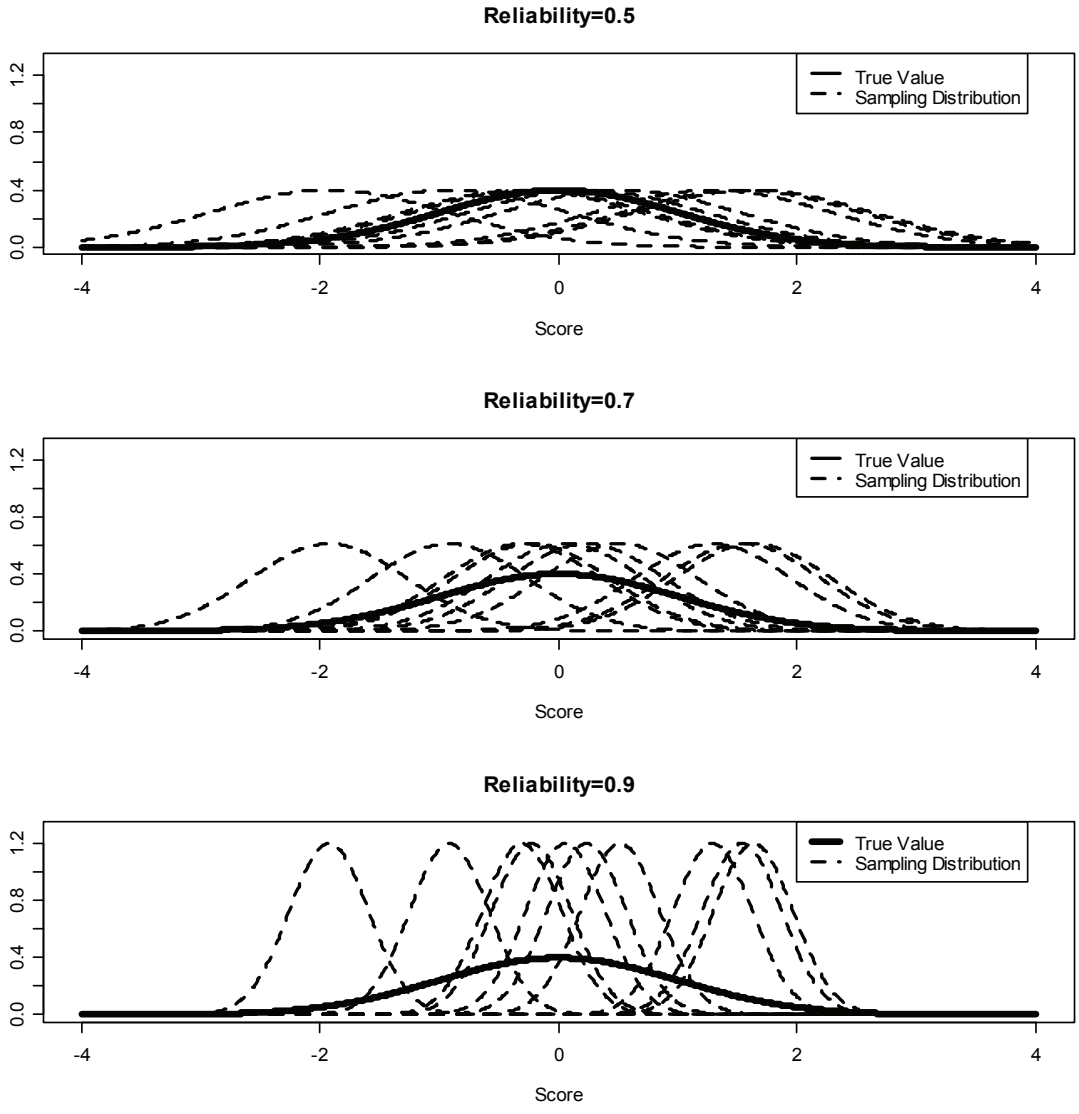
[23] Raudenbush SW and Bryk AS, *Hierarchical Linear Models: Applications and Data Analysis Methods*, Newbury Park, Calif.: Sage, 2nd ed., 2002.

of reliability because a simple two-level hierarchical model will estimate the two variance components required. The provider-to-provider variance is the variance we would get if we were able to calculate the variance of the true values. The provider-specific error variance is the sampling error. This sampling error can be determined from the sampling properties of the indicators. Conceptually, the HLM estimates the provider-to-provider variance by subtracting known measurement error variances from the observed variance of the provider scores. How to calculate reliability is taken up in Appendix A for both continuous and binary performance measures.

Although the model used is a simple two-level HLM, it is not estimated in the typical way. This is a consequence of the aggregate form of the data, which is not the typical HLM data form. Rather than using multiple observations within the physician as with the usual two-level HLM, we have a single score for each physician and its corresponding standard error. This aggregate data requires special software, as discussed later in this report. Although the special software issues can be challenging, there is a corresponding advantage to this approach. In many cases, only aggregated physician scores are available. The availability of a software solution for this problem makes it possible to implement these models even when the disaggregated data are not available. In particular, scores that have already been case-mix adjusted and aggregated may be used. An additional advantage of this approach can be the reduced computational burden of working with small aggregated data sets at the physician level.

Figure 1 shows the relationship between the physicians' score distributions and the underlying reliability. Each panel of Figure 1 shows the true distribution of the physicians' scores as a solid bell-shaped curve. The dashed bell-shaped curves show the sampling distribution (based on multiple observations for each physician) for ten physicians selected at random. At a reliability of 0.5, it is difficult to detect differences between physicians. At a 0.7 level of reliability, we can start to see differences between some physicians and the mean. At a reliability of 0.9, we can start to see significant differences between pairs of physicians. The sampling variation is the variation typically expressed by confidence intervals.

Figure 1: Relationship Between Reliability and Physicians' Score Distribution



Higher reliability increases the likelihood that you will assign a physician to the "right" group within the distribution of physician scores. Sample size, while often used as a proxy for reliability, may be insufficient to overcome small physician-to-physician variation. Therefore, simple minimum sample-size cut points (e.g., a rule that a physician's profile must include at least 30 events) may not guarantee that physicians are correctly labeled. Higher sample sizes are always useful. If two providers have a particular difference in scores, the ability to tell them apart depends only on their standard errors. But if provider-to-provider variances are small and there are small differences between providers, much larger sample sizes are required to distinguish one

provider from another. What changes with smaller provider-to-provider variances is that the typical differences between providers are smaller. Either bigger provider-to-provider variances or bigger sample sizes can achieve increased reliability. The relationship between reliability and the misclassification of physicians is explored in greater detail in a later section.

The statistical model that motivates reliability is that observed physician performance scores are subject to sampling error. The measures assigned to a provider can be considered a sample of indicators from a much larger "theoretical" set of indicators that are not available to observe. The true provider score is the score we would observe if we had this much larger set of indicators available. What we call *measurement error* here is the difference between this true score and the score we observe from the sample of indicators and observations that are available.

How Does Reliability Translate into Misclassification?

Fundamentally, reliability is the measure of the degree of confidence that performance results distinguish a physician from his or her peers. One concern is that, for most readers, interpreting reliability is not intuitive. Additionally, there is no agreement on a gold standard level of reliability in the context of provider profiling. In this section, we describe how we used reliability to estimate a more intuitive concept: the rate at which physicians are misclassified for a particular application of cost profiling.

The most commonly used applications of cost profiles (e.g., public reporting, pay for performance, tiering) require putting physicians into categories. Reliability can be used to calculate the probability that a physician will be correctly or incorrectly classified in a particular application. If assignment to categories (e.g., above-average costs) is based on relative comparisons (e.g., the top 10 percent of the distribution of physicians' scores), reliability can be used to estimate the probability of misclassification.[24]

In this section, we use a number of scenarios to illustrate how reliability can be used to estimate misclassification. We start with the simplest categorization system and some unrealistic

[24] If assignment to categories is based on a fixed external standard (e.g., cost profile less than 0.5) reliability can be used to estimate misclassification probabilities after the fixed external standard is transformed to a percentile of the scale.

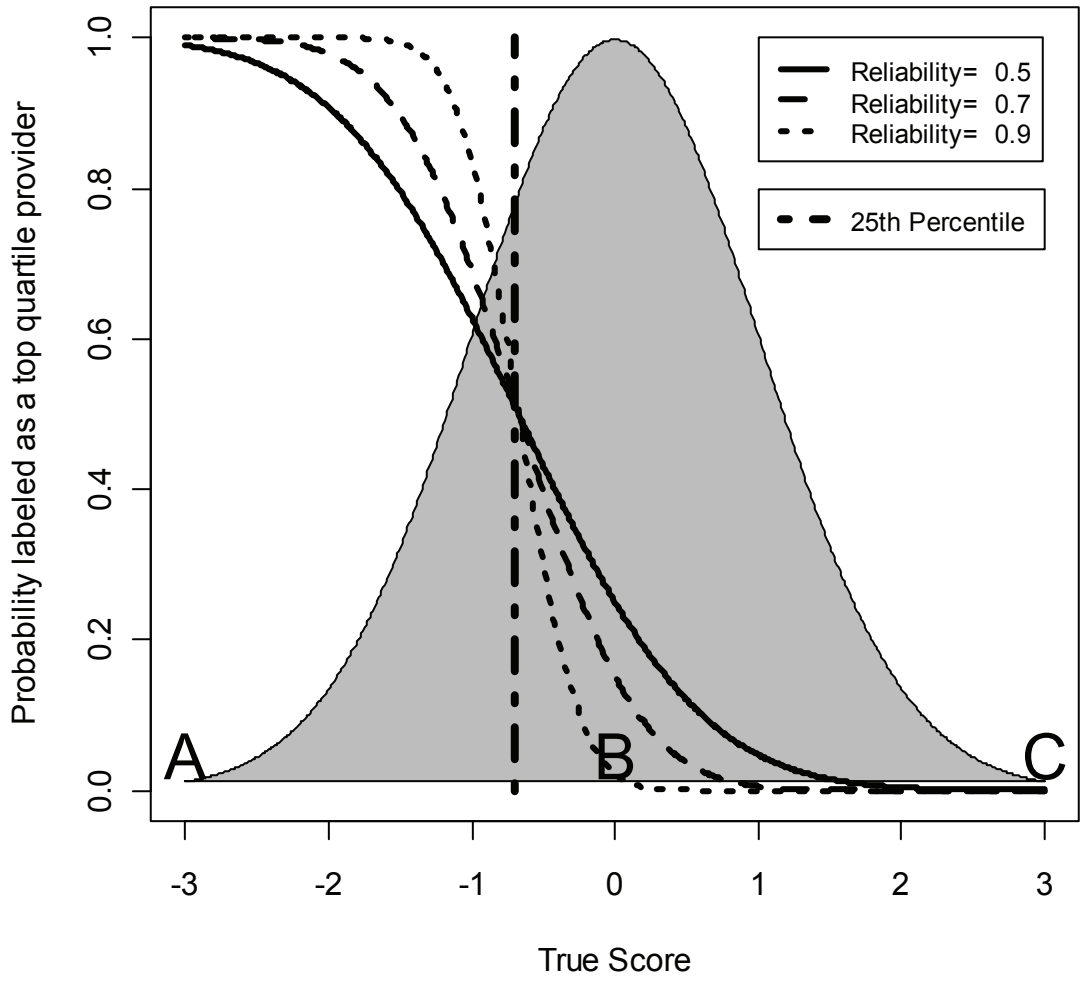
assumptions. In the following scenarios, we use more complex categorization systems and relax those assumptions.

Scenario 1: A Two-Category System in Which All Providers Have the Same Reliability and We Know True Cut Points in Distribution

In our first illustration, we assume that all physician cost profiles have the same reliability and that we know the true cut points. By "true cut points," we mean that we know the 25th percentile of the true distribution of physician profile scores in the absence of sample error. Obviously, we do not know the true cost profiles. Later, we will use a more realistic model that includes estimated cut points and mixtures of physician reliabilities. We first illustrate the relationship between misclassification and reliability using the simplest two-category system.[25] In this illustrative application of profiling, the 25 percent of physicians who have the lowest cost profiles are labeled as "lower cost" (in Figure 2 these are the physicians in each specialty in the left of the plot) and the remaining 75 percent are labeled as "not lower cost." This is a process often used by health plans to create what are termed "high-performance" networks. The illustration can easily be translated to other types of performance measures, including quality measures.

[25] Hussey et al., 2009.

Figure 2: Probability of Being Labeled as a Lower-Cost Physician Based on True Score and Reliability of Cost Profile Score



There are several points to emphasize. If a physician is far enough to the lower end of the distribution (shown as A in Figure 2), he or she will be labeled as lower cost at any level of reliability. At a reliability of 0.5 or 0.7, even an average physician (shown as B in Figure 2) can be labeled as lower cost with a probability of up to 25 percent. A physician who is clearly high cost (shown as C in Figure 2) has only a low probability of being labeled as lower cost at any reliability.

To understand the relationship between reliability and misclassification, it is useful to return to the components of between-physician and within-physician variation. If the between-physician variation is large, on average, physicians will be further from any given cut point. Large between-physician variation makes it easier to correctly classify physicians. Within-physician variation makes it harder to classify physicians correctly by making it less clear where the physician's true score is. Consider a physician with a particular true value, like the 75th percentile of the true provider distribution. Consider a cut point at the 50th percentile. We can calculate the probability that the provider will be correctly classified by looking at the fraction of a normal distribution that is above the 50th percentile (based on the within-provider variation). The fraction below the 50th percentile is the misclassification probability. If we calculate these probabilities for all possible true scores, we can use a weighted average of the probabilities to calculate an overall misclassification rate.

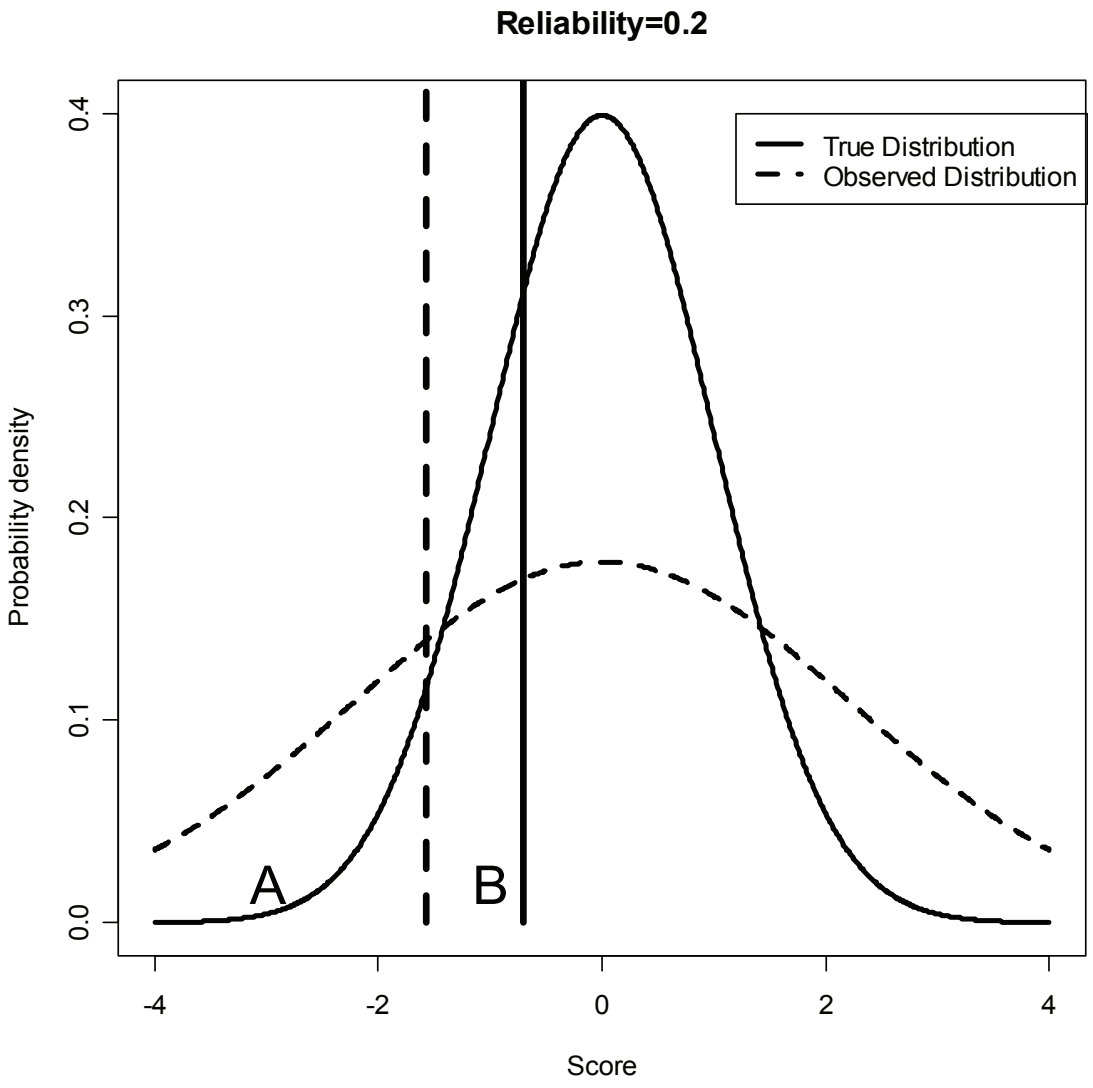
Scenario 1 demonstrates the relationship between reliability and misclassification where every physician has the same reliability and we know the physician's "true score." To calculate the misclassification rate in a more realistic situation, we need to address varying reliability and the fact that the cut point will change based on reliability. These "real world" details will be addressed in the next scenario.

Scenario 2: We No Longer Know the True Cut Points in Distribution and Instead Use Expected Cut Points

The first challenge is relaxing the assumption of a known cut point in the true physician distribution. Figure 3 illustrates the issue. The solid bell-shaped curve is the true score distribution. This graph is on the *Z* scale, running from roughly –2 to +2 standard deviations. The solid vertical line marks the 25th percentile of the true distribution. The dashed bell-shaped curve is what we would expect the observed distribution to look like for a reliability of 0.2 (for reliabilities less than 1, the observed distribution will be wider than the true distribution). The dashed vertical line marks the 25th percentile of the observed distribution. Note that this is a much lower percentile of the true distribution. It is a common misconception that the observed distribution of scores is the same shape, or width, as the true distribution. This is not generally true if reliability is less than perfect. In most cases the observed distribution will be wider than

the true distribution estimated with a hierarchical model. One of the major sources of misclassification can be that the selection of the cut point is driven by the wider dispersion of the observed distribution, which includes observations of physician scores that have low reliability because of sample size or low provider-to-provider variance.

Figure 3: Probability of Being Labeled Lower Cost Based on True Score and Reliability of Physician's Cost Profile



The physician labeled as A is far enough into the left tail of the distribution that he or she will be to the left of the misestimated cut point if his or her reliability is high enough. The difficult problem is physician B, where the situation is somewhat counterintuitive. Because physician B's true value lies between the true and the observed 25th percentiles, the better the reliability of physician B's score, the *more* likely he or she is to be misclassified.

There is an important point to emphasize. One of the major sources of misclassification is that selection of the cut point is driven by the wider dispersion of the physician score distribution, which includes observations of physician scores that have low reliability because of sample size or low provider-to-provider variance. The 25th percentile line based on the observed distribution is at a much lower percentile of the equivalent 25th percentile line, based on true distribution. For example, let us return to Figure 3 and physician B. Because physician B's true value lies between the true and the observed 25th percentiles, the better physician B's reliability, the more likely he or she is to be misclassified. This is in contrast to physician A, who is far enough into the left tail of the distribution that it is unlikely the physician will be misclassified.

Table 1 presents traditional names for various summaries of correct classification and misclassification rates. We will refer to the true positive, false positive, false negative, and true negative probabilities as the four basic classification probabilities. In addition to these traditional summaries, there are two additional physician misclassification quantities that are useful for policy purposes. The first question is: "What is the probability that a physician who is actually low cost is labeled as higher cost?" This is just 1 minus sensitivity. The second question is: "What is the probability that a physician labeled as lower cost is actually higher cost?" This is just 1 minus the positive predictive value. These are the most salient examples of a mistake that is bad for physicians and a mistake that is bad for patients.

Table 1: Definitions of Classification Probabilities

		True		
		Low Cost (Positive)	Higher Cost (Negative)	
Observed	Low Cost (Positive)	True Positive (TP)	False Positive (FP)	Positive Predictive Value: TP / (TP + FP)
	Higher Cost (Negative)	False Negative (FN)	True Negative (TN)	Negative Predictive Value: TN / (FN + TN)
		Sensitivity: TP / (TP + FN)	Specificity: TN / (FP + TN)	

We must know four things to calculate the misclassification probabilities for specialist physicians with a particular physician reliability: the mean for the physicians' specialty, the provider-to-provider variance of the specialty, where the observed cut point lies in the estimated true distribution, and the reliability. To illustrate the probability of misclassification, we will use the data from family and general practice physicians in Massachusetts, which we reported in the 2010 *New England Journal of Medicine* paper.[26] This specialty had a median reliability of 0.61 for its cost profile scores. In this analysis, we have done the reliability and misclassification calculations separately for each physician specialty. Separate calculations by specialty allow separate estimates of both the physician-to-physician and within-physician variances. Here, we will use the single specialty as an example.

First, we need to determine where the observed value lies in normal distribution of true scores. Recall that the true score distribution is centered at the value we call the true mean (which can be estimated from the sample mean) and has a variance equal to the variance of the between-provider variance. To find the location of the observed value in a normal distribution, we need to

[26] Adams et al., 2010b.

estimate its deviation from the center, the mean, in terms of the standard deviation units. We do this by calculating the Z score:

$$Z = \frac{Observed_{25th} - Mean}{\sqrt{\sigma^2_{provider-to-provider}}}$$

We can now use a standard normal distribution table to obtain the percentile of the true score distribution that corresponds to the 25th percentile of the observed scores. For the family practice physicians, the observed 25th percentile cost profile is 0.856, the mean is 1.049, and the provider-to-provider variance is 0.0224. Filling in the family practice values, we get:

$$Z = \frac{0.856 - 1.049}{\sqrt{0.0224}} = -1.29$$

Although we could stop here and move on to calculating misclassification probabilities using this Z value, it is convenient to go one step further and turn this into a percentile of the Z distribution. Percentiles are easier to think about than Z values. Since most cut points are expressed as percentiles, this makes understanding the difference between the goal and actual percentiles clearer. A Z value of –1.29 corresponds to the 10th percentile of the Z distribution. This value can be obtained from the Z distribution tables in the back of most statistics textbooks. Alternatively, the NORMSDIST function in Microsoft Excel may be used. Note that the 10th percentile is further from the mean of the distribution than the 25th goal percentile. In general, the observed percentile will result in an estimated true percentile than is further from the mean than desired. It may be counterintuitive that after turning the observed 25th percentile into a Z score and then back into a percentile that it becomes the 10th percentile of the true distribution. This is a consequence of the true variance (estimated by the HLM) being smaller than the observed variance, since the noise always makes the distribution wider.

We can now calculate the probabilities of the various outcomes in Table 1. The mathematical and computational steps in this process are provided in Appendix B. In the appendix, we show how to produce a table to look up the misclassification probabilities for any cut point and any goal percentile. Table 2 illustrates our family practice example.

Table 2: Four Basic Classification Probabilities for Family Practice Physicians with a Reliability of 0.61

		True	
		Low Cost (Positive)	Higher Cost (Negative)
Observed	Low Cost (Positive)	0.12	0.04
	Higher Cost (Negative)	0.13	0.71

This table gives the four basic classification probabilities for any family and general practice physicians in Massachusetts who have a reliability of 0.61. These classification probabilities will be different among different specialties or in other populations of physicians, because the mean score and the provider-to-provider variance will be different. Note that the top row here does not sum to 0.25. This is because the observed cut point is for a mixture of reliabilities, not the 0.61 reliability.

Once we can calculate the four basic classification probabilities for physicians with any given reliability, it is simple to calculate the classification probabilities for entire populations of physicians. For each physician in the sample, we look up his or her specific classification probabilities and then average across all physicians. For example, the Massachusetts data has 1,065 family practice physicians. Using the tables described in Appendix B, we looked up the classification probabilities for the 1,065 family practice physicians based on the reliability of each. Table 3 lists the results for the specialty as a whole.

Table 3: Four Basic Classification Probabilities for Family Practice Physicians

		True	
		Low Cost (Positive)	Higher Cost (Negative)
Observed	Low Cost (Positive)	0.15	0.10
	Higher Cost (Negative)	0.10	0.65

The four basic classification probabilities can be used to calculate other summaries (e.g., positive predictive value). The estimated misclassification rate for this example is 20 percent. Note that the numbers in this table are slightly different than those that appear in our earlier work. This is a consequence of the newer misclassification estimation algorithms developed for this report.

Scenario 3: We Move from a Two-Category System to a Three-Category System

Our next example is of a three-tiered system. This three-tiered system identifies the 25 percent of physicians with the lowest cost profiles as tier 1 and the next lowest cost profile 25 percent of physicians as tier 2. A more elaborate table is required for this problem. As before, the details of calculating the probabilities are in Appendix B. Using the same family practice example with reliability of 0.61, we now need to translate two cut points into the estimated true Z scale. The 25th percentile cut point is unchanged. The 50th percentile cut point corresponds to a Z value of zero. Table 4 shows a basic three-category classification system.

Table 4: Basic Classification Probabilities for Family Practice Physicians with a Reliability of 0.61 in a Three-Tiered System

		True		
		Lowest Cost	Low Cost	Higher Cost
Observed	Lowest Cost	0.12	0.03	0.01
	Low Cost	0.11	0.13	0.10
	Higher Cost	0.02	0.09	0.39

As in the previous example, it is straightforward to match tables like this one by reliability to the Massachusetts family practice physicians. Table 5 presents the overall results.

Table 5: Basic Classification Probabilities for Family Practice Physicians in a Three-Tiered System

		True		
		Lowest Cost	Low Cost	Higher Cost
Observed	Lowest Cost	0.12	0.04	0.04
	Low Cost	0.10	0.12	0.08
	Higher Cost	0.03	0.09	0.38

The three cells on the diagonal are correctly classified. This table can be used to calculate error rates or any other summary of interest.

A three-by-three table like Table 5 does not have traditional labels, such as false positive or false negative. But we can compare tables of different dimensions by summarizing the misclassification rates. The overall misclassification rate in Table 5 is 38 percent. In contrast, in the two-category system (Table 3) the overall misclassification rate is 20 percent. Some readers may be surprised to see that finer classifications do not help with misclassification. The mistaken intuition may be that finer misclassification is more like a continuous measure that has the full

information. But the decrease in lost information is more than offset by the addition of categories into which physicians can be misclassified.

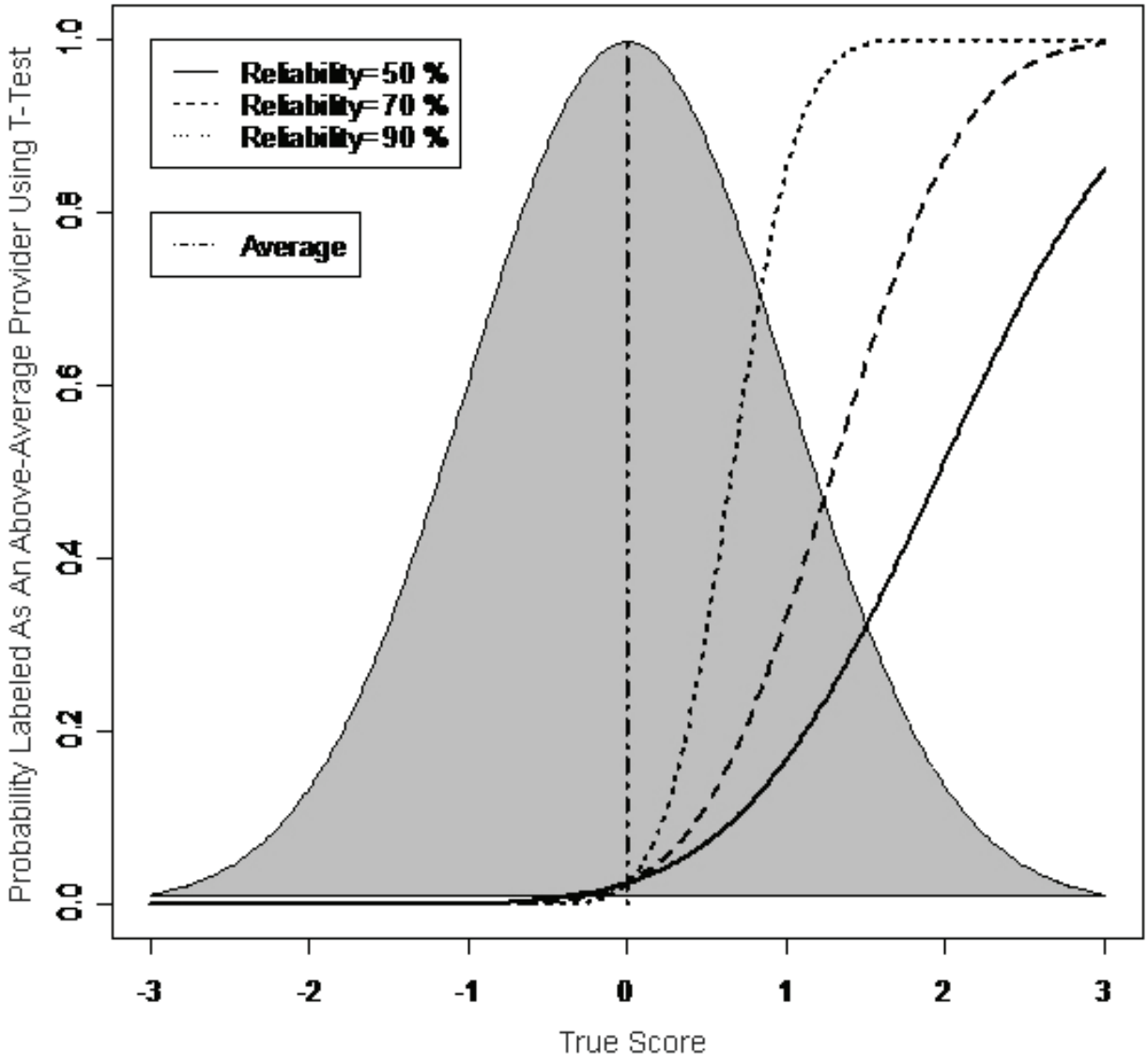
Scenario 4: We Categorize Physicians Using Statistical Testing Instead of Cut Points

The most common way to incorporate statistical uncertainty into a physician categorization system is the use of T or Z tests. These tests attempt to ensure that there is reasonably strong evidence that the physician's performance score is different from the average before identifying him or her as high- or low-performing. In this section, we will use the QAT-C quality measurement system to profile providers. This system is described elsewhere and consists of a collection of binary (pass or fail) quality of care measures.[27] The physician's score is his or her average pass rate across all of the measures.

Figure 4 shows the probability of being labeled as a high-performing physician at various values of the true score distribution for reliabilities of 50 percent, 70 percent, and 90 percent if a statistical test versus the mean is used. Note that the statistical test is quite stringent, in the sense that it is less likely to flag a physician as high-performing than is the percentile cutoff method. Using the terminology of misclassification, statistical testing will decrease false positives. This is the reason why the application of statistical tests is often thought of as a more rigorous or "conservative" approach. They are conservative in the sense that they do not flag a physician as different from the mean without strong evidence to do so. On the other hand, they will typically increase false negatives. There is a higher likelihood that a truly high performing physician will be misclassified as average. This tradeoff is unavoidable and is a consequence of limitations in the amount of information from the available observations.

[27] McGlynn EA, Asch SM, Adams J, Keesey J, Hicks J, DeCristofaro A, and Kerr EA, "The Quality of Health Care Delivered to Adults in the United States," *New England Journal of Medicine*, Vol. 348, No. 26, 2003, pp. 2635–2645.

Figure 4. Physician Labeled as an Above-Average Provider by a T Test



We must know three things to calculate the misclassification probabilities for physicians with a particular reliability: the provider-to-provider variance of the specialty, where the testing point lies in the estimated true distribution, and the reliability. Let us again consider Massachusetts family practice physicians as an illustration of how to determine the misclassification probabilities. The median reliability is 0.84 for the quality scores. Note that this is higher than the reliability for the cost scores. In the Massachusetts data in general, the quality scores have

higher reliability than the cost scores (although this is not inevitably the case). First, we need to determine where the testing point lies in the estimated true distribution. Most statistical tests are tests relative to the mean. This is very convenient for misclassification problems, since the mean of the standard normal distribution has a Z value of zero. This allows us to skip the steps used in the earlier examples to find the test point in the estimated true distribution. If the test was not versus the mean, the earlier steps would need to be repeated. Tests against points other than the mean appear to be rare in physician profiling applications.

We can now calculate the probabilities of the various outcomes in Table 6. The mathematical and computational steps in this process are provided in Appendix B. In the appendix, we show how to produce a table to look up the misclassification probabilities for any cut point and any goal percentiles. (Note that a higher value of the quality performance measure is now "good," so the table has been relabeled and higher quality scores are now labeled as positive.) Here is the table for our family practice example:

Table 6: Four Basic Classification Probabilities for Statistical Testing of Family Practice Physicians with a Reliability of 0.84

		True	
		Lower Quality (Negative)	Higher Quality (Positive)
Observed	Lower Quality (Negative)	0.50	0.35
	Higher Quality (Positive)	0.00	0.15

This table gives the four basic classification probabilities for any family practice physicians in Massachusetts who have a reliability of 0.84. This table makes it easy to see the conservative nature of statistical testing. Statistical testing tolerates higher rates of false negatives in order to avoid false positives. For a statistical test, we have assumed the true definition as greater than the

mean. This is most consistent with the literal definition of a statistical test. Other definitions, such as keeping the 25 percent definition from the earlier examples, are certainly possible.

Once we can calculate the four basic classification probabilities for physicians with any given reliability, it is simple to calculate the four basic classification probabilities for entire populations of physicians. All that is required is to average the reliability-specific tables together. For example, the Massachusetts data has 1,065 family practice physicians. Using the statistical testing tables described in Appendix B, we can match the four basic classification probabilities to a data set with the 1,065 family practice physicians by matching on reliability. The four basic classification probabilities can then be averaged across the data set to obtain specialty averages (Table 7).

Table 7: Four Basic Classification Probabilities for Statistical Testing of Family Practice Physicians

		True	
		Lower Quality (Negative)	Higher Quality (Positive)
Observed	Lower Quality (Negative)	0.50	0.34
	Higher Quality (Positive)	0.00	0.16

These results are very similar to the results for the median reliability. The zero cell is only zero to the presented digits. These four basic classification probabilities can be used to calculate other summaries (e.g., positive predictive value).

Contrary to a common misperception, statistical testing is not uniformly superior to cut point testing. From a misclassification point of view, the price of lower probabilities of labeling a

below-average physician as above average is that fewer above-average physicians are labeled as high-performing.

Conclusions

Reliability is a topic of increasing importance in profiling applications. Reliability may be particularly important when the risks of misclassifying a physician are high—for example, if the classification of the physician were to lead to sanctions. We hope this tutorial will reduce some of the confusion about what reliability is and how it relates to misclassification. The methods we describe enable explicit calculation of the misclassification probabilities of the system. Ultimately, whether reliability is "good enough" for any given application must be judged by the developers and users of a profiling system. Once a specific system has been proposed, the methods presented here should enable an evaluator to calculate the reliabilities and, consequently, the misclassification probabilities for use by policymakers.

From the examples here, it should be clear that low reliability is problematic across many commonly used classification systems. Even statistically based methods, for all their rigor, can only be used to compensate for low reliability by incurring as a tradeoff an increase in the false negative rate.

We hope that our appendixes assist others in calculating reliability and misclassification rates with their own data. Such evaluations will allow the debate to focus on means of improving the reliability of cost and quality measures.

Appendix A: Calculating Reliability

In this appendix, we present two ways of calculating reliability for physician performance measures. One method is based on a simple two-level hierarchical linear model (HLM) and is appropriate for normally distributed continuous data. It is also the easiest model to use for composites of binary performance measures that can be approximated by a continuous normal distribution. This method can even be used approximately for binary performance measurement data. The second method is only for binary performance measures like QAT-C or HEDIS measures. This method is based on the beta-binomial model, a simple two-level model for binary data with a beta distribution at the second level. This method has the advantage of very stable computational properties as well as a useful interpretation of the estimated beta distribution as prior counts of successes and failures.

For the continuous method, we will use cost profiling as a motivating example. Each provider has a score and a standard error for that score from the profiling system. See the *New England Journal of Medicine* technical appendix for details of how one such score can be constructed.[28] In this case, we estimate reliability as a function of the components of a simple HLM.[29] A simple two-level HLM separates the observed variability in physician scores into two components: variability between physicians and variability within a physician. The equivalent definition of reliability from this framework is:

$$reliability_{MD} = \frac{\sigma^2_{physician-to-physician}}{\sigma^2_{physician-to-physician} + \sigma^2_{physician-specific-error}}$$

For each specialty we fit a two-level HLM to estimate the physician-to-physician variance:

$$O/E \sim Normal(\mu_j, \sigma^2_{physician-specific-error})$$

$$\mu_j \sim Normal(\mu, \sigma^2_{physician-to-physician})$$

Since each physician cost profile score potentially has a different variance, this must be incorporated into the modeling. This is sometimes referred to as a "variance-known" HLM estimation problem. This type of problem is most often seen in meta-analysis applications. The estimated variances are incorporated into the model fit to address this problem. It is possible to

[28] Adams et al., 2010a.
[29] Adams, 2009.

get a boundary solution for the estimate of the physician-to-physician variance corresponding to an estimate of zero. The interpretation of this result is that there is no evidence of physician-to-physician variance for the specialty. This results in zero reliability for all physicians in the specialty.

Variance-known problems can be tricky to compute. The most basic HLM software may not have a way to incorporate the known variances into the problem. A useful approach in SAS is to use the gdata argument with proc mixed.

The SAS code is as follows:

```
data gdata;
  set scoredata;
  col = _n_;
  row = _n_;
  value = errorvariance;
  keep col row value;
run;
proc mixed data=scoredata METHOD=REML ;
  class physician;
  model score =;
  random physician / gdata=gdata ;
run;
```

In this code the composite is “score,” and the error variance of score is “errorvariance.” The gdata data set is the mechanism for telling proc mixed about the different error variances for each physician. Once the physician-to-physician variance estimate is obtained, the standard formulas can use the estimate and the error variance estimates to calculate provider-specific reliabilities.

Our second example is how to calculate reliability for simple binary measures like HEDIS. This is covered in more detail elsewhere.[30] In this section, we discuss how to estimate these parameters from real data sets. There are three steps in the process:

1) Build a data file of the proper form for physician-to-physician variance estimation.
2) Use the Betabin SAS macro to estimate the physician-to-physician variance.
3) Use the physician-to-physician variance estimate and the physician-specific information to calculate the physician-specific reliability scores.

To use the Betabin SAS macro, the data must be in the form of one record per physician. The records should include a physician ID, the number of eligible events (denominator), and the number of events passed (numerator). Here are the first few records of the beta55 file used in our example:

Physician	eligible	passed
1	10	6
2	10	4
3	10	8
4	10	5
5	10	5
6	10	6
7	10	6
8	10	8

The second step is to get an estimate of the physician-to-physician variance. The best way we have found so far is a publicly available SAS macro:

MACRO BETABIN Version 2.2, March 2005
Summary: Fits a beta binomial model
Author: Ian Wakeling, Qi Statistics
Website: www.qistatistics.co.uk

[30] Adams, 2009.

As with all free software, caveat emptor! We have tested this by simulating data sets like beta55, including different values of alpha and beta, as well as cases with varying sample sizes by physician. The macro has always reproduced the known simulation values within the bounds of statistical error.

Here is the macro call for beta55:

```
%betabin(data=rel.beta55collapsed,ntrials=eligible,nsucc=passed)
```

Table A.1 shows the relevant portion of the macro output (much of the output has been deleted here).

Table A.1: BETABIN Macro, Beta-Binomial Model Parameters

Parameter	Estimate	Standard Error	T Value	Pr > | T |	Alpha	Lower	Upper
Mu	0.5112	0.006891	74.18	<0.0001	0.05	0.4977	0.5247
Alpha	4.5865	0.4096	11.20	<0.0001	0.05	3.7837	5.3893
Beta	4.3862	0.3908	11.22	<0.0001	0.05	3.6201	5.1524

Mu is the model's estimate of the pass rate. Note that the alpha and beta estimates are within two standard errors of the values used to generate the data.

The third step is to use the physician-to-physician variance to calculate the reliabilities. These values are what we need to calculate the reliabilities from an observed data set. The following formulas show how they are used.

Here is the formula for the physician-to-physician variation:

$$\sigma^2_{provider-to-provider} = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}.$$

Then just plug in the numbers from the SAS output:

$$\sigma^2_{provider-to-provider} = \frac{4.5865 * 4.3862}{(4.5865+4.3862+1)(4.5865+4.3862)^2} = 0.025 \quad .$$

To get the reliability, we can start with the original formula:

$$reliability = \frac{\sigma^2_{provider-to-provider}}{\sigma^2_{provider-to-provider} + \frac{p(1-p)}{n}}.$$

Then we can plug in our physician-to-physician variance estimate:

$$reliability = \frac{0.025}{0.025 + \frac{p(1-p)}{n}}.$$

For each physician, we will need to use an estimate of their pass rate:

$$\hat{p} = \frac{\#passed}{\#events}.$$

Plugging this into the reliability formula, we get:

$$reliability = \frac{0.025}{0.025 + \frac{\hat{p}(1-\hat{p})}{n}}.$$

Consider a physician with a score of 7 out of 10. The pass rate would be 70 percent. We can plug this into the formula and get:

$$reliability = \frac{0.025}{0.025 + \frac{0.7(1-0.7)}{10}} = 0.54.$$

It is worth underscoring that every physician gets his or her own rate estimate. Consequently, even with equal sample sizes, reliability varies from physician to physician.

Appendix B: How to Produce Reliability-Misclassification Tables for Particular Physician Populations and Classification Systems

In this appendix, we detail how to calculate the misclassification probabilities for two- and three-category observed cut point systems and a statistical testing system. The SAS code required to perform the calculations is available from the authors. This appendix is written to provide more explicit mathematical detail for readers who are interested. We presume at least an advanced undergraduate course in mathematical statistics as background for this appendix.

Misclassification Probabilities for a Two-Category Profiling System

For any provider reliability, the probability of being correctly or incorrectly classified is a function of the goal percentile cut point and the observed percentile cut point in the estimated true score distribution. In the example used in this report, family practice physicians had a goal percentile of 25 percent and an observed percentile of 10 percent.

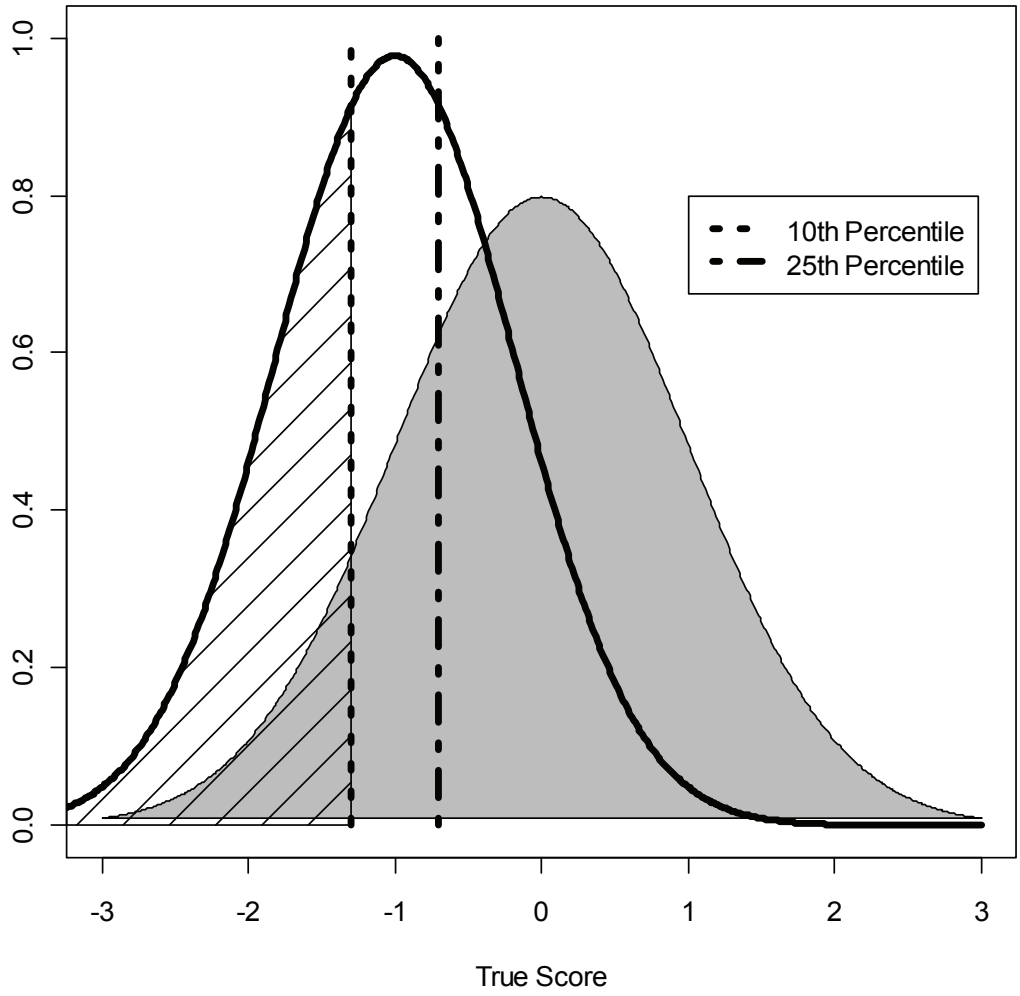
The estimated true score distribution comes from the two-level HLM used to calculate the physician-to-physician variance. The misclassification calculations assume this model is correct for calculation purposes. This is a very simple model based on normal distribution assumptions. It assumes that the physician-to-physician variance is constant. If, for example, high volume providers have a different physician-to-physician variation than low volume providers, the simple two-level HLM will not be adequate. In this case the physician population could be divided into two or more subsets based on volume, and the methods described here could be applied in each subset and then aggregated. Conceptually, other physician-level covariates could affect provider-to-provider variation and necessitate the use of more elaborate models than are used here.

The calculation of misclassification probabilities for a given reliability is conditional on the estimated true distribution. For each point on the true distribution, the correct and misclassification probabilities are calculated by integrating the sampling error distribution of physicians with the given reliability over the correct and incorrect classification regions. These values are then integrated with respect to the estimated true distribution to get correct and

misclassification probabilities for physicians with the given reliability. Although it is possible to do these calculations using the original physician-to-physician variance and physician-specific error variances, working with reliabilities makes it possible for us to use a standard normal distribution as the estimated true distribution. This ultimately makes it possible for us to produce standard tables that can be used to estimate misclassification rates in different settings without rerunning the calculations presented here.

Figure B.1 illustrates the process. The gray bell-shaped curve represents the estimated true distribution for the specialty. The taller bell-shaped curve is the sampling distribution for a physician with a reliability of 0.6 whose true score is between the 10th and 25th percentiles. This physician is in the best 25 percent of his or her specialty. The area under this provider's bell-shaped curve that is cross-hatched is the probability that this provider will be labeled as high performing.

Figure B.1: Correct Classification for a Family Practice Physician Who Is in the Best 25 Percent



The probability represented in Figure B.1 can be calculated for any value in the true distribution. For values greater than the 25th percentile, the area to the left of the observed cut point will be the *misclassification* probability.

Using the intuition from Figure B.1, we can write the integrals for the specialty average classification probabilities. The first step is to calculate the correct classification probability for a provider who is below the goal percentile:

$$TruePositive = \int_{-\infty}^{cutpo\,\mathrm{int}} \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} dx \quad \text{where } \sigma^2 = \frac{1}{\mathrm{Re}liability} - 1$$

This is just the cumulative normal distribution.

To get the average true positive rate for physicians with this reliability, we now average this quantity over the portion of the estimated true distribution where the physicians meet the goal in the true distribution. In the example in Figure B.1, this is the portion of the distribution below the 25th percentile of the gray bell-shaped curve.

$$AverageTruePositive = \int_{-\infty}^{goalpercentile} \left(\int_{-\infty}^{cutpo\,\mathrm{int}} \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} dx \right) \frac{e^{-\frac{(z)^2}{2}}}{\sqrt{2\pi}} dz$$

The average false negative can be obtained by subtracting this from 1.

The average false positive is a similar calculation with the integral over the remainder of the distribution:

$$AverageFalsePositive = \int_{goalptile}^{\infty} \left(\int_{-\infty}^{cutpo\,\mathrm{int}} \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} dx \right) \frac{e^{-\frac{(z)^2}{2}}}{\sqrt{2\pi}} dz$$

The average true negative can be obtained by subtracting this from 1.