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 research organization providing objective analysis and effective solutions that address the challenges facing the public and private sectors around the world.

Support RAND

Browse Books & Publications

Make a charitable contribution

For More Information

Visit RAND at www.rand.org

Explore RAND Health

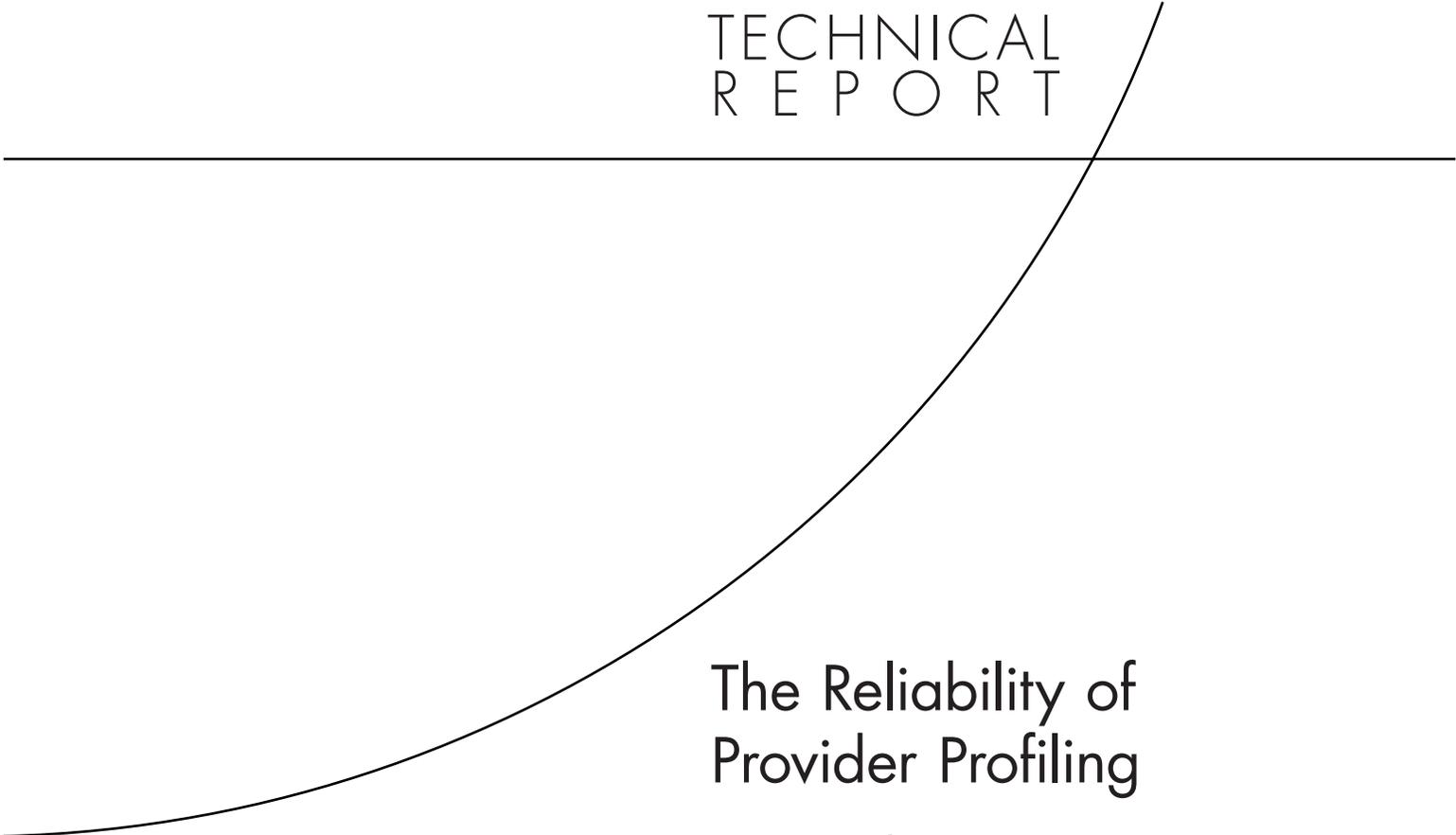View document details

Limited Electronic Distribution Rights

This 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.

The Reliability of Provider Profiling

A Tutorial

John L. Adams

RAND HEALTH

Preface

This technical report explains the use and implementation of reliability measurement for quality measures in provider profiling in health care. It provides details and a practical method of how to calculate reliability measures from the sort of data typically available. It also explains why reliability measurement is an important component of evaluating a profiling system. 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 quality health care.

Contents

Figures

Tables

Summary

Public and private purchasers and health plans are demanding more information about the quality and relative costliness of U.S. physicians to increase physician accountability and aid in value-based purchasing. Although performance measurement has been in place for some time in hospitals and managed care organizations (MCOs), the focus on physician profiling is a relatively new development. The inherent limitations of the available data at the physician level have brought to the fore technical issues that were less important at higher levels of aggregation in hospitals and MCOs. One of these technical issues is the reliability of a physician's performance measurement. Although a number of research efforts have evaluated quality measures for their reliability, the methods for doing so in practice may seem daunting to those designing and running performance measurement systems.

This report focuses on simple binary (pass/fail) measures. A patient may trigger one or more of these measures. These measures are then attributed to physicians, using various rules. HEDIS© measures are one example of the types of measures considered here.

Reliability is a key metric of the suitability of a measure for profiling because it describes how well one can confidently distinguish the performance of one physician from another. Conceptually, it is the 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.

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.

For analysts not focused on the technical intricacies of physician profiling, the sudden emergence of reliability as an important property may come as a surprise. After years of focus on sample size, power, and confidence intervals, reliability has recently come into focus as a key issue. The reason this has occurred is the strong relationship between reliability and the key policy issues in the physician profiling debate. This is a logical consequence of the way that stakeholders are using physician profiling information.

There has recently been more interest in public reporting and pay for performance. The focus has been on putting physicians into categories. High-performance networks and three-star public reporting systems have emerged. A three-star system might give one star to physicians in the bottom quartile, two stars to physicians in the middle 50 percent of the data, and three stars to physicians in the top quartile.

Given the discrete categories that are being used, the big question becomes the probability of misclassification. If the categories are based on relative comparisons, reliability tells you most of what you need to know about misclassification in these systems. For a simple high-performance network system that flags a subset of the physicians as high-performing, there are two types of errors: (1) flagging a lower-performance physician as high-performance and (2) failing to flag a high-performance physician as high-performance.

In this report, we estimate reliability with a beta-binomial model. The beta-binomial is a natural model for estimating the reliability of simple pass/fail rate measures. There are also computational advantages to using the beta-binomial model, which is based on the beta distribution for the "true" physician scores. The beta distribution is a very flexible distribution on the interval from 0 to 1. The beta-binomial model assumes the physician's score is a binomial random variable conditional on the physician's true value that comes from the beta distribution.

This tutorial underscores that reliability is not just a property of a measure set but also depends what population is used to estimate the reliability. Whether a set of measures is useful for profiling providers depends on how different the providers are from one another. Measures that may be useful in one group of providers may not be useful in another group with little provider-to-provider variation. Similarly, as the providers under study increase their performance, the reliability may decrease if the provider-to-provider variance decreases over time. This is especially true as measures hit the upper limits of their ranges.

There are unanswered questions regarding how to calculate reliability for more complicated measures. Conceptual challenges remain, especially when multiple levels in the system may influence measure scores.

Introduction

Public and private purchasers and health plans are demanding more information about the quality and relative costliness of U.S. physicians to increase physician accountability and 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 physician profiling is a relatively new development. The inherent limitations of the available data 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 reliability of a physician's performance measurement. Although a number of research efforts have evaluated quality measures for their reliability, the methods for doing so in practice may seem daunting to those designing and running performance measurement systems.

I will use the term *performance measurement* to generically refer to scores reported at the physician level. I will use the terms physician and provider interchangeably. Within this broad label, there are many possible domains of measurement, including quality of care, patient satisfaction, and relative costs and efficiency. Each of these domains can be further divided into subdomains. For example, quality of care may be divided into outcomes and process of care. Because the Healthcare Effectiveness Data and Information Set (HEDIS) measures are a combination of process and intermediate outcome quality of care measures, I will use these types

---

[1] McKethan A, Gitterman D, Feezor A, Enthoven A. New directions for public health care purchasers? Responses to looming challenges. *Health Aff (Millwood)* 2006;25(6):1518-28.

[2] Milstein A, Lee TH. Comparing physicians on efficiency. *N Engl J Med* 2007;357(26):2649-52.

[3] Associated Press. Regence BlueShield, WA doctors group settle lawsuit. *Seattle Post-Intelligencer*. 2007; August 8.

[4] Kazel R. Tiered physician network pits organized medicine vs. United. *American Medical News*. 2005; March 7.

of measures as examples. But the findings can be expanded to other domains in the same way or with minor variations.

This report will focus on simple binary (pass/fail) measures. A patient may trigger one or more of these measures. These measures are then attributed to physicians with various rules. Event counts equal the number events triggered among the eligible patients (e.g., diabetic patients.) HEDIS measures are one example of the types of measures considered here. To the extent we consider composite scales they are scales constructed by combining these simple binary measures. These composites can be across all types of patient care or they could be specific to subsets of patient care (e.g., chronic care) or subsets of the patient population (e.g., diabetics.)

In this tutorial I will not address several other issues and challenges that arise in creating physician level scores.[5,6] These include

- reporting individual measure versus condition versus composite measures

- scoring individual measures as pass/fail (0/1) versus partial credit

- attributing care to a physician by assigning patients versus condition versus measures

- creating composite measures by average of the average versus the sum of passes divided by the sum of eligibilities versus all-or-nothing scoring

- whether to use case-mix adjustment

- the role of validity in developing measures.

---

[5] Hofer TP, Hayward RA, Greenfield S, Wagner EH, Kaplan SH, Manning WG. The unreliability of individual physician "report cards" for assessing the costs and quality of care of a chronic disease. *JAMA* 1999;281(22):2098-105.
[6] Reeves D, Campbell S, Adams JL, Shekelle PG, Kontopantelis E, Roland MO. Combining multiple indicators of clinical quality: an evaluation of different analytic approaches. *Med Care*. 2007;45(6):489-96.

Even though each of these choices can have a significant effect on the reliability of a measure, these issues will only be addressed in passing here. The focus of the tutorial is on how to measure reliability after these decisions have been made and a measurement strategy has been proposed.

What Is Reliability?

What is reliability? It is a key metric of the suitability of a measure for profiling because it describes how well one can confidently 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 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.

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.

Measures of physician clinical quality, patient experience, peer review, medical errors, and utilization have been evaluated for their reliability.[7,8,9,10]

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

> *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. In particular, developing lower bounds for the reliability in various settings is common in the literature.[11] For example, test-retest reliability is a common lower bound for this reliability.

One way to think about reliability that is helpful for researchers who are familiar with regression analysis is to consider a hypothetical regression model:

$$measurement = \beta_0 + \beta_1(truevalue) + \varepsilon \,.$$

---

[7] 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;21(1):13-21.

[8] Hofer TP, Bernstein SJ, DeMonner S, Hayward RA. Discussion between reviewers does not improve reliability of peer review of hospital quality. *Med Care* 2000;38(2):152-61.

[9] Hofer et al., 1999.

[10] Hayward RA, Hofer TP. Estimating hospital deaths due to medical errors: Preventability is in the eye of the reviewer. *JAMA* 2001;286(4):415-20.

[11] Fleiss J, Levin B, Paik M. *Statistical Methods for Rates & Proportions.* Indianapolis, IN: Wiley-Interscience; 2003.

The R-squared from this regression would be the reliability. Both of these characterizations of reliability are useful to build intuition. Unfortunately, neither of these views of reliability is particularly useful for calculating reliability in real single-period-of-observation problems.

The most common way to make reliability an implementable quantity is to characterize it as a function of the components of a simple hierarchical model (HLM).[12] A simple two-level HLM separates the observed variability in physician scores into two components, variability between physicians and variability within 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 look, it is not obvious to most analysts why this version of reliability is equivalent to the basic definition. 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 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 or measurement error. This measurement error can be determined from the sampling

---

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

properties of the indicators. For example, binomial distribution properties will be used for the pass/fail measures in this report. Conceptually, the HLM subtracts known measurement error variances from the over all observed variance of the provider scores to estimate the provider-to-provider variance.

To understand some of the special issues in applying reliability to physician profiling, it is useful to add another level of detail to this reliability formula. In particular, a closer look at what is in the error variance can provide insight:

$$reliability = \frac{\sigma^2_{provider-to-provider}}{\sigma^2_{provider-to-provider} + \frac{\sigma^2_{average-item-error}}{n}} \; .$$

In this equation, the provider-specific error variance has been rewritten as the average error variance for a single item (average item error) in a physician's score where n is the number of items. Here an item would be a single pass/fail event (0/1). This form of the equation makes it easier to see the typical form of the variance of a mean depending on the population variance and the sample size. This formulation also makes it easier to separate the effects of measurement error in the items from the number of items. In many measurement examples other than physician profiling (e.g., functional status questionnaires), this additional detail is less important since all of the respondents have the same n. In some important physician profiling problems the number of items (or patient outcomes) in the physicians' scores can vary widely from physician to physician.

The more-detailed formula makes it easier to understand some common misconceptions about reliability, which is often mistakenly thought of as a property of a measurement system (e.g., the SF-12 survey.) This common misunderstanding does not make much trouble in settings where the things that affect reliability are often held constant. But beyond items and their measurement

properties, reliability is a function of physician-to-physician variation and within-physician sample size. Both of these important components of reliability will vary from setting to setting. The same measurement system will have different reliabilities when applied nationally or applied to a single metropolitan region. Even within region, different reliabilities may be obtained from combined datasets versus calculations for a single managed care organization.

Higher reliability increases the likelihood that you will assign a physician to the "right" group. Sample size, while often used as a proxy for reliability, may be insufficient. Therefore, simple minimum sample size cut points (e.g., a rule that a physician's profile must include at least 30 events) may not solve this problem. 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 set of indicators that we do not have available. 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 available sample.

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 passed or failed 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.

What Is a "Sufficient" Level of Reliability?

Different levels of reliability have been advocated for different purposes. Psychometricians use a rule of thumb of 90 percent for drawing conclusions about individuals.[13] Lower levels (70-80 percent) are considered acceptable for drawing conclusions about groups. The choice of minimum reliability level raises questions about the trade-off between feasibility and scientific soundness.

Figure 1 shows the relationship between the physicians' true score distributions and the observed distributions as reliability changes from 50 percent to 70 percent to 90 percent. Each panel of Figure 1 shows the true distribution of the physicians' scores. The solid bell-shaped curve represents the provider-to-provider variance, and the dashed bell-shaped curves show the physician-specific error distribution for ten random physicians. At a reliability of 50 percent, it is difficult to detect differences between physicians. At 70 percent reliability, we can start to see differences between some physicians and the mean. At a reliability of 90 percent, we can start to see significant differences between pairs of physicians.

---

[13] Hays RD, Revicki D. Reliability and validity (including responsiveness). In: Fayers P, Hays R, eds. *Assessing Quality of Life In Clinical Trials*. New York: Oxford University Press; 2005.
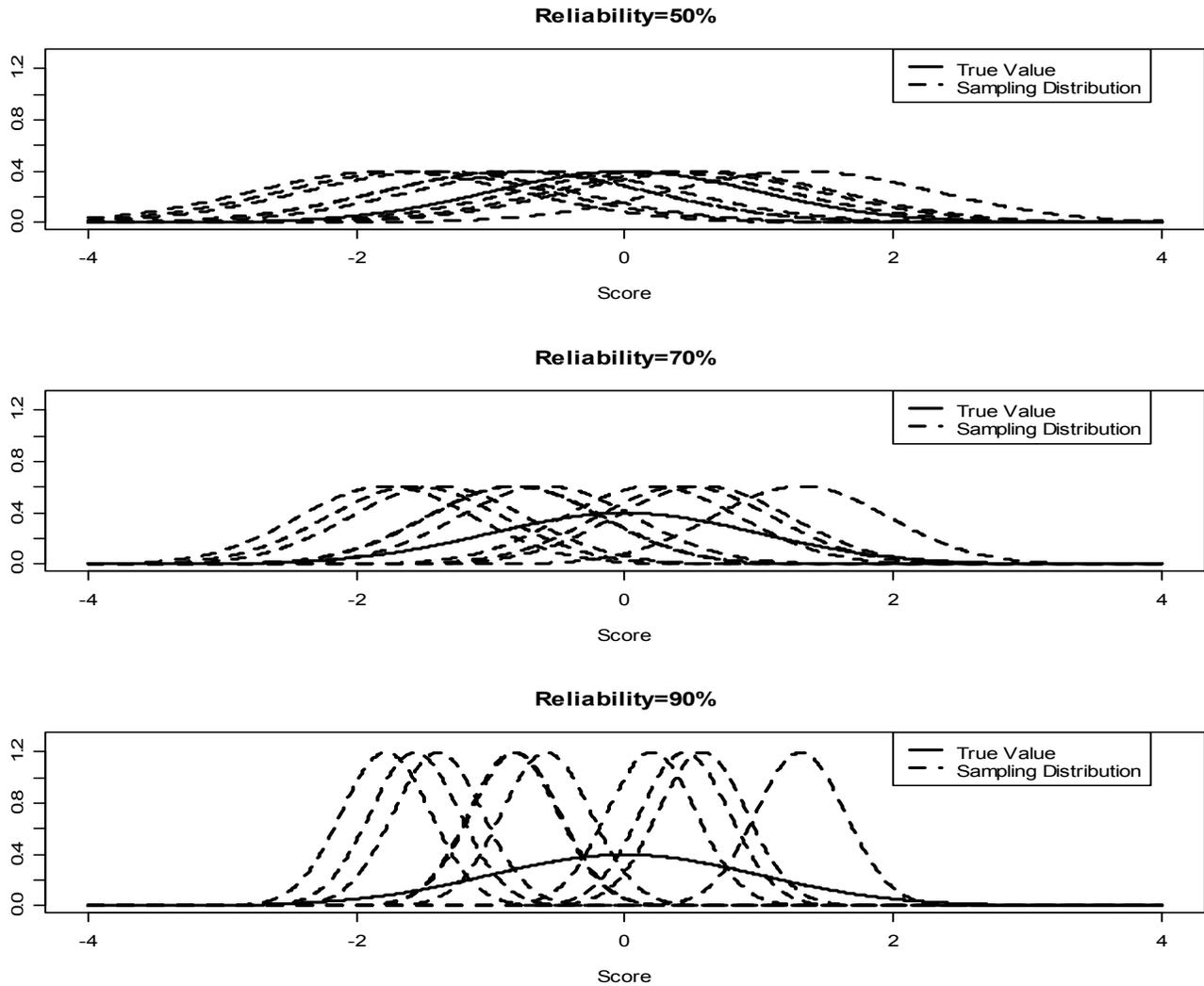
Figure 1: How Reliability Is Related to the Physicians' Score Distribution

Beyond the simple constant variance reliability illustrated in Figure 1, this tutorial will address

two features of physician profiling data that are not commonly found in other reliability

calculations. These are lack of balance and heterogeneity. Lack of balance is a consequence of

the widely varying sample sizes from physician to physician. In a typical survey measure (e.g.,

SF-12) everyone answers the same questions, each question only once, so the number of items is

the same for each respondent. Heterogeneity is the variation in measurement error that is due to the change in variance of the binomial distribution as the probability changes.

Why Is Reliability Important?

For analysts not focused on the technical intricacies of physician profiling, the sudden emergence of reliability as an important property may come as a surprise. After years of focus on sample size, power, and confidence intervals, reliability has recently come into focus as a key issue. The reason this has occurred is the strong relationship between reliability and the key policy issues in the physician profiling debate. This is a logical consequence of the way that stakeholders are using physician profiling information.

Fundamentally, reliability is the measure of whether you can tell one physician, from another. There has recently been more interest in public reporting and "pay for performance." The focus has been on putting physicians into categories. High-performance networks and three-star public reporting systems have emerged. A three-star system might give one star to physicians in the bottom quartile, two stars to physicians in the middle 50 percent of the data, and three stars to physicians in the top quartile.

Given the discrete categories that are being used, the big question becomes the probability of misclassification. If the categories are based on relative comparisons, reliability tells you most of what you need to know about misclassification in these systems. There are several types of misclassification that depend on the number of classification categories. For a simple high-performance network system that flags a subset of the physicians as high-performing, there are two types of errors: (1) flagging a lower-performance physician as high-performance and (2)

failing to flag a high-performance physician as high-performance. Three-star systems have several possible types of misclassifications.

All of these misclassification probabilities are reduced by increasing reliability. As an illustration, we will look at the reliability-misclassification relationship for a high-performance network system that assigns physicians to the high-performance category with one of two rules: a cut-point rule where the observed score is at or above the 75th percentile or a statistical t-test that rejects the hypothesis that the physician is average (one-sided on the positive side). For simplicity, we will explore groups of physicians with the same reliability but examine different reliability levels.

Figure 2 shows the probability of being labeled as a high-performance physician using a cut point at the 75th percentile at various values of the true score distribution for reliabilities of 50 percent, 70 percent, and 90 percent. The gray bell shaped curve represents the true score for the physicians. The true score is expressed in units of standard deviations from the mean. The average physician here has a score of zero. If a physician is far enough into the right tail of the distribution, he or she will be labeled as high-performing at any level of reliability. But note that even average physicians can be labeled as high-performing if the reliability is poor.
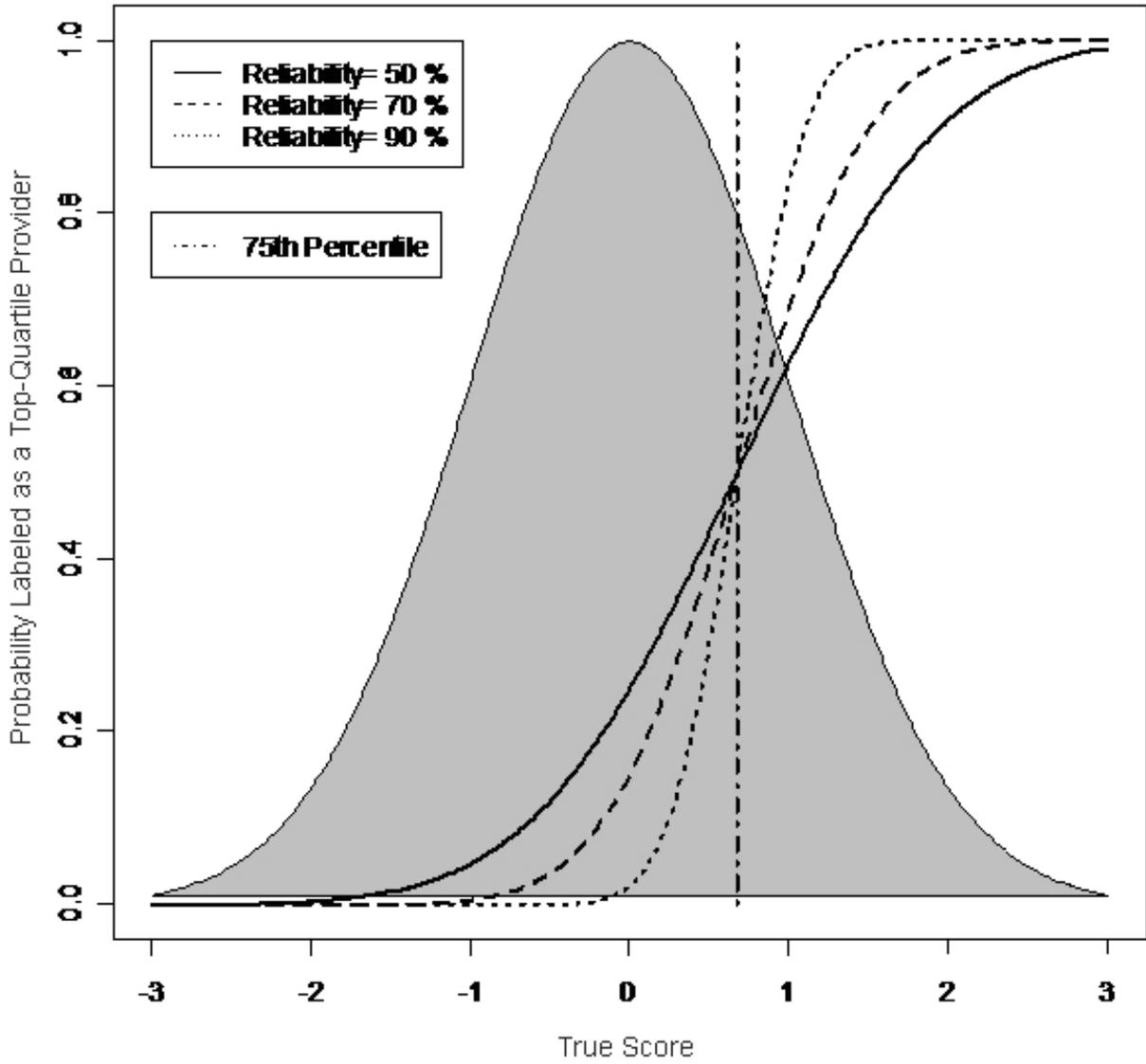
Figure 2. Physician Labeled as a Top-Quartile Provider

Table 1 summarizes the misclassification probabilities for various levels of reliability. Note that

even at a reliability of 90 percent a substantial number of misclassifications can occur.

Table 1: Misclassification Probabilities for a 75th Percentile Cut Point at Various Levels of Reliability

Reliability %	True score above the cut point labeled as below (misclassification)	True score below the cut point labeled as above (misclassification)
5	64.3	21.5
10	60.9	20.1
15	57.2	18.9
20	54.4	18.2
25	51.8	17.2
30	49.2	16.2
35	46.6	15.4
40	44.3	14.8
45	41.9	13.9
50	39	13.1
55	36.5	12.4
60	34.4	11.4
65	31.8	10.6
70	28.9	9.8
75	26.5	8.7
80	23.5	7.7
85	19.6	6.7
90	16.3	5.5
95	11.5	3.8
100	0	0

Figure 3 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. This is the reason statistical tests are often thought of as more rigorous. They do not flag a physician as different from the mean unless there is strong evidence to do so.
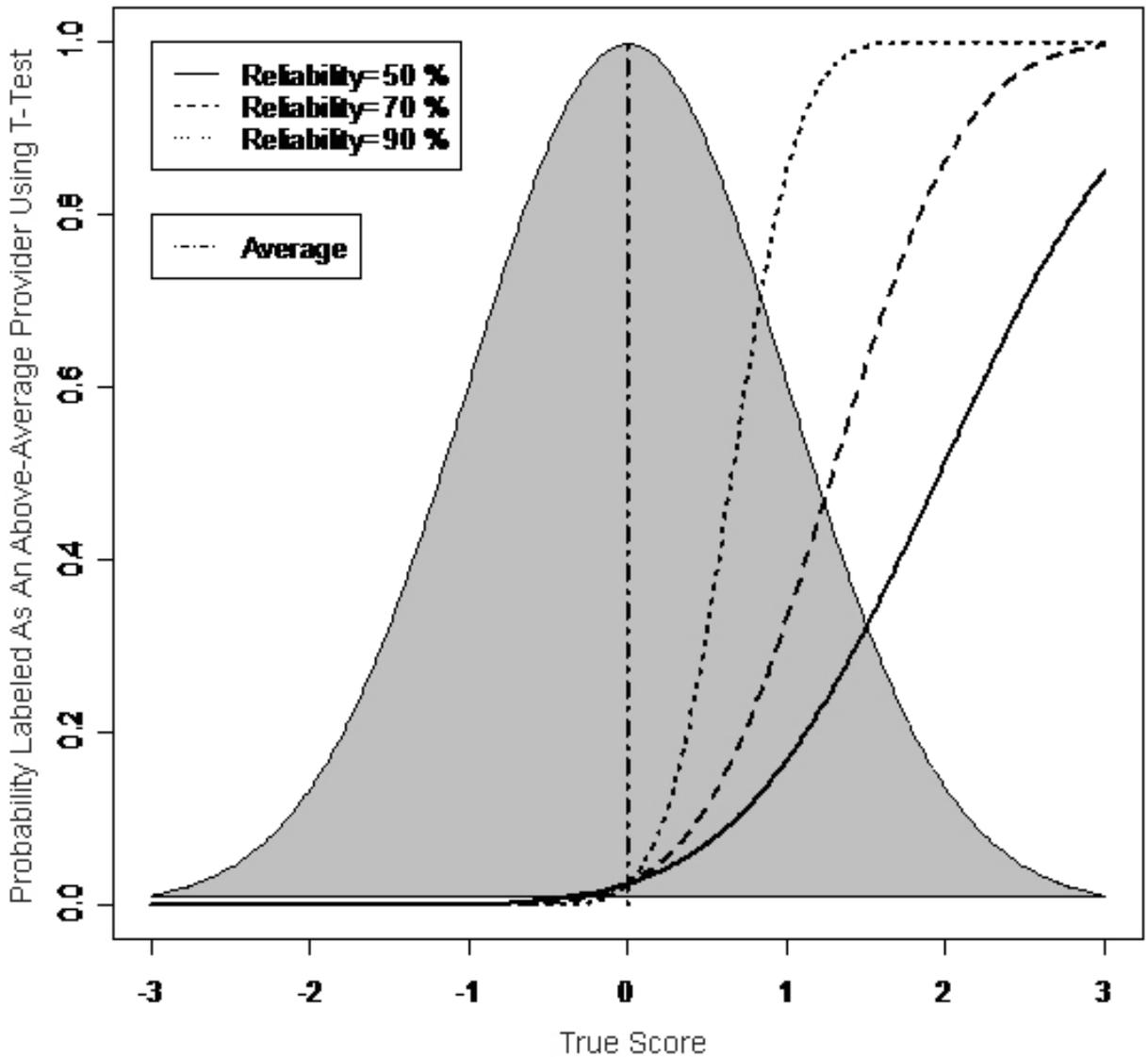
Figure 3. Physician Labeled as an Above-Average Provider by a t-test

Table 2 summarizes the correct and misclassification probabilities for various levels of reliability

for statistical tests versus the mean or 50th percentile. A literal reading of misclassification for

this rule would suggest that being above or below zero is the cut point for

correct/misclassification. But this may be too literal. For this reason, I present correct

classification rates separately for the third and fourth quartiles. From a policy perspective, failing

to label a third-quartile physician as high-performing may be a less costly error than failing to label a fourth quartile physician as high-performing. 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.

Table 2: Correct and Misclassification Probabilities for a Statistical Test at Various Levels of Reliability

Reliability %	True score in second quartile labeled as above average (% misclassified)	True score in third quartile labeled as above average (% correctly classified)	True score in fourth quartile labeled as above average (% correctly classified)
5	1.3	1.9	3.5
10	1.5	2.6	5.7
15	1.5	3	7.5
20	1.5	3.3	9.4
25	1.5	3.5	11.4
30	1.4	3.9	13.8
35	1.3	4.2	16.4
40	1.3	4.5	19.1
45	1.2	4.9	22.4
50	1.2	5.3	26.1
55	1.1	6	30.3
60	1.1	6.3	35.4
65	1	6.9	40.8
70	0.9	7.9	48
75	0.8	9.2	55.3
80	0.7	11.2	64.3
85	0.6	14.2	74.9
90	0.5	19.8	86.8
95	0.3	34.4	97.7
100	0	99.6	100

These are only two of many possible rules for assigning physicians to categories for payment or public reporting purposes. It is straightforward to calculate correct and misclassification probabilities for most systems based on relative scores. But in all the cases we have observed, reliability is the essential quantity for estimating the correct and misclassification probabilities.

How Does Reliability Relate to Validity and Precision?

Validity is the most important property of a measurement system. In nontechnical terms, validity is whether the measure actually measures what it claims to measure. If the answer is yes, the measure is valid. This may be an important question for physician profiling. For example, what if a measure of quality of care is dominated by patient adherence to treatment rather than by physician actions? Labeling the measure as quality of care measure does not necessarily make it so.

Although this is not an exhaustive list, here are several important determinants of validity of physician performance measures:

- Is the measure fully controllable by the physician?
- Is the measure properly adjusted for variation in the case-mix of patients among physicians?
- Is the measure partially controlled by some other level of the system?
- Is the measure correlated with other established quality measures?

Reliability assumes validity as a necessary precondition. Although the reliability calculations can still be performed for measures that are not valid, subsequent interpretation is problematic. Consider intermediate outcome measures that depend on patient characteristics that are not available for analysis. A high reliability for this type of measure would leave unresolved whether the physician-to-physician variance was due to true physician effects or due to the clustering of patients with different patient characteristics within physician.

Precision is a measure of how repeatable measurements are. Common measures of precision include standard errors and confidence intervals. For the binomial cases in this report, the

precision is estimated by the variance of the binomial mean, $\dfrac{\hat{p}*(1-\hat{p})}{n}$. Precision is one

component of reliability represented by the measurement error variance. What precision does not

consider is the physician-to-physician variation.

A common question in this area is: "Why isn't it good enough to make sure the confidence

intervals (or standard errors) are small?" Small confidence intervals are indeed desirable and can

be adequate to make sure the scores are above or below a fixed cut point with a certain

misclassification rate. But reliability is about relative comparisons. Small confidence intervals

cannot distinguish one physician from another if reliability is small. Consider a "topped out"

measure in which the population pass rate is 95 percent. Even large sample sizes and their

corresponding narrow confidence intervals can't distinguish one physician from another if the

range of physician scores is narrow enough. Large physician-to-physician variation makes it

possible to distinguish physicians even if confidence intervals are larger.

How Do You Calculate Reliability?

Reliability is usually calculated by fitting the appropriate hierarchical model. Most of the time

this will be a hierarchical linear model assuming normal errors and normal mixing distributions.

These are typically estimated with SAS's PROC MIXED or with one of several specialized

computer packages (e.g., HLM or MLM.) Usually these models are linear and use normal

distributional assumptions. We could use the traditional extensions of these models to

hierarchical logit or probit models for our binary outcomes. The approach we will take here is

somewhat different. We will introduce the estimation of reliability by using the beta-binomial

model. Although less common than normal HLMs, the beta-binomial model is a more natural fit

for estimating the reliability of simple pass/fail rate measures (e.g., HEDIS measures). There are also computational advantages to using the beta-binomial model. The normal HLM will be introduced later when we consider the estimation of reliability for composites and continuous measures.

The approach assumes that each physician has a true pass rate, p. This true pass rate varies from physician to physician presumably as a consequence of variation in physicians' practice styles. The observed pass rate $\hat{p}$ will vary in any given time period because the number of events is small and subject to random variation around the true rate.

The beta-binomial model is based on the beta distribution for the "true" physician scores. The beta distribution is a very flexible distribution on the interval from 0 to 1. This distribution can have any mean in the interval and can be skewed left or right or even U-shaped. It is the most common distribution for probabilities on the 0-1 interval. The beta-binomial model assumes the physician's score is a binomial random variable conditional on the physician's true value that comes from the beta distribution.

The beta distribution is usually defined by two parameters, alpha and beta. This is a historical parameterization that is somewhat nonintuitive in this application. Most users do not need to develop an intuition for how to interpret alpha and beta. Alpha and beta can be thought of as intermediate calculations to get to the needed variance estimates. The translation from alpha and beta to the mean and variance of the distribution is useful for our purposes:

$$\mu = \frac{\alpha}{(\alpha + \beta)}$$

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

Figure 4 shows what the beta distribution looks like for various combinations of alpha and beta. The beta distribution can be symmetric and centered on 0.5 to look like the center of a normal distribution. It can be skewed left or right to represent populations of physicians with low or high averages. It can even be U-shaped to represent populations of physicians that have nearly separate low and high subpopulations.
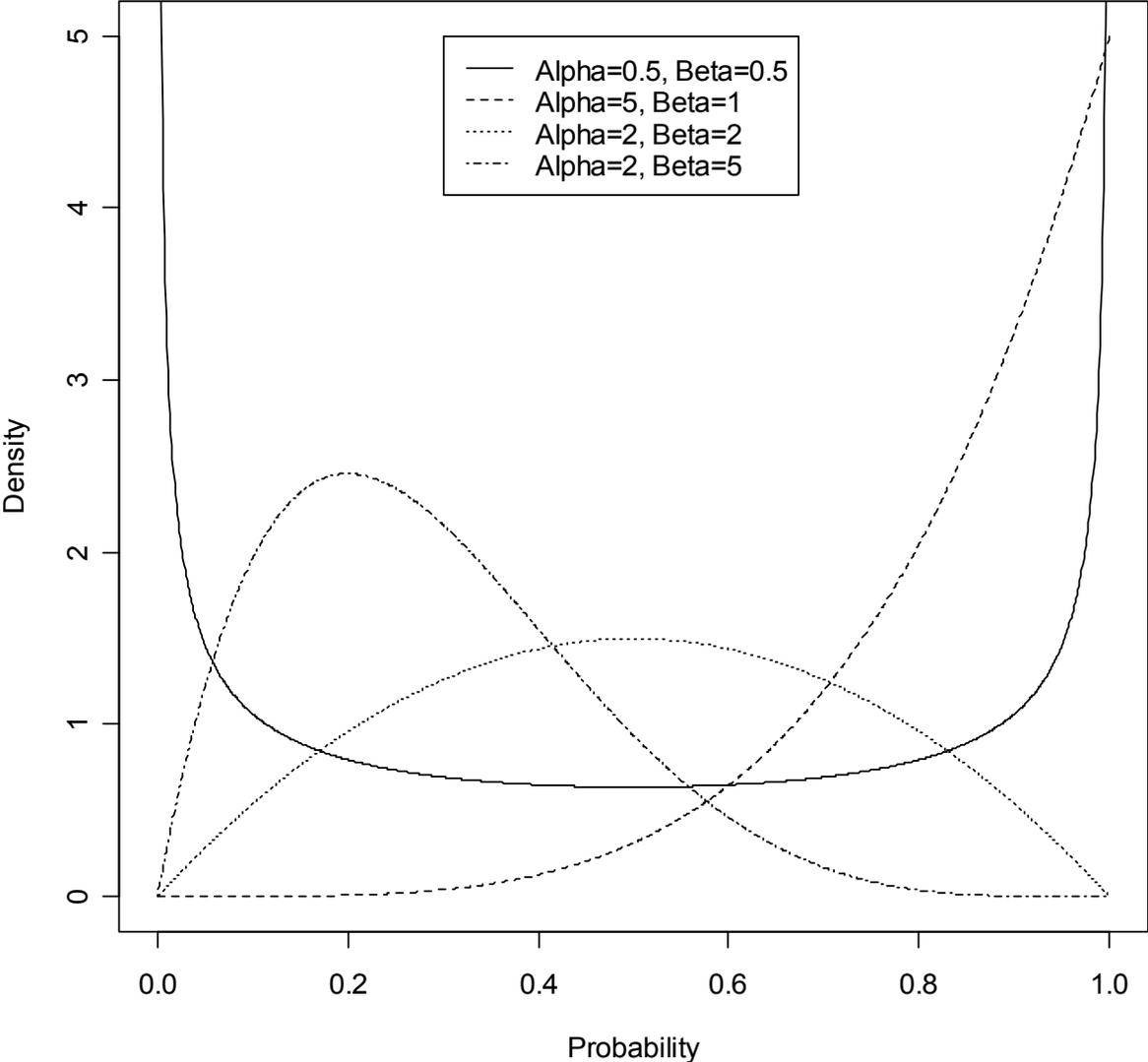

Figure 4. Beta Distributions

Calculating the reliability from the beta-binomial is straightforward if you know alpha and beta. You need to calculate the physician variance:

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

Then you need the usual binomial variance for the error where p is the provider specific probability of passing the indicator:

$$\sigma^2_{error} = \frac{\hat{p}(1-\hat{p})}{n}.$$

Where $\hat{p}$ is the observed pass rate for the provider. Then the standard formula for reliability is applied.

As an example to help illustrate the estimation and software issues, I will build a simulated dataset to explore. The simulated dataset is called beta55. The "55" is to remind us that both alpha and beta are set to 5 for the beta distribution. Figure 5 shows what the true distribution looks like for beta55.

Here are the steps in generating beta55:

- One thousand simulated physicians are drawn from a beta (5,5). This generates a "true" performance rate for the physician.
- For each physician, ten binary (pass/fail) eligibility events are simulated from their known rate.

These two steps are the beta and the binomial parts of the beta-binomial model.
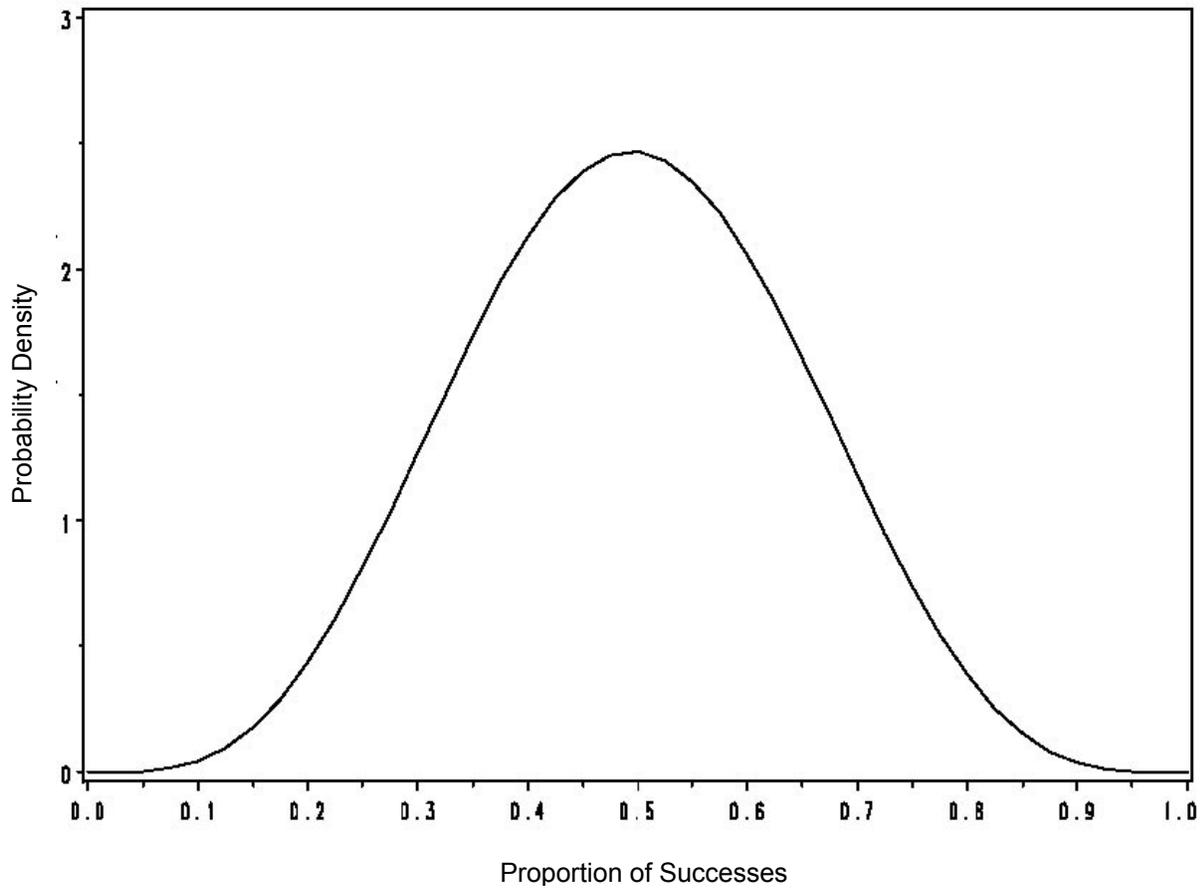
Figure 5. Beta Distribution for the Binomial Proportion

Using the earlier formulas for the beta distribution, we can calculate the mean and variance for

beta55:

Mean:

$$\mu = \frac{5}{(5+5)} = \frac{1}{2};$$

variance:

$$\sigma^2_{provider-to-provider} = \frac{5*5}{(5+5+1)(5+5)^2} = \frac{1}{44} = 0.023$$

.

Although the variance is the most useful quantity for reliability calculations, most people find the

standard deviation, the square root of the variance, a more intuitive quantity. The standard

deviation for beta55 is 0.151. Although the normal-theory rule of thumb that approximately 95 percent of the probability lies within +/-2 standard deviations of the mean does not generally apply to beta distributions, it isn't a bad approximation for beta55.

We can now calculate the theoretical reliability for physicians from the beta55 dataset. We know the physician-to-physician variance, 0.023.

The binomial error is:

$$\sigma^2_{binomial} = \frac{p(1-p)}{n}$$

where p is the passing rate for a physician.

And the reliability is

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

So the reliability depends on p! This is different from the usual scale development situation.

There is no uniform answer to the question: "What is the reliability of the scores?"

Some examples:

Physician-to-physician variance	n	p	Reliability
0.023	10	0.5	0.48
0.023	10	0.2	0.59
0.023	10	0.8	0.59
0.023	10	0.9	0.72

The 1,000 physicians in the beta55 dataset have a mean reliability of 0.51. Figure 6 presents a histogram of the 1,000 physicians' reliabilities. The bulk of the physicians have rates near 50 percent where the binomial variance term is largest and the reliability lowest. But for physicians with rates nearer to 0 or 1, the reliability can be substantially better.

0.023/(0.023 + Beta55 * (1 – Beta55)/10)
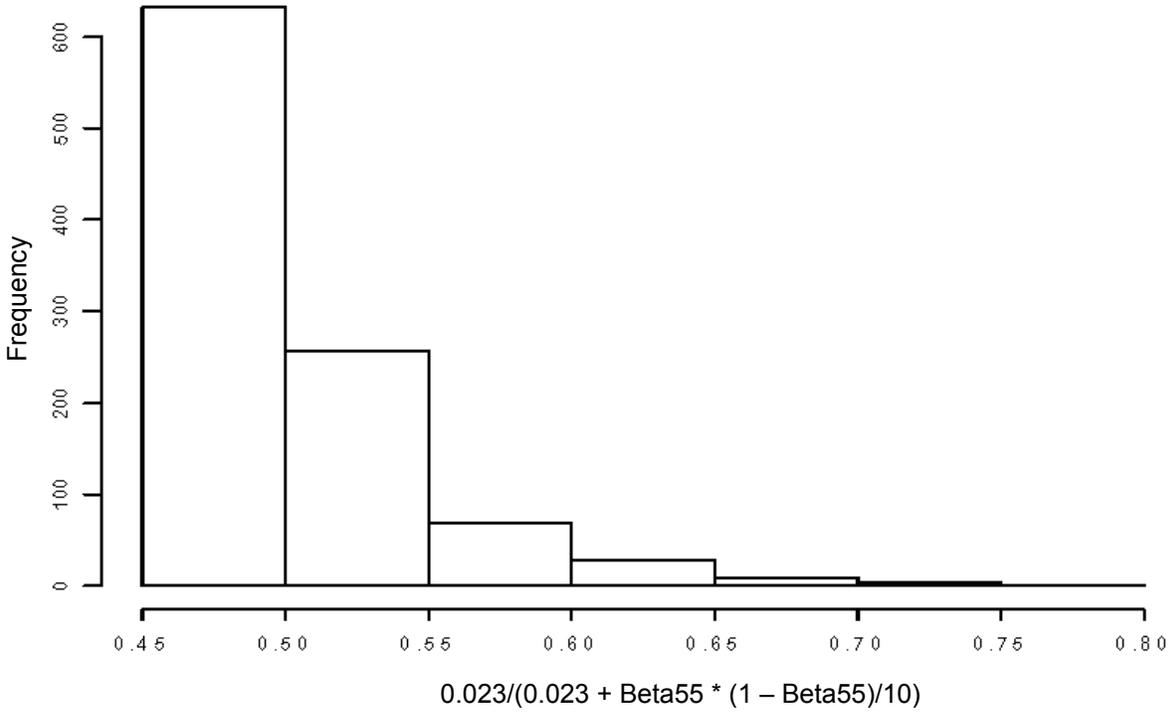
Figure 6. Histogram of the Simulated Values

Since these are simulated data, we know the true pass rates for the physicians. If you know the

true value, you can do the hypothetical regression mentioned earlier. Recall that

$$measurement = \beta_0 + \beta_1(truevalue) + \varepsilon.$$

SAS code:

```
proc reg data=rel.truth;
model measurement=truth;
```

Results:

```
Root MSE             0.15337    R-Square      0.4724
Dependent Mean        0.5099    Adj R-Sq      0.4718
Coeff Var            30.0783
```

```
Variable     DF   Estimate   Error    t Value   Pr > |t|
Intercept    1    0.02073    0.01707  1.21      0.2247
True Value   1    0.97337    0.03256  29.89     <.0001
```

24

Note that the R-squared from this regression, 0.4724, is similar to the theoretical mean reliability of 0.51. The difference is due to the homoskedastic normal error assumptions of the regression model. The point is that they are quite similar despite the assumption violations.

Calculating the Reliability

The previous calculations use our knowledge of the true parameters of the beta distribution. 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, 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 I 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

Web site: www.qistatistics.co.uk

As with all free software, caveat emptor! I have tested this by simulating datasets 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)

Here is the relevant portion of the macro output (much of the output has been deleted here):

BETABIN Macro: Beta-Binomial Model Parameters

Parameter	Estimate	Standard Error	t value	Pr > | t |	Alpha	Lower	Upper
mu	0.5112	0.006891	74.18	<.0001	0.05	0.4977	0.5247
alpha	4.5865	0.4096	11.20	<.0001	0.05	3.7837	5.3893
beta	4.3862	0.3908	11.22	<.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 beta55.

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. Here is how they are used:

Remember 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:

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

Note that the estimated physician-to-physician variance is nearly identical to the known variance

used to simulate beta55.

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.

The Reliability of Composite Scores

Calculating the reliability of composite scores developed from binary measures presents different challenges from the simple binary score case. Composite scores are built from several different binary scores averaged together, possibly with weights. For example, the pass rates of several diabetes measures could be averaged to obtain a diabetes composite score. In general, it is not possible to use an exact distributional model like the beta-binomial to do the modeling. Instead, approximate normal theory models are used. In some ways, these models are simpler and more familiar than the beta-binomial model. However, they can be computationally more challenging, and some assessment of the appropriateness of the normal approximations is required. Alternatively, extensions of the logistic or probit hierarchical models could be used, but these more advanced methods are beyond the scope of this tutorial.

In this section, three issues will be addressed: (1) a brief review of traditional scale methods useful for developing composites, (2) an introduction to the normal-normal HLM, and (3) reliability calculation for composites.

Traditional scale methods include factor analysis and related methods. These are commonly used in developing scales from several items for a given respondent. The analog is scale development from person-level questionnaire data. Consider the SF-12 questionnaire to estimate person-level physical and mental functional status scores. The roughly equivalent problem in physician profiling is to take several measures on a physician and develop a score that combines those measures.

It is possible to treat physician data in the same way that person-level questionnaire data are analyzed. Standard factor analysis computer packages could be run, and the factor scores could

be used for physician profiling. Several features of most physician profiling data can make the traditional factor methods problematic. First, physician profiling data can have wildly varying sample sizes for the measures that go into a physician's score. For example, one physician may have 50 patients who are eligible for diabetes measures and another physician may have only a few. This produces a wide range of variances from physician to physician that is not readily handled in standard factor analysis computer packages. Second, there can be a lack of or only small overlap between measures. For example, there may be no physician who has eligible patients for both well-child visit measures and geriatric measures. Most factor analysis programs work with the correlations between scores to develop factors. These packages cannot cope with some common data patterns in profiling applications. Again, extensions to more elaborate latent class models like the Rasch or item-response theory models are possible but beyond the scope of this tutorial.

Factor analysis may still be a workable solution for some profiling applications. It has the advantage of standard tools as well as a built-in methodology for determining the number of latent factors and testing for a one-dimensional underlying construct.

In the remainder of this section, I will focus on the cases where the standard factor analysis methodology is not necessarily appropriate. I will take as given a set of rules for developing a physician score and then work through how to calculate reliability using a normal-normal HLM to calculate the physician-to-physician variance. Note that we could have used a more elaborate methodology extending factor analysis to binary outcomes, but we will focus on simpler, more familiar approaches.

Consider building a composite of HEDIS measures at the physician level. Each physician will have different numbers of eligibility events for each of several measures. Some measures may not have any eligibility events for some physicians. Several decisions could be made in developing a composite:

- Omit a measure for a physician if a minimum number of eligibility events do not occur.

- Consider weights of various sorts: equal weights, importance weights, market basket weights.

- Use complex scoring rubrics (e.g., no credit for screening measures unless action is taken afterwards) or possibly all-or-nothing scoring.

In this section, I will not focus on the optimality of the way the composite is constructed. The composite can be ad hoc or driven by nonstatistical considerations. The focus here is how to calculate reliability given a scoring method. A good example of this process can be found in Scholle et al.[14]

The primary tool for calculating reliability from composites is the normal-normal model. This is the simplest hierarchical linear model.[15] I will assume that the true physician score is distributed normally with some overall mean and variance. The physician's observed score is distributed normally with the true score as the mean and measurement noise around the true mean. These models can be fit with mixed model or specialty software. I will focus on estimation with SAS's PROC MIXED.

[14] Scholle SH, Roski J, Adams J, Dunn D, Kerr EA, Dugan DP, Jensen R. Benchmarking physician performance: Reliability of individual and composite measures. *Am J Manage Care.* 2008;14(12):833-838.
[15] Raudenbush and Bryk, 2002.

Two things are required for each physician as inputs to the estimation process: a score and an error variance for that score. It is the error variance calculation that may be a challenge for some scoring rules. Methods of calculating error variances are typically obtained by thinking of the sampling error of a single physician's score without consideration of the between variance issues. For the weighted combination of several binary measures, the sampling variance can be calculated by multiplying the variances of individual measures by the squared weights and summing the components. A statistician may be consulted to develop an error variance formula for more complicated composites, perhaps using the delta method. It may be possible to implement a bootstrap calculation to determine the error variance computationally.

Once the scores and their error variances are obtained for all the physicians, the HLM can be fit to estimate the physician-to-physician variance. The difference in the standard errors, typically a result of different sample sizes by measure, requires some special software considerations. The chosen software must be able to handle observations with different variances. Coincidently, this is the same computational problem that meta-analysts have when combining information from studies with different variances.

As an example, we sketch how to do the calculation in SAS's PROC MIXED. The dataset:

ID	Score	Errorvariance
1	0.91	0.0039
2	0.33	0.0105
3	0.56	0.0117

...

The SAS code:

```
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 dataset 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 provide specific

reliabilities.

There is a growing interest in calculating the reliability of physician cost profiles. This is a

challenging application of reliability because of the skewed distribution of cost data. The

calculation of reliability for cost data parallels the calculation for composites presented here.

Scores and their error variances can be analyzed using PROC MIXED in the same way PROC

MIXED is used for composites.

Conclusions

Reliability is a topic of increasing importance in profiling applications. I hope this tutorial will

reduce some of the confusion about what reliability is and what it means. The simple and robust

computational methods presented here will make it easier for those designing performance

measurement systems to routinely examine the reliability of their measures.

This tutorial has underscored that reliability is not just a property of a measure set but also

depends what population is used to estimate the reliability. Whether a set of measures is useful

for profiling providers depends on how different the providers are from one another. Measures

that may be useful in one group of providers may not be useful in another group with little

provider-to-provider variation. Similarly, as the providers under study increase their performance

the reliability may decrease if the provider-to-provider variance decreases over time. This is

especially true as measures hit the upper limits of their ranges.

There are certainly unanswered questions regarding how to calculate reliability for more

complicated composites. Conceptual challenges remain, especially when there are multiple levels

in the system that may influence measure scores.