

Supplementary Appendix

This appendix has been provided by the authors to give readers additional information about their work.

Supplement to: Adams JL, Mehrotra A, Thomas JW, McGlynn EA. Physician cost profiling — reliability and risk of misclassification. *N Engl J Med* 2010;362:1014-21.

TECHNICAL
R E P O R T



Physician Cost Profiling— Reliability and Risk of Misclassification

Detailed Methodology and Sensitivity
Analyses

TECHNICAL APPENDIX

John L. Adams, Ateev Mehrotra, J. William Thomas,
Elizabeth A. McGlynn

Sponsored by the U.S. Department of Labor

This work was sponsored by the U.S. Department of Labor. The research was conducted in RAND Health, a division of the RAND Corporation.

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. RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

RAND[®] is a registered trademark.

© Copyright 2010 Massachusetts Medical Society

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 provides more detail about the methods outlined in the article “Physician Cost Profiling—Reliability and Risk of Misclassification,” published in the *New England Journal of Medicine* (Adams et al., 2010). In an effort to rein in rising health care costs, purchasers are experimenting with a number of policy interventions that depend on physician cost profiling. For example, some health plans are using cost profiles to identify and then exclude high-cost physicians from their networks. In this report, we describe in more depth our data sources, how we assessed the reliability of physician cost profiling tools, how we calculated the misclassification of physician performance, and the results of our sensitivity analyses.

This report will be of interest to health plan representatives, policymakers, and researchers with an interest in more details on cost profiling in health care.

This work was funded by the U.S. Department of Labor. 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.

Contents

Preface	iii
Figures	vii
Tables	ix
Summary	xi
Abbreviations	xiii
CHAPTER ONE	
Introduction	1
CHAPTER TWO	
Data Sources and Criteria for Inclusion of Patients and Physicians	3
Source of Claims Data	3
Selecting Patients for Inclusion in the Study Population	3
Physician Database and Selecting Physicians for Inclusion in the Study	3
CHAPTER THREE	
Constructing Cost Profiles	7
Creating Episodes of Care and Risk Adjustment Methodology	7
Computing the Costs of Episodes	8
Assigning Episodes to Physicians	9
Constructing a Summary Cost Profile Score	10
CHAPTER FOUR	
Calculating the Reliability of Cost Profiles	13
Overview of the Rationale for Examining Reliability	13
Detailed Description of the Method for Calculating Reliability	15
CHAPTER FIVE	
The Relationship Between Reliability and Misclassification	23
CHAPTER SIX	
Sensitivity Analyses	29
References	35

Figures

2.1.	Steps for Selecting Patients for Inclusion in the Study	4
2.2.	Constructing the Physician Sample	5
3.1.	Comparison of the Mix of Assigned Versus Unassigned Episodes, by Major Disease Category	10
3.2.	Score Distribution of Physicians, by Specialty	11
4.1.	Relationship Between Number of Assigned Episodes and Reliability of Physicians' Cost Profile Scores.....	14
5.1.	Relationship Between Reliability and Physicians' Score Distribution.....	24
5.2.	Probability of Being Labeled a Lower-Cost Physician Based on True Score and Reliability of Cost Profile Score.....	25
5.3.	Probability of Being Labeled Lower-Cost Based on True Score and Reliability of Physician's Cost Profile.....	27
6.1.	Three Illustrative Classification Systems	33

Tables

2.1.	Characteristics of Massachusetts Physicians Compared to U.S. Physicians	6
4.1.	Estimates of the Physician-to-Physician Variance Components	18
4.2.	Relative Ranks for the Components of Reliability, by Specialty	19
5.1.	Misclassification Probabilities for a 25th-Percentile Cut Point at Various Levels of Reliability	26
5.2.	Rates of Misclassification, by Specialty, for a Two-Tiered Network.....	28
6.1.	Median Reliability, by Specialty, for Each Sensitivity Analysis	30
6.2.	Overall Misclassification Rates for Three Systems.....	33

Summary

This technical report accompanies the article “Physician Cost Profiling—Reliability and Risk of Misclassification,” published in the *New England Journal of Medicine* (Adams et al., 2010). In this report, we provide more detail about the methods used to assess the reliability of physician cost profiling tools and the potential for misclassification of physician performance. We also present the results of our sensitivity analyses.

Purchasers are experimenting with a variety of approaches to control health care costs, including limiting network contracts to lower-cost physicians and offering patients differential copayments to encourage them to visit “high-performance” (i.e., higher-quality, lower-cost) physicians. These approaches require a method for analyzing physicians’ costs and a classification system for determining which physicians have lower relative costs. To date, many aspects of the scientific soundness of these methods have not been evaluated.

One important measure of scientific soundness is reliability. 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 that of 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.

The overall finding of the research is that the majority of physicians in our data sample did not have cost profiles that met common thresholds of reliability and that the reliability of cost profiles varied greatly by specialty. In an illustrative two-tiered insurance product, a large fraction of physicians were misclassified as low-cost when they were actually not, or vice versa. Our findings raise concerns about the use of cost profiling tools, because consumers, physicians, and purchasers are at risk of being misled by the results.

Abbreviations

DRG	diagnosis-related group
ERG	episode risk group
ETG	episode treatment group
HLM	hierarchical linear model
MHQP	Massachusetts Health Quality Partners
PPO	preferred provider organization

Introduction

Purchasers are experimenting with a variety of approaches to control costs (Sandy, Rattray, and Thomas, 2008). Recently, their focus has shifted to physicians because they write the orders that drive spending (Milstein and Lee, 2007). Prior research suggests that if physicians adopted less intensive practice patterns, health care spending would decrease (Sirovich et al., 2008). Health plans are limiting the number of physicians who receive “in-network” contracts, offering patients differential copayments to encourage them to visit “high-performance” (i.e., higher-quality, lower-cost) physicians, paying bonuses to physicians whose resource use patterns are lower than average (Sorbero et al., 2006), and publicly reporting the relative costliness of physicians (Draper, Liebhaber, and Ginsburg, 2007). Legislation under consideration in the 111th Congress calls for the use of cost profiling in value-based purchasing strategies.

All of these applications require a method for analyzing physicians’ costs and a classification system for determining which physicians have lower relative costs. Quality and other performance measures are traditionally evaluated for scientific soundness by assessing validity and reliability (Institute of Medicine, 2006). The validity of episode grouping tools as a method for constructing clinically homogeneous cost groups is widely accepted (Centers for Medicare and Medicaid Services, undated; National Quality Forum, 2009). A second validity question is whether the method of assigning episodes to physicians and creating summary scores accurately represents physicians’ economic performance. We have evaluated the convergent validity of different methods of assigning episodes to physicians (Mehrotra et al., in press). The reliability of physician cost profiling, however, has not previously been addressed.

Reliability of cost profiles is determined by three factors: the number of observations (i.e., episodes), the variation between physicians in their use of resources to manage similar episodes, and random variation in the scores. For cost profiles, reliability is measured at the individual physician level because the factors that are used to estimate reliability are different for each physician. For any specific application of cost profiling, we can estimate the likelihood that a physician’s performance will be inaccurately classified based on the reliability of the physician’s profile score.

We evaluated the reliability of current methods of physician cost profiling and what those levels of reliability suggest about the risk that physicians’ performance will be misclassified. We conducted the analysis separately by specialty because patterns of practice differ by specialty, and most applications, such as high-performance networks, have been implemented by specialty (Brennan et al., 2008; Bridges to Excellence and Leapfrog Group, 2004).

In this report, we provide more detail about the methods described in Adams et al. (2010). We focus on the methods used to assess the reliability of physician cost profiling tools and the potential for misclassification of physician performance. We also present the results of our sen-

sitivity analyses. Some material that appears in Adams et al. (2010) is also reported here, so this report is a self-contained document.

In Chapter Two, we describe the Massachusetts health plan claims data and provider databases used in the study. We also provide details on the inclusion criteria for patients and physicians. In Chapter Three, we explain how we constructed the physician cost profiles. For the purposes of this research, we made choices that reflect what health plans commonly do in constructing such profiles. In Chapter Four, we provide some background on reliability. Reliability is determined by three factors: the number of observations (i.e., episodes), the variation between physicians in their use of resources to manage similar episodes, and random variation in the scores. We describe how we calculated reliability and provide an analysis of how the three determinants of reliability combine to produce different results by specialty. In Chapter Five, we describe how we translated our reliability calculations into estimates on the rates of misclassification. For most readers, interpreting reliability is not intuitive. Therefore, we used reliability to estimate a more intuitive concept—the rate at which physicians are misclassified—for an illustrative application of cost profiling. In our illustrative application of cost profiling, 25 percent of physicians who have the lowest-cost profiles are labeled as “lower-cost,” and the remaining 75 percent are labeled as “not lower-cost.”

In Chapter Six, we present the results of our sensitivity analyses. In our sensitivity analyses, we compared Winsorizing versus not Winsorizing outliers, using allowed versus standardized costs, constructing profiles separately for each health plan, using different attribution rules, restricting the evaluation to physicians with at least 30 observations, and evaluating two alternate classification applications, in contrast to the two-category system presented in Adams et al. (2010). Compared to Winsorizing, not using Winsorizing decreased median reliability for 11 specialties and increased median reliability for seven. Using actual costs rather than standardized costs improved median reliability for only three specialties. If the four plans had produced physician cost profiles separately, three of the plans would have had substantially lower reliabilities for all specialties, and the fourth plan would have had higher median reliabilities for 15 of 28 specialties and lower median reliabilities for two. Requiring physicians to have at least 30 episodes to be included in the profiling analyses increased the median reliability for 18 of the 28 specialties but substantially decreased the number of physicians who could be profiled (8,689 versus 12,789). We examined two alternative episode assignment rules, both of which had lower reliabilities. We also evaluated two alternative classification methods, both of which had higher rates of misclassification.

Data Sources and Criteria for Inclusion of Patients and Physicians

Source of Claims Data

We obtained all commercial professional, inpatient, other facility, and pharmaceutical claims for calendar years 2004 and 2005 from four health plans in Massachusetts. Because most physicians contract with multiple health plans, we aggregated data at the physician level across health plans to construct a larger number of observations on each physician than would have been possible with data from a single health plan. The data set included claims from managed care, preferred provider organization (PPO), and indemnity product lines. At the time of the study, the four health plans enrolled more than 80 percent of Massachusetts residents with any type of commercial health insurance.

Selecting Patients for Inclusion in the Study Population

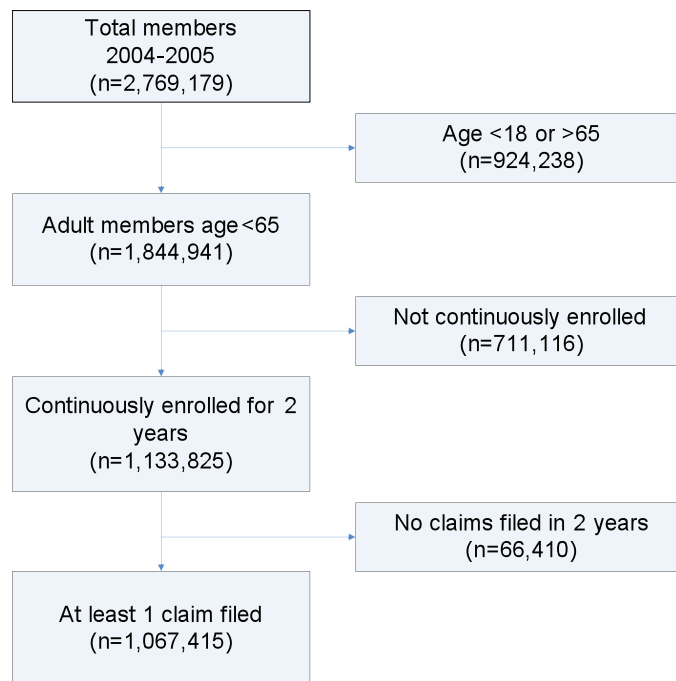
Our analyses used all claims for adult enrollees between the ages of 18 and 65 who were continuously enrolled for the two-year period (2004–2005) and who filed at least one claim. Figure 2.1 shows the steps we took to arrive at the final study sample and the implications of those steps for the final sample size.

We excluded patients over the age of 65 because these patients are eligible for Medicare and the plans could not reliably identify those for whom Medicare was the primary payer. Another aspect of the project involved quality measurement, and our quality measures included only adult measures and required that patients were continuously enrolled. We therefore excluded children (<18) and those who were not continuously enrolled. As a result, we included 58 percent of nonelderly adult enrollees across the four plans in the study. The average age of included patients was 43.3 years, and 53.7 percent were female.

Physician Database and Selecting Physicians for Inclusion in the Study

Massachusetts Health Quality Partners (MHQP) has created a master physician database for the state to facilitate aggregating data across health plans and to enable public reporting of quality information. This database includes all physicians who are listed on the provider files of the largest health plans in Massachusetts. For each physician, MHQP creates a master physician identifier that is linked to the physician identifier used by each of the health plans. MHQP uses unique identifiers (e.g., Massachusetts license number), names, and addresses to create the linkages. MHQP also determines physician specialty using the specialty listed in the

Figure 2.1
Steps for Selecting Patients for Inclusion in the Study



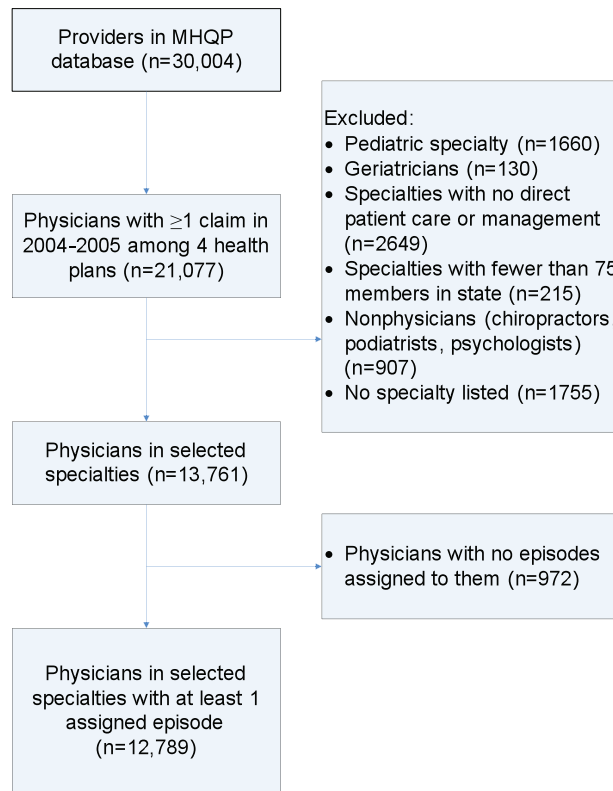
health plan provider files. In the final reconciliation, MHQP uses the physician's Massachusetts Board of Registration file to verify mismatched license numbers and clinical specialties.

In the master physician database created by MHQP, approximately 20 percent of physicians had two specialties. For this study, we assigned each physician to a single specialty using the following logic:

1. In most cases, the combinations were a general specialty (e.g., internal medicine) and a related subspecialty (e.g., cardiology). Because subspecialty costs are typically higher and the main use of the specialty category in our analyses was to compare each physician's cost profile to the average of other physicians in the same specialty, we assigned these physicians to the subspecialty, decreasing the likelihood that a physician would be identified as a high-cost outlier.
2. If one of the two specialties was a non-direct patient care specialty (e.g., pathology and internal medicine), we selected the direct patient care specialty (e.g., internal medicine).
3. In the very rare cases in which there was no clear hierarchy in the two specialties (e.g., general surgery versus pulmonary and critical care), we selected the first listed specialty (the order had been previously assigned at random by MHQP).

The use of this master physician directory allowed us to link a physician's claims across health plans. There are 30,004 providers listed in the database. From this list, we first restricted our sample to the 21,077 providers with a Massachusetts address who had filed at least one claim with any of the four health plans in 2004–2005 (see Figure 2.2). This step excluded physicians who retired, practiced outside the state, left the state, or had temporary licenses, such as medical residents.

Figure 2.2
Constructing the Physician Sample



We then excluded the following providers: (1) pediatricians and geriatricians, to be consistent with our patient exclusion rules; (2) physicians in non–direct patient care specialties only (e.g., pathologists, radiologists, anesthesiologists); (3) physicians who were not assigned a specialty in the database; (4) nonphysicians (e.g., chiropractors, clinical psychologists); (5) physicians in specialties with fewer than 75 total members in the state (to allow an adequate sample for constructing peer-group comparisons); and (6) physicians who were not assigned any episodes. The final sample contained 12,789 physicians, or 93 percent of the 13,761 physicians in the state in the selected specialties who had any claims in the two-year period.

The claims and MHQP databases do not include descriptive information on the physician sample other than specialty. We linked physicians in our database to publicly available data from the Massachusetts Board of Registration, which includes information on gender, board certification status, location of medical school, years since graduation from medical school, and type of medical degree. We were able to link 12,223 physicians, or 95.6 percent of our study population. The characteristics of the Massachusetts physicians compared to those of U.S. physicians overall are shown in Table 2.1. We note that physicians in Massachusetts are more likely to be board certified, to have trained in a U.S. medical school, to have been in practice for a shorter period, and to have an allopathic medical degree.

Table 2.1
Characteristics of Massachusetts Physicians Compared to U.S. Physicians

Characteristic	Massachusetts		U.S. (2005 AMA Masterfile)
	n	%	%
Gender			
Female	3,687	30	27
Male	8,536	70	73
Board certification			
Yes	11,250	92	69
No	973	8	31
Medical school			
Domestic	10,205	84	75
International	2,018	17	25
Years in practice ^a			
<15	4,156	34	35
16–25	3,789	31	24
26–35	2,689	22	18
36–45	1,222	10	12
≥45	367	3	11
Degree ^b			
D.O.	267	2	6
M.D.	11,943	98	94

SOURCE: Massachusetts data are from the MHQP and claims data sets. U.S. physician data are from the 2005 American Medical Association Physician Masterfile.

NOTE: The sample size in the table (12,223) represents 95.6% of physicians in our study sample (12,789), because some physicians could not be matched to the Massachusetts Board of Registration database.

^a Calculated from year of medical school graduation to January 1, 2005 (the midpoint of the time period studied).

^b 13 physicians were missing data on the type of medical degree.

Constructing Cost Profiles

In this chapter, we describe the steps we used to construct physician-level cost profiles. For the purposes of this study, we made choices that reflect the way in which health plans commonly construct these profiles. These profiles should be viewed as typical or common cost profiles. It was outside of the scope of our study to design an optimal method for constructing physician cost profiles; rather, we set out to understand the implications of the methods that were in common use. We made no assumptions at the outset about the methodological adequacy of the profiling tools currently in use.

The basic steps involve (1) selecting a method for aggregating claims into meaningful clinical groupings, which we refer to as *episodes* in our analysis; (2) developing a method for computing the costs of episodes; (3) assigning episodes to physicians; and (4) constructing a summary cost profile for each physician. Each step is described in detail in the following sections.

Creating Episodes of Care and Risk Adjustment Methodology

We used Ingenix's Episode Treatment Groups® program, version 6.0, a commercial product commonly used by health plans to aggregate claims into episodes. We chose this program because it was being used by the health plans that were participating in our study, the Massachusetts Group Insurance Commission, and other insurers to create tiered insurance products (Lake, Colby, and Peterson, 2007). We used version 6.0 because we had a research license for the study.

The episode treatment group (ETG) methodology is described in detail in previous publications (see, e.g., Dang, Pont, and Portnoy, 1996). The software evaluates all types of claims (facility, professional, and ancillary) and sorts them into mutually exclusive and exhaustive episode types, or ETGs. There are more than 600 episode types; examples include “hypo-functioning thyroid gland,” “viral meningitis,” and “cataract with surgery.” For some episode types, such as bone fractures or pregnancy, the ETG system uses sub-ETGs to designate the site of the injury (e.g., bone fracture of the foot or ankle versus bone fracture of the thigh or hip) or the presence of complications (e.g., normal pregnancy versus pregnancy with complications). For all our analyses, we used sub-ETGs when applicable.

The duration of an episode is flexible. For an acute illness, the start and end of an episode is defined by a “clean period.” This is a prespecified period during which no treatment occurs, both before the claim that identified or “triggered” the episode and after the claim that ends the episode (often 30 days). For chronic-illness episodes, no clean period is required, and the

duration is generally one year. A patient can have concurrent episodes that account for different illnesses occurring at the same time. For example, a chest X-ray and blood glucose exam performed during the same encounter may be assigned to separate episodes if the patient is treated simultaneously for both bronchitis and diabetes.

As is standard, we used only complete episodes or, for chronic illnesses, episodes that spanned a full year (designated by ETG types 0, 1, 2, or 3). Except for our method of assigning costs to each service, we used the default settings for the program (i.e., we used the default clean periods). We created more than 5.6 million episodes for analysis.

We risk-adjusted the cost profiles using Ingenix's Episode Risk Groups[®] software, version 6.0. As is typical of similar risk-adjustment systems (e.g., Diagnostic Cost Groups, Adjusted Clinical Groups), the Episode Risk Groups system uses medical claims, pharmaceutical claims, and demographic variables to determine health risk, but it differs in that it uses episodes of care as markers of risk rather than the diagnoses listed on individual medical claims (Thomas, Grazier, and Ward, 2004). For each patient, episodes experienced during a given period are mapped into 119 episode risk groups (ERGs), and then a patient-specific risk score is determined based on age, gender, and mix of ERGs. For our analyses, we used the ERG retrospective risk score.

Computing the Costs of Episodes

In computing the costs of episodes, we had to choose between using the actual reimbursement paid by the health plans and creating a standardized or average price across all plans. In our article (Adams et al., 2010), we present the results using standardized prices. The results using actual reimbursement levels are reported in Chapter Six of this report. Here, we describe the method for creating standardized prices.

We used the “allowed cost” field from the claims, which captures the total amount that can be reimbursed for a service, including copayments (Bridges to Excellence and Leapfrog Group, 2004; Rattray, 2008). We summed the allowed costs for each type of service (procedure, visit, service, and drug) across all health plans and divided by the number of services of that type, arriving at an average unit price.

Although all health plans paid for inpatient admissions based on diagnosis-related groups (DRGs), the four plans used different and incompatible systems, so the standardized price applied to a hospitalization was the average lump-sum reimbursement for a DRG *within a given health plan*. Only facility reimbursements were included. A small fraction of hospitalizations were assigned more than one DRG. These hospitalizations were not included in the calculations of standardized prices, and the standardized price that we applied to these hospitalizations was for the DRG with the highest average price.

We then created a standardized cost for each episode assigned to a physician by multiplying the number of units of each service assigned to the episode by the standardized price for that service and summing the components. We refer to the resulting total cost of an episode as the *observed cost*. We tested the sensitivity of our reliability results to the use of standardized prices versus actual reimbursements (see Chapter Six for the results).

There are different opinions about the validity of including extreme values in the calculation of cost profiles; however, most health plans use some method to exclude outliers, so we followed this convention. We set all allowed costs below the 2.5th and above the 97.5th

percentile of each service price distribution to the values at those cut points, a process known as Winsorizing. We selected Winsorizing over other methods of dealing with outliers because we found this to be a superior method in our prior work (Thomas and Ward, 2006). We also Winsorized the standardized cost for each episode by setting total episode costs falling below the 2.5th and above the 97.5th percentile distribution to the values at those cut points. We calculated the reliability of cost profiles using this method. We also tested the sensitivity of our reliability results to the use of Winsorizing by estimating reliability without excluding extreme values.

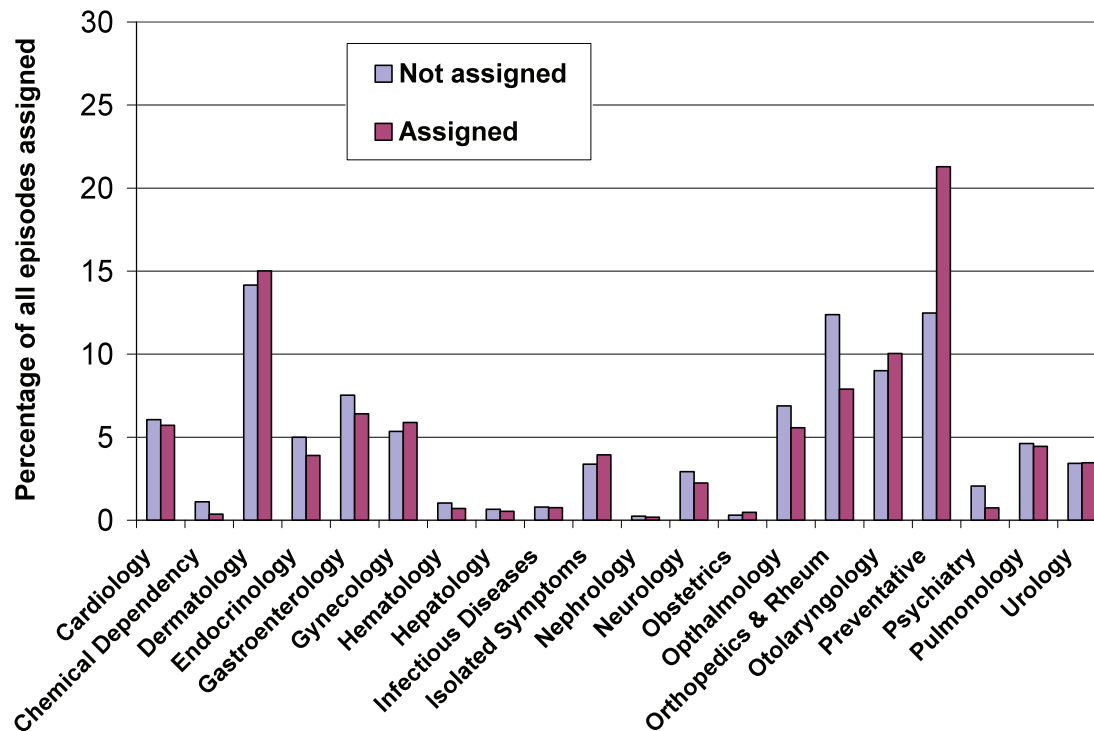
Assigning Episodes to Physicians

Because, in most situations, no accountability is assigned a priori to a physician for a patient's care, algorithms have been developed to make such assignments based on different patterns of utilization. These algorithms are broadly referred to as *attribution rules*.

In other work, we evaluated 12 potential attribution rules that used different combinations of the unit of analysis (episode or patient), signal for responsibility (visits or costs), thresholds for assigning responsibility (plurality or majority), and number of physicians to whom episodes are attributed (single or multiple) (Mehrotra et al., in press). In this report, we present our results using a common attribution rule that assigns an episode to the physician who bills the highest proportion of professional costs in that episode as long as the proportion is greater than 30 percent (Bridges to Excellence and Leapfrog Group, 2004; Rattray, 2008). If no physician met the criteria, the episode was dropped from our analyses. Overall, 52 percent of episodes could be assigned to a physician. We compared the distribution of episodes by major disease category (see Figure 3.1) and found that orthopedic and rheumatology episodes had a higher representation among unassigned than assigned episodes (12.4 percent versus 7.9 percent of all episodes), and preventive care had a lower representation among unassigned compared to assigned episodes (12.5 percent versus 21.3 percent). With these exceptions, we found no real differences between assigned and unassigned episodes.

To test the sensitivity of our findings to the attribution rule chosen, we examined two other commonly used rules. In the first, episodes are attributed to the physician who accounts for the highest fraction (minimum of 30 percent) of face-to-face encounters (based on the number of evaluation and management visits) within the episode; 50 percent of episodes could be assigned under this rule (MaCurdy et al., 2008). In the second alternative rule, we used a "patient-based" rule and assigned all episodes for a patient to the physician who accounted for the highest fraction of professional costs for that patient (minimum of 30 percent) over the two-year study period; 39 percent of episodes could be attributed under this rule (Pham et al., 2007; Centers for Medicare and Medicaid Services, 2007). At the time of this study, none of the health plans in Massachusetts was using an attribution rule that assigned care to multiple physicians, so we did not evaluate the reliability of such a rule.

Figure 3.1
Comparison of the Mix of Assigned Versus Unassigned Episodes, by Major Disease Category



Constructing a Summary Cost Profile Score

We calculated the summary cost profile score as a ratio based on all episodes attributed to each physician:

$$\text{Summary cost profile} = \frac{\text{Sum of the observed costs}}{\text{Sum of the expected costs}},$$

or, in mathematical notation:

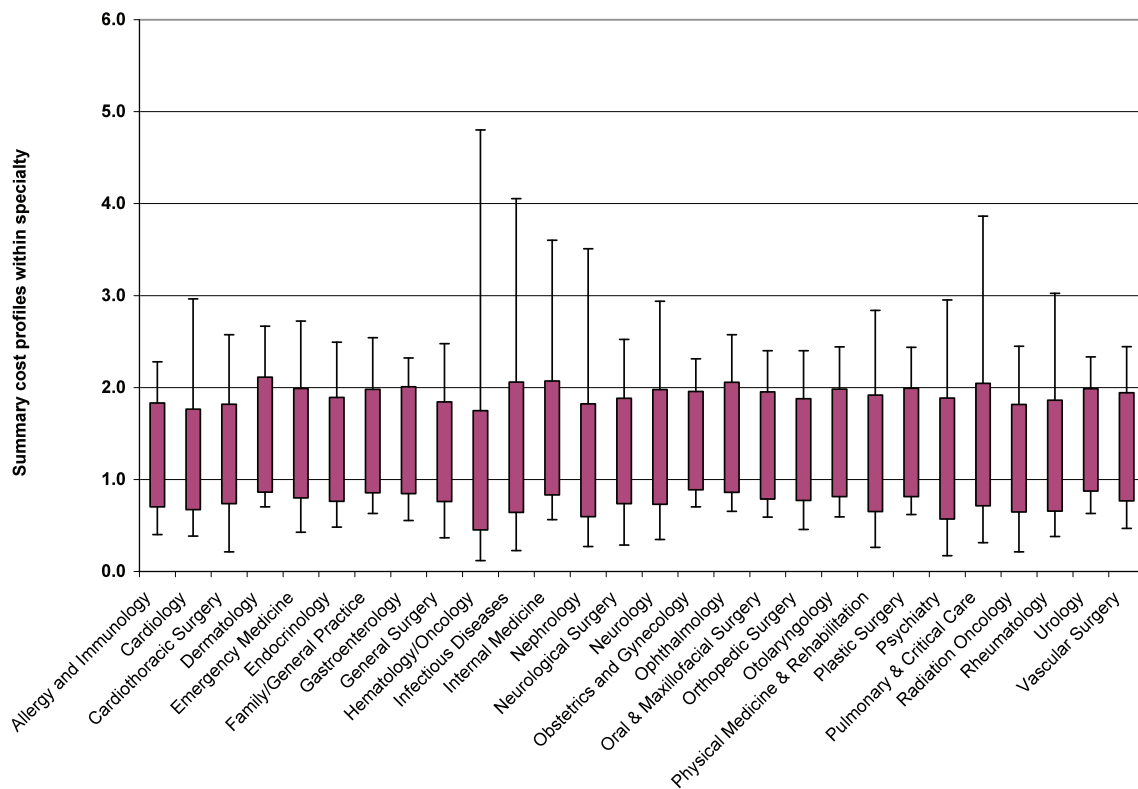
$$O/E = \frac{\left(\sum \text{observed}_i\right)}{\left(\sum \text{expected}_i\right)},$$

where the sum (i) of the observed costs is over all episodes assigned to the physician and the sum (i) of the expected costs is the sum of the averages of the equivalent set of episodes (“case-mix matched”) assigned to all physicians in the same specialty. As discussed earlier, the ERG classification system is a risk-adjustment methodology that assigns a patient’s episode to a discrete risk level. Our calculations of expected costs were specialty-specific, as recommended in prior reports (e.g., Bridges to Excellence and Leapfrog Group, 2004), because this approach

also indirectly adjusts for severity of illness (e.g., hypertension episodes assigned to nephrologists are likely different from hypertension episodes assigned to family physicians).

If the sum of observed costs exceeds the sum of expected costs (i.e., the physician is more costly than his or her peers), the physician’s cost profile score will be greater than 1. If the sum of observed costs is lower than the sum of the expected costs (i.e., the physician is less costly than his or her peers), the cost profile score will be less than 1. The summary cost profile is a continuous variable with a median near 1, a minimum value of zero, and no bound on the maximum value. The minimum value of zero is seldom observed in our data, and the maximum value rarely exceeds 10. The use of division for the ETG adjustment is analogous to the standardized mortality ratio or other adjusted metrics designed to maintain a mean near 1 and a proportion or percentage interpretation. It is perhaps surprising that the distributions are reasonably symmetric. The mean is rarely more than 20-percent higher than the median. Most cost distributions are skewed even after case-mix adjustment. The symmetry is a consequence of the *O/E* metric and the detailed adjustment of the ETGs. Figure 3.2 shows the distributions of the summary cost profile scores for each specialty.

Figure 3.2
Score Distribution of Physicians, by Specialty



NOTE: The boxes represent the 25th- to 75th-percentile spans and the whiskers represent the 5th- to 95th-percentile spans.

Calculating the Reliability of Cost Profiles

Overview of the Rationale for Examining Reliability

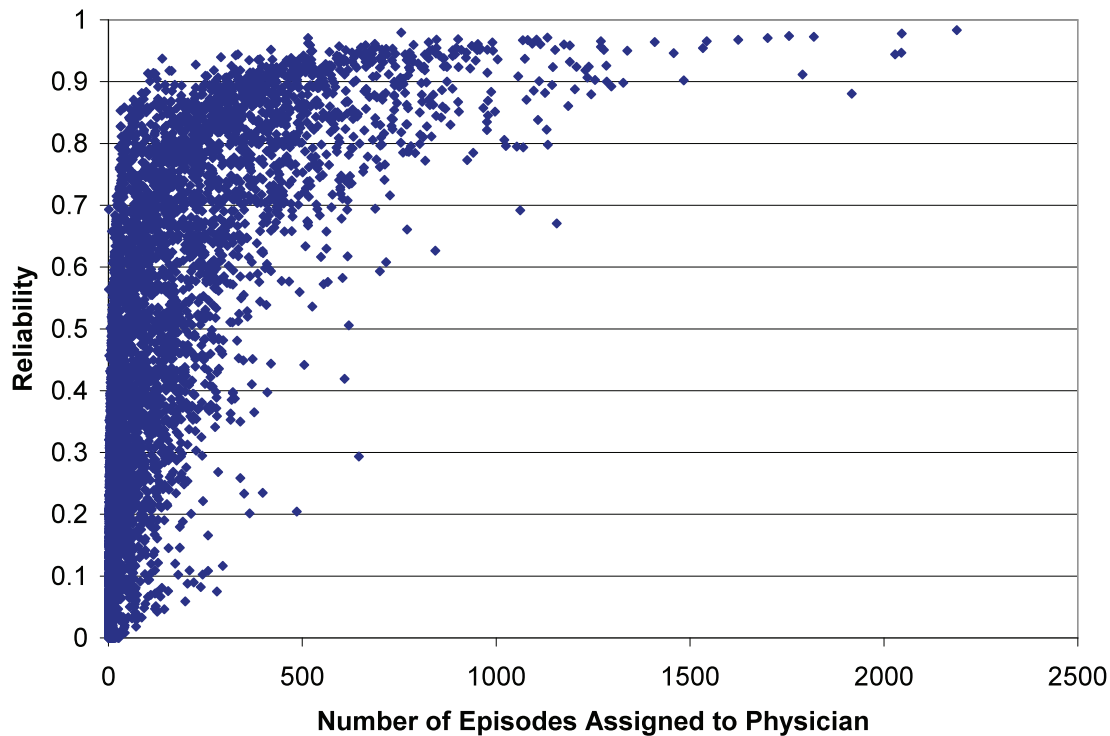
Measures are traditionally evaluated for scientific soundness by assessing validity and reliability. *Validity* refers to the degree to which a measure reflects the construct that it is intended to capture. In this case, the construct of interest is the relative use of resources by physicians. Although we are not aware of any formal validity studies on the methods for creating episodes (and this is difficult because there is no gold standard for comparison), their widespread use in a variety of settings suggests that they have face validity. Further, in examining discussions in the literature, two key issues are raised with respect to validity: whether the applications produce an adequate number of observations to accurately reflect physician performance and whether the method of assigning episodes to physicians is appropriate. The first issue is part of what motivated us to construct an aggregate data set, which captured the vast majority of physicians' commercial business. Sample size is also a relevant consideration in reliability, as discussed later. The second issue led us to test a variety of attribution methods being used by health plans and to select for inclusion in this study a commonly used method that had the highest reliability. The literature, however, is silent on the issue of reliability, with the exception of some efforts to set minimum sample sizes for inclusion in profiling as a proxy for reliability but without any formal evaluation of the success of these methods. This is the reason we chose to focus on this aspect of evaluating cost profiling measures.

Reliability is a key method for evaluating whether a measure is suitable for profiling because it indicates the proportion of signal versus noise contained in a physician's cost profile. 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 level of zero can be interpreted as meaning that all the variability in the scores is attributable to measurement or sampling error. A reliability level of 1 can be interpreted as meaning that all the variability is attributable to real differences in performance. 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 a physician's performance is good, but rather that one can confidently classify that physician's performance relative to his or her peers. Measures of physician clinical quality, patient experience, peer review, medical errors, and utilization have been evaluated for their reliability (Safran et al., 2006; Hofer, Hayward, et al., 1999; Hofer, Bernstein, et al., 2000; Hayward and Hofer, 2001).

Reliability is a function of three factors: sample size, differences between physicians in terms of performance, and measurement or sampling error. At the physician level, increasing the number of episodes assigned to a physician can increase the sample size. This was the main reason we developed an aggregated claims database for this study. Real differences in performance between physicians are the quantity of interest for the cost profile: It is a measure of heterogeneity in practice style. Measurement error also may be improved with more accurate data or improved case-mix adjustment. However, our task was to evaluate how well the methods being used by health plans perform on the test of reliability rather than to design an optimal method for cost profiling.

Sample size, while often used as a proxy for reliability, is often insufficient for this purpose. Figure 4.1 shows the relationship between sample size (i.e., the number of attributed episodes) and reliability of the cost profile for each physician in the data set (each diamond in the figure represents a physician). Although there is a relationship between sample size and reliability, sample size alone cannot predict reliability. Therefore, simple minimum sample-size rules (e.g., a physician's profile must include at least 30 episodes) are not a substitute for direct calculation of reliability. The figure also makes it clear that minimum sample sizes (e.g., 30) will not ensure sufficient reliability for a majority of physicians (National Committee for Quality Assurance, 2008).

Figure 4.1
Relationship Between Number of Assigned Episodes and Reliability of Physicians' Cost Profile Scores



Detailed Description of the Method for Calculating Reliability

In this section, we elaborate on how we calculated the reliability of physicians' cost profile scores. *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 we knew the true value. Most of the complications of reliability calculations come from various workarounds for estimating the true value. Instead of measuring the true value, researchers will often estimate the lower bound of reliability (Fleiss, Levin, and Paik, 2003). For example, the use of test-retest reliability when evaluating survey instruments is a common method for determining the lower bound of 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:

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

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 when there is a single observed period such as we face with the measurement of cost profiles.

In the case of cost profiles, we estimate reliability as a function of the components of a simple hierarchical linear model (HLM) (Raudenbush and Bryk, 2002). A simple two-level HLM separates the observed variability in physician scores into two components: variability between physicians and variability within the physician. The equivalent definition of reliability from this framework is:

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

Or, with a more intuitive labeling:

$$reliability = \frac{\sigma_{signal}^2}{\sigma_{signal}^2 + \sigma_{noise}^2}.$$

Or, made more specific to our setting:

$$reliability = \frac{\sigma_{physician-to-physician}^2}{\sigma_{physician-to-physician}^2 + \sigma_{physician-specific-error}^2}.$$

It may not be obvious to many analysts why these two versions of reliability are 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 between-physician variance is an estimate of the variance if we were able to calculate the variance of the true values. The physician-specific-error variance is the sampling or measurement error.

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:

$$reliability = \frac{\sigma^2_{physician-to-physician}}{\sigma^2_{physician-to-physician} + \frac{\sigma^2_{average_episode_error}}{n}}.$$

In this equation, the error variance has been rewritten as the average error variance for the episodes attributed to a physician divided by n , where n is the number of episodes. Each episode type has a different error variance. This notation makes clear the necessity of uncoupling the issue of “noisy” episodes from the number of episodes.

The more detailed formula makes it easier to understand some common misconceptions about reliability in the context of physician cost profiling, which is often mistakenly thought of as a property of a measurement system (e.g., the SF-12 survey). The formula demonstrates why *each* physician’s cost profile has a different reliability. First, each physician is assigned a different mix of episodes (e.g., three hypertension, two diabetes, one preventive care). This mix will affect both the physician-to-physician variance and the episode-error variance. Also, with physician cost profiles, the number of episodes attributed can vary widely from physician to physician, as shown in Figure 4.1.

This line of analysis also illustrates why reliability will vary depending on the data set and could vary when applied nationally or applied to a single metropolitan region. The physician-to-physician variation and item-error variance will be different in each data set. Even within a region, different reliabilities may be obtained from a combined data set versus a single health plan, as we demonstrate in one of the sensitivity analyses in Chapter Six.

Conceptually, the between-physician variance is the variance we would expect if we had very large sample sizes for every physician. Although the computation is done with a simple HLM estimation, the between-physician variance can be interpreted as the variance that remains after we remove the variance that we would expect from the within-physician variances. For two specialties, we have estimated a between-physician variance of zero. The interpretation of a zero variance is that there is no evidence from the model that physicians differ from each other beyond the differences we would expect to see by chance alone.

To our knowledge, there is no literature on how to calculate the standard errors for *O/E* cost profiles based on the ETG system. We used a simple episode type variance estimator to estimate the variance of the observed value. For each ETG, by sub-ETG, by ERG level in a given specialty, we estimated the variance of the observed value using the variance of all episodes of that type. In a few cases, there was only one episode of that type in the data set. Therefore, in these cases, we imputed the variance using an assumed coefficient of variation of 1 and the observed cost value. When we examined all the ETG combinations across the specialties, the assumption that the coefficient of variation was equal to 1 appeared to be generally reason-

able. We then used a simple delta method calculation to translate the variance of the observed score to the O/E scale. Bootstrap calculations produced similar values.

In more mathematical detail, the variance is as follows:

$$\begin{aligned} \text{Var}(O/E) &= \text{var}\left(\left(\sum \text{observed}_i\right) / \left(\sum \text{expected}_i\right)\right) \\ \text{Var}(O/E) &= \text{var}\left(\sum \text{observed}_i\right) / \left(\sum \text{expected}_i\right)^{-2} \\ \text{Var}(O/E) &= \left(\sum \text{var}\left(\text{observed}_i\right)\right) / \left(\sum \text{expected}_i\right)^{-2} \\ \text{Var}(O/E) &= \left(\sum \sigma_i^2\right) / \left(\sum \text{expected}_i\right)^{-2}. \end{aligned}$$

For σ_i^2 , we can use the variance of the entire population for each particular ETG, as discussed earlier. The standard error is the square root of this variance.

This simple calculation assumes that the variance of the expected values is small compared to the variance of the observed values. This assumption is justified by the much larger sample sizes used to calculate the expected values. Every σ_i^2 in the numerator has a σ_i^2/n_i in the denominator, where the sample size is the number of episodes of the type in the specialty. A more elaborate standard error can be developed by incorporating the denominator variance into the delta method calculation. We found only a modest increase in variance using the more elaborate method, ranging from 0.5 percent for internal medicine to 1.2 percent for family practice.

We created the models within each specialty because our cost profile measure is specialty-specific (i.e., “expected” costs are constructed within the specialty). Also, the applications of cost profiles are often specialty-specific. For example, tiered networks are created separately within each specialty. All of this suggests that a within-specialty analysis is an appropriate approach.

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

$$\begin{aligned} \mu_j &\sim \text{Normal}\left(\mu, \sigma_{\text{physician-to-physician}}^2\right) \\ O/E &\sim \text{Normal}\left(\mu, \sigma_{\text{physician-specific-error}}^2\right). \end{aligned}$$

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 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, resulting in zero reliability for all physicians in the specialty.

Table 4.1 presents the HLM estimates of the physician-to-physician variance components by specialty. Note that two specialties, cardiothoracic surgery and radiation oncology, have estimated variance components of zero. Note, too, that vascular surgery’s variance component is not statistically significantly different from zero.

Table 4.1
Estimates of the Physician-to-Physician Variance Components

Specialty	Physician-to-Physician Variance	Standard Error	p-value
Allergy and Immunology	0.069	0.013	<0.0001
Cardiology	0.130	0.013	<0.0001
Cardiothoracic Surgery	0.000	—	—
Dermatology	0.096	0.009	<0.0001
Emergency Medicine	0.049	0.005	<0.0001
Endocrinology	0.026	0.006	<0.0001
Family/General Practice	0.022	0.002	<0.0001
Gastroenterology	0.036	0.003	<0.0001
General Surgery	0.056	0.007	<0.0001
Hematology/Oncology	0.088	0.021	<0.0001
Infectious Diseases	0.066	0.017	<0.0001
Internal Medicine	0.037	0.002	<0.0001
Nephrology	0.058	0.016	0.0002
Neurological Surgery	0.024	0.010	0.0060
Neurology	0.067	0.009	<0.0001
Obstetrics and Gynecology	0.029	0.002	<0.0001
Ophthalmology	0.043	0.004	<0.0001
Oral and Maxillofacial Surgery	0.027	0.006	<0.0001
Orthopedic Surgery	0.026	0.004	<0.0001
Otolaryngology	0.044	0.006	<0.0001
Physical Medicine and Rehabilitation	0.083	0.019	<0.0001
Plastic Surgery	0.036	0.009	<0.0001
Psychiatry	0.089	0.012	<0.0001
Pulmonary and Critical Care	0.061	0.015	<0.0001
Radiation Oncology	0.000	—	—
Rheumatology	0.136	0.023	<0.0001
Urology	0.017	0.003	<0.0001
Vascular Surgery	0.005	0.009	0.3015

Table 4.2 presents the average quantities and ranks for each specialty of the three key factors that determine reliability: the physician-to-physician variance, the median number of episodes assigned, and the variance of the average episode error. The table is sorted from highest to lowest reliability. In each of the dimensions, a rank of 1 is best (and 28 is worst). For example, dermatology has the number one rank on n because it has the largest number of attributed

Table 4.2
Relative Ranks for the Components of Reliability, by Speciality

Speciality	Median Reliability	Rank, Median Reliability	Average Number of Episodes per Physician	Rank, Average Number of Episodes per Physician	Physician-to-Physician Standard Deviation	Rank, Physician-to-Physician Standard Deviation	Median Standard Error of Physician Cost Profile Score	Rank, Median Standard Error of Physician Cost Profile Score	Median Average Episode Error	Rank, Median Average Episode Error
Dermatology	0.97	1	732	1	0.31	3	0.06	1	1.62	6
Allergy and Immunology	0.86	2	192	10	0.26	7	0.11	5	1.52	3
Ophthalmology	0.83	3	311	5	0.21	14	0.09	2	1.59	4
Gastroenterology	0.79	4	264	7	0.19	16	0.1	6	1.62	7
Otolaryngology	0.79	4	266	6	0.21	14	0.11	3	1.79	10
Obstetrics and Gynecology	0.74	6	325	4	0.17	19	0.1	4	1.80	11
Rheumatology	0.69	7	195	9	0.37	1	0.24	16	3.35	24
Internal Medicine	0.66	8	339	3	0.19	16	0.14	9	2.58	19
Family/General Practice	0.61	9	383	2	0.15	23	0.12	7	2.35	18
Cardiology	0.58	10	104	17	0.36	2	0.31	20	3.16	23
Urology	0.54	11	230	8	0.13	25	0.12	8	1.82	12
Neurology	0.52	12	78	20	0.26	7	0.25	17	2.21	14
Emergency Medicine	0.48	13	94	19	0.22	13	0.23	14	2.23	15
Plastic Surgery	0.47	14	126	11	0.19	16	0.2	10	2.24	16
General Surgery	0.46	15	114	14	0.24	11	0.25	18	2.67	20
Oral and Maxillofacial Surgery	0.4	16	55	23	0.16	20	0.2	11	1.48	2

Table 4.2—Continued

Specialty	Median Reliability	Rank, Median Reliability	Average Number of Attributed Episodes per Physician	Rank, Average Number of Attributed Episodes per Physician	Physician-to-Physician Standard Deviation	Rank, Physician-to-Physician Standard Deviation	Median Standard Error of Physician Cost Profile Score	Rank, Median Standard Error of Physician Cost Profile Score	Median Average Episode Error	Rank, Median Average Episode Error
Endocrinology	0.37	17	107	16	0.16	20	0.21	12	2.17	13
Orthopedic Surgery	0.36	18	123	13	0.16	20	0.21	13	2.33	17
Physical Medicine and Rehabilitation	0.34	19	50	24	0.29	6	0.41	22	2.90	21
Psychiatry	0.33	20	14	27	0.3	4	0.43	23	1.61	5
Neurological Surgery	0.31	21	41	25	0.15	23	0.23	15	1.47	1
Nephrology	0.23	22	73	21	0.24	11	0.44	24	3.76	25
Pulmonary and Critical Care	0.2	23	125	12	0.25	10	0.5	25	5.59	27
Hematology/Oncology	0.18	24	60	22	0.3	4	0.63	26	4.88	26
Infectious Diseases	0.13	25	109	15	0.26	7	0.65	27	6.79	28
Vascular Surgery	0.05	26	96	18	0.07	26	0.31	19	3.04	22
Cardiothoracic Surgery	0	27	24	26	0	27	0.34	21	1.67	8
Radiation Oncology	0	27	6	28	0	27	0.68	28	1.67	8

episodes per M.D. The number one rank for average episode error goes to neurological surgery, which has the smallest average episode error.

Examining the ranks, it is easier to see why one specialty has a higher reliability than another specialty. Consider dermatology. Dermatology has the highest average number of episodes and ranks third in provider-to-provider variation and sixth in episode average standard error. All of these ranks contribute to a high reliability for dermatology, but it is sample size that carries dermatology into first place. The bottom of the reliability list is dominated by specialties with low or zero provider-to-provider variation: vascular surgery, cardiothoracic surgery, and radiation oncology. These specialties have little or no evidence of providers being different from each other. In the middle of the range, the three elements trade off to determine the specialty's reliability. Consider gastroenterology: Despite a bottom-half ranking on provider-to-provider variation, top quartile sample sizes combined with top quartile average episode standard errors carry it to a fourth-place ranking in reliability. On the lower end of the scale, consider hematology/oncology: The fourth-highest physician-to-physician variation cannot overcome noisy episodes and small sample sizes, resulting in the fifth-lowest reliability.

The Relationship Between Reliability and Misclassification

Fundamentally, reliability is the measure of whether it can be determined that a physician is different 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 for cost profiling. In this chapter, 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 likelihood that a physician will be correctly or incorrectly classified in a particular application.

Figure 5.1 shows the relationship between the physicians' score distributions and the underlying reliability. Each panel of Figure 5.1 shows the true distribution of the physicians' scores (the solid bell-shaped curve), with dashed bell-shaped curves showing the sampling distribution for ten randomly selected physicians. At a reliability level of 0.5, it is difficult to detect differences between physicians. At a level of 0.7, we start to see differences between some physicians and the mean. At a reliability of 0.9, we start to see significant differences between pairs of physicians.

If assignment to categories (e.g., above-average costs) is based on relative comparisons (e.g., top 10 percent of the distribution of physicians' scores), reliability can be used to estimate the probability of misclassification. (If assignment to categories is based on a fixed external standard, e.g., a cost profile of 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.)

We now illustrate the relationship between misclassification and reliability using the simplest categorization system—a high-performance or two-category network (Draper, Liebhaber, and Ginsburg, 2007). In this illustrative application of profiling, the 25 percent of physicians who have the lowest cost profiles are labeled “lower-cost” (in Figure 3.2 in Chapter Three, these are the physicians in each specialty in the bottom “whisker” part of the plot), and the remaining 75 percent are labeled “not lower-cost.”

In such a categorization system, there are two types of misclassification errors: (1) flagging a not-lower-cost physician as lower-cost (equivalent to a false positive) and (2) failing to flag a lower-cost physician as lower-cost (equivalent to a false negative).

To start, we explore groups of physicians with the same reliability and known cut points. By *known cut points* we mean cut points that can be expressed as percentiles (e.g., 25th percentile) of the true distribution of physician profile scores in the absence of sample error. Later, we use a more realistic model that includes estimated cut points and mixtures of physician reliabilities.

Figure 5.1
Relationship Between Reliability and Physicians' Score Distribution

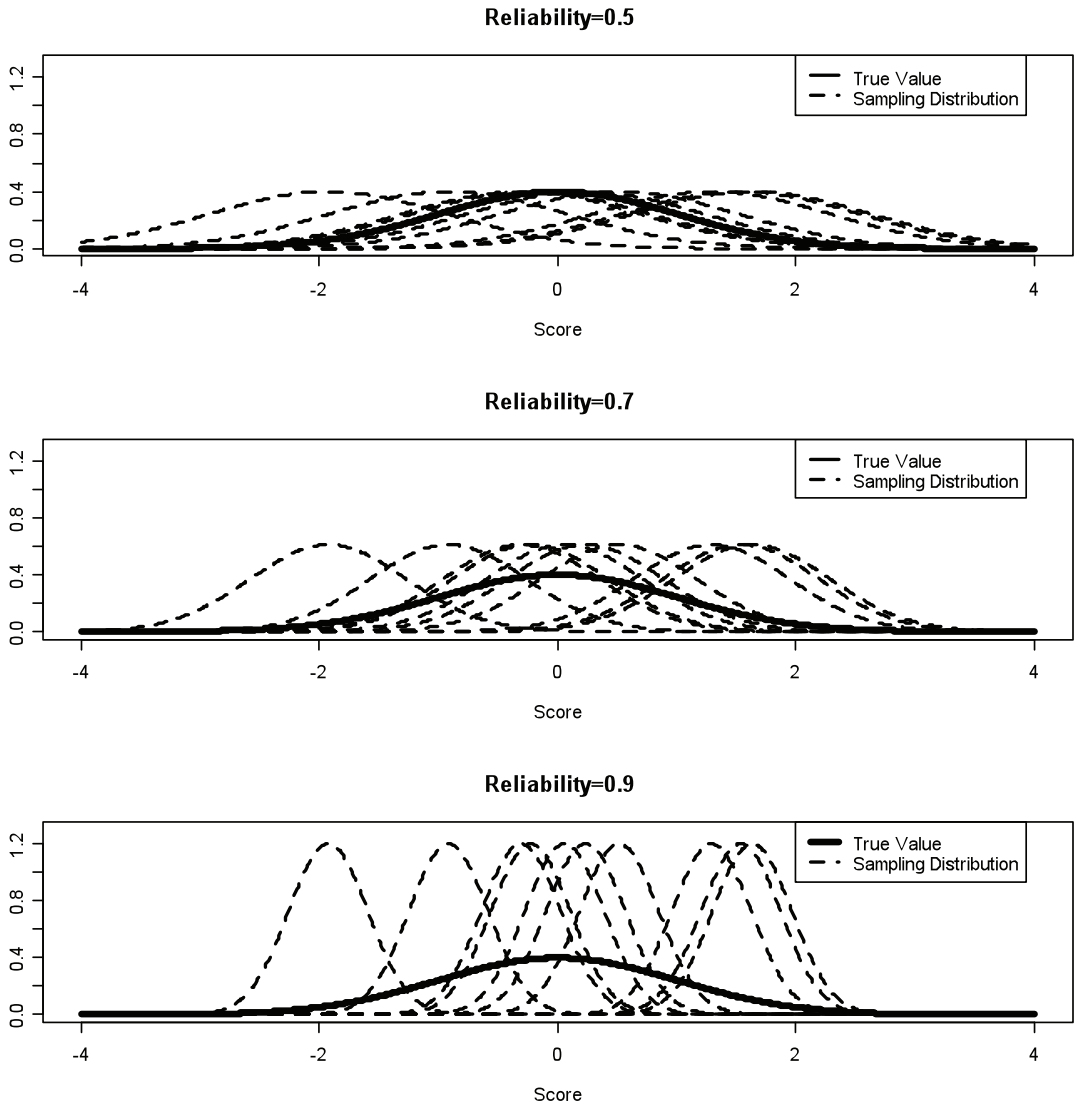
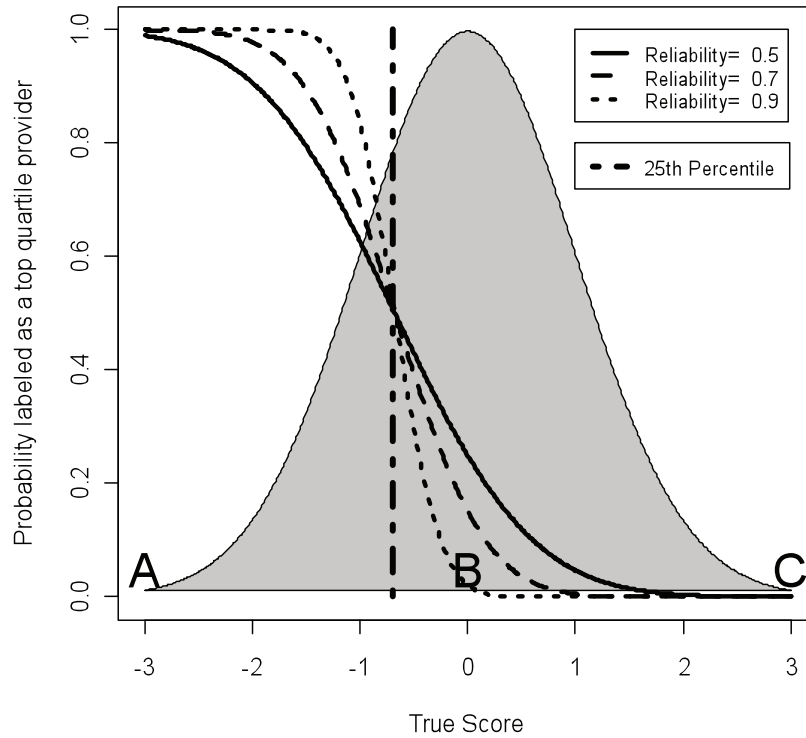


Figure 5.2 shows the probability of being labeled a high-performing physician when *every* physician's cost profile has a reliability of 0.5, 0.7, or 0.9. The gray bell-shaped curve shows the distribution of the true scores of the physicians. A dashed vertical line shows the 25th percentile cutoff in this distribution (25 percent of the physicians' scores are to the left of the dark line). The true score is expressed in units of standard deviations from the mean. Each of the curves (0.5, 0.7, and 0.9) shows, for a given true score, the probability of being classified as lower-cost.

Figure 5.2
Probability of Being Labeled 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 into the left tail of the distribution (shown as A in the figure), he or she will be labeled lower-cost at any level of reliability. At a reliability of 0.5 or 0.7, even an average physician (shown as B in the figure) can be labeled lower-cost with a probability of up to 25 percent. A physician who is clearly high-cost (shown as C in the figure) has only a low probability of being labeled lower-cost at any reliability.

Table 5.1 summarizes across the entire true-score distribution the misclassification probabilities at various levels of reliability. Note that even at a reliability of 0.9, a substantial rate of misclassification can occur. The table demonstrates the relationship between reliability and misclassification when 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.

The first issue is relaxing the assumption of a known cut point in the true physician distribution, as illustrated in Figure 5.3. The solid bell-shaped curve is the true score distribution. 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 would 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. One of the major sources of misclassification can be that the selection of the cut point is driven by the wider dispersion of the low-reliability physicians.

Table 5.1
Misclassification Probabilities for a 25th-Percentile Cut Point at Various Levels of Reliability

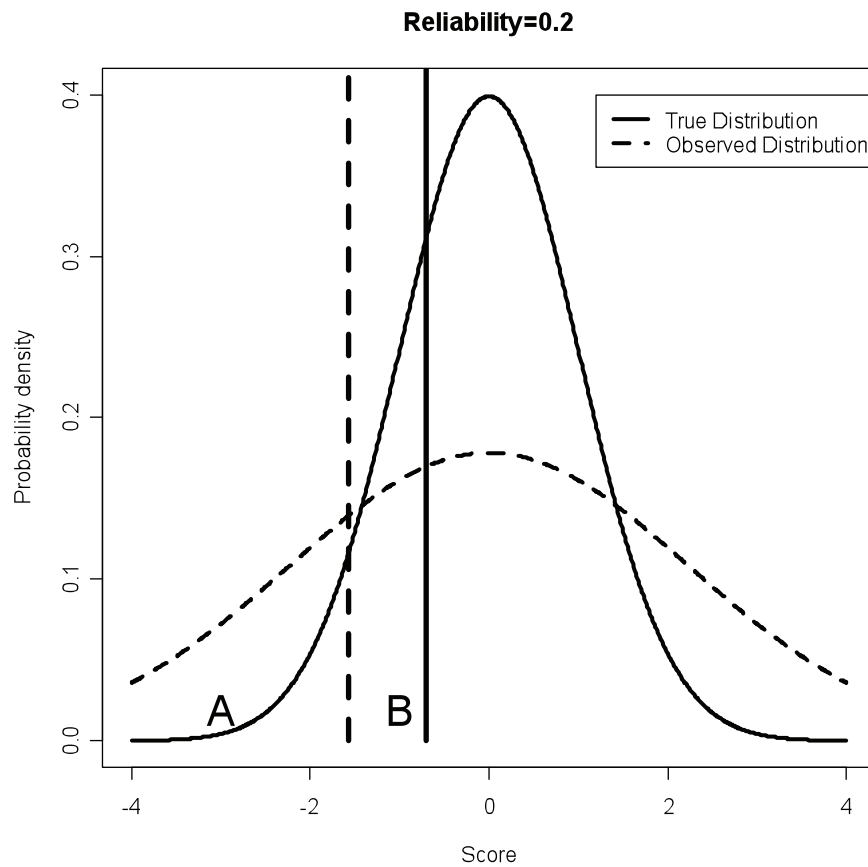
Reliability	True Lower-Cost, Labeled Not Lower-Cost (%)	True Not-Lower-Cost Labeled Lower-Cost (%)	Overall Misclassification Rate (%)	Labeled Lower-Cost and Actually Lower-Cost (%)
0.05	64.3	21.5	32.2	35.6
0.10	60.9	20.1	30.3	39.3
0.15	57.2	18.9	28.5	43.0
0.20	54.4	18.2	27.3	45.5
0.25	51.8	17.2	25.9	48.3
0.30	49.2	16.2	24.5	51.1
0.35	46.6	15.4	23.2	53.6
0.40	44.3	14.8	22.2	55.6
0.45	41.9	13.9	20.9	58.2
0.50	39.0	13.1	19.6	60.8
0.55	36.5	12.4	18.4	63.1
0.60	34.4	11.4	17.2	65.7
0.65	31.8	10.6	15.9	68.2
0.70	28.9	9.8	14.6	70.7
0.75	26.5	8.7	13.2	73.8
0.80	23.5	7.7	11.7	76.8
0.85	19.6	6.7	9.9	80.0
0.90	16.3	5.5	8.2	83.5
0.95	11.5	3.8	5.7	88.6
1.0	0	0	0.0	100.0

The physician labeled 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 problem is physician B. Because physician B's true value lies between the true and the empirical 25th percentiles, the better physician B's reliability, the more likely he or she is to be misclassified.

To estimate the misclassification probabilities for a given specialty (see Adams et al., 2010), we performed the following steps:

1. Estimate the 25th percentile of the observed cost score for each specialty from the data.
2. Estimate the physician-to-physician variance for the specialty (one of the components of the reliability calculation).
3. Determine where the 25th percentile of the observed cost score falls in the true scale. (As illustrated earlier, it will always be to the left of the 25th percentile of the true score distribution.) This is done by selecting the 25th percentile of the distribution with the observed mean and the physician-to-physician variance estimate.

Figure 5.3
Probability of Being Labeled Lower-Cost Based on True Score and Reliability
of Physician's Cost Profile



4. For each reliability level, calculate the misclassification probability.
5. Merge the misclassification probabilities with each physician. Each physician is assigned the misclassification probability that would be expected for his or her cost profile score reliability.
6. Sum the misclassification probabilities within the specialty to get specialty-wide misclassification rates.

Table 5.2 shows actual rates of misclassification, by specialty, for the sample. Among physicians who would be labeled low-cost, the proportion within each specialty who are not low-cost ranges from 11 percent for dermatology to 67 percent for vascular surgery. Among those who are labeled not low-cost, the proportion who are actually low-cost ranges from 8 percent for dermatology to 22 percent for four specialties. The overall misclassification rates range from 8 percent for dermatology (the specialty with the highest median reliability) to 36 percent for vascular surgery (the specialty with the lowest non-zero median reliability). When reliability is equal to zero, we cannot estimate a misclassification rate because misclassification is not defined.

Table 5.2
Rates of Misclassification, by Specialty, for a Two-Tiered Network

Specialty	Number of Physicians in Specialty	Of Those Classified as Low-Cost, Proportion Not Low-Cost (%)	Of Those Classified as Not Low-Cost, Proportion Low-Cost (%)	Overall Misclassification Rate (%)
Allergy and Immunology	96	29	11	15
Cardiology	708	40	13	20
Cardiothoracic Surgery ^a	99	—	—	—
Dermatology	343	11	8	8
Emergency Medicine	710	43	16	22
Endocrinology	169	50	19	25
Family/General Practice	1,065	39	16	21
Gastroenterology	426	32	11	16
General Surgery	579	47	14	23
Hematology/Oncology	308	58	22	28
Infectious Diseases	234	61	22	29
Internal Medicine	2,979	50	22	25
Nephrology	199	55	21	27
Neurological Surgery	106	53	18	26
Neurology	434	42	16	22
Obstetrics and Gynecology	922	36	10	17
Ophthalmology	548	28	12	16
Oral and Maxillofacial Surgery	174	47	17	23
Orthopedic Surgery	580	50	17	25
Otolaryngology	229	29	13	16
Physical Medicine and Rehabilitation	143	48	17	24
Plastic Surgery	119	43	15	21
Psychiatry	727	48	19	24
Pulmonary and Critical Care	362	58	21	28
Radiation Oncology ^a	65	—	—	—
Rheumatology	177	33	14	18
Urology	216	40	15	20
Vascular Surgery	72	67	22	36

^a Because the physician-to-physician variance is zero, we cannot distinguish physicians' performance from one another in this specialty. Thus, it is not possible to estimate a misclassification rate.

Sensitivity Analyses

We tested the sensitivity of our results by calculating the reliability for the physicians in our sample and varying the five design choices. We also examined the effect on misclassification of two alternative categorization schemes. After each choice, we provide the rationale for the alternative method:

1. Do not Winsorize the standardized price and episode costs. Including extreme values might increase reliability by increasing physician-to-physician variation.
2. Do not use standardized prices. Using actual reimbursements might increase variability, which, in turn, would improve reliability by increasing physician-to-physician variation.
3. Calculate reliability separately for each of the four health plans. Aggregation of data across health plans may affect physician-to-physician variability because of differences in benefit design, plan management, and data-handling techniques. Thus, reliability might increase using individual health plan data.
4. Vary attribution rules. Because different attribution rules might change which episodes are assigned to which physicians, reliability could change. We tested two alternative rules: a patient-based rule (all costs for a patient are used to assign a physician) and an episode-based rule that used the relative proportion of face-to-face visits within an episode, rather than costs, to determine episode assignment.
5. Require physicians to have at least 30 assigned episodes. Removing physicians likely to have low-reliability cost profile scores based on sample size alone might make the problem of reliability among the remaining physicians less pronounced.
6. Test classification systems with a different cut point and more than two categories. The problem of misclassification might be reduced if different choices were made.

For ease of comparison, we compare the median reliability for each specialty to results for the first five of the sensitivity analyses (see Table 6.1). Because of confidentiality commitments, we do not list the health plans by name. In addition, because there is no clear way to determine the level of difference in reliability that might be important, we took a two-step approach. First, we flagged specialties for which the alternative analytic approach resulted in the median reliability changing by more than five percentage points (either an increase or decrease). These specialties are identified by shading in Table 6.1. Second, among the specialties for which the change was at least this large, we flagged those for which the alternate analysis resulted in the median being above one of the thresholds provided in our analysis (0.7). These specialties are marked with an asterisk in Table 6.1. Two asterisks indicate that the median moved from above 0.7 to below.

Table 6.1
Median Reliability, by Specialty, for Each Sensitivity Analysis (compared to results reported in Adams et al., 2010)

Specialty	Median Reliability Results in Adams et al. (2010)	With 30-Episode Cutoff	>30-Episode Cutoff	Non-Winsorized	Allowed Cost	Health Plan A	Health Plan B	Health Plan C	Health Plan D	Patient Cost Plurality Attribution	Episode Visit Plurality Attribution
Allergy and Immunology	0.86	0.90	0.90	0.80	0.76	0.86	0.34**	0.00**	0.75	0.28**	0.75
Cardiology	0.58	0.66	0.65	0.65	0.56	0.61	0.04	0.07	0.00	0.40	0.40
Cardiothoracic Surgery	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Dermatology	0.97	0.97	0.97	0.94	0.93	0.98	0.80	0.16**	0.82	0.55**	0.94
Emergency Medicine	0.48	0.55	0.56	0.42	0.30	0.54	0.04	0.00	0.00	0.13	0.43
Endocrinology	0.37	0.62	0.55	0.30	0.27	0.50	0.00	0.00	0.00	0.11	0.28
Family/General Practice	0.61	0.70*	0.66	0.37	0.47	0.70*	0.14	0.23	0.23	0.25	0.31
Gastroenterology	0.79	0.83	0.82	0.68**	0.69**	0.86	0.31**	0.09**	0.18**	0.44**	0.45**
General Surgery	0.46	0.61	0.62	0.57	0.35	0.38	0.02	0.03	0.05	0.12	0.33
Hematology/Oncology	0.18	0.53	0.37	0.14	0.15	0.16	0.00	0.00	0.00	0.08	0.18
Infectious Diseases	0.13	0.62	0.59	0.10	0.12	0.17	0.01	0.12	0.00	0.15	0.07
Internal Medicine	0.66	0.78*	0.75*	0.94*	0.40	0.73*	0.25	0.17	0.11	0.35	0.45
Nephrology	0.23	0.37	0.36	0.14	0.16	0.29	0.03	0.00	0.00	0.09	0.04
Neurological Surgery	0.31	0.42	0.65	0.37	0.32	0.44	0.00	0.00	0.15	0.00	0.40
Neurology	0.52	0.67	0.42	0.49	0.38	0.57	0.02	0.00	0.20	0.20	0.25
Obstetrics and Gynecology	0.74	0.77	0.77	0.75	0.73	0.62*	0.16**	0.16**	0.16**	0.26**	0.62**

Table 6.1—Continued

Specialty	Median Reliability Results in Adams et al. (2010)	With 30-Episode Cutoff	>30-Episode Cutoff	Non-Winsorized	Allowed Cost	Health Plan A	Health Plan B	Health Plan C	Health Plan D	Patient Cost Plurality Attribution	Episode Visit Plurality Attribution
Ophthalmology	0.83	0.85	0.85	0.76	0.75	0.81	0.34**	0.13**	0.16**	0.26**	0.72
Oral and Maxillofacial Surgery	0.40	0.55	0.54	0.32	0.20	0.47	0.00	0.00	0.00	0.21	0.29
Orthopedic Surgery	0.36	0.40	0.64	0.45	0.52	0.51	0.04	0.06	0.00	0.21	0.36
Otolaryngology	0.79	0.84	0.84	0.13**	0.60**	0.83	0.25**	0.39**	0.37**	0.24**	0.71
Physical Medicine and Rehabilitation	0.34	0.75	0.63	0.39	0.28	0.47	0.00	0.00	0.00	0.22	0.33
Plastic Surgery	0.47	0.52	0.74*	0.42	0.47	0.57	0.08	0.05	0.22	0.07	0.41
Psychiatry	0.33	0.64	0.67	0.32	0.31	0.34	0.00	0.00	0.28	0.11	0.34
Pulmonary and Critical Care	0.20	0.32	0.33	0.14	0.20	0.31	0.05	0.00	0.03	0.13	0.12
Radiation Oncology	0.00	—	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02
Rheumatology	0.69	0.75	0.75*	0.59	0.62	0.70	0.11	0.00	0.36	0.27	0.62
Urology	0.54	0.58	0.56	0.63	0.62	0.70*	0.01	0.13	0.00	0.20	0.43
Vascular Surgery	0.05	0.05	0.07	0.15	0.17	0.30	0.00	0.00	0.00	0.00	0.06

NOTE: Shading indicates that the change in the median is greater than 5 percentage points. The shading does not identify differences that were due exclusively to rounding. * = median threshold increased from below to above 0.70. ** = median threshold decreased from above to below 0.70. The asterisks do not account for rounding.

In the first sensitivity analysis, when we did not Winsorize costs, we found that 18 of 28 specialties had median reliabilities that changed by more than five percentage points. Among these specialties, 11 had lower median reliabilities and seven had higher median reliabilities. For only three specialties did the new median change the threshold assessment; one (internal medicine) moved above the 0.7 threshold, and two (gastroenterology and otolaryngology) moved below the 0.7 threshold. Thus, we conclude that Winsorizing leads to somewhat higher reliability estimates overall.

For the second sensitivity analysis, when we used actual reimbursements rather than standardized prices, the median reliability for 17 specialties changed more than five percentage points. Of those specialties, only three had higher median reliabilities (orthopedic surgery, urology, and vascular surgery), and none of the changes moved the median above the 0.7 threshold. Thus, we conclude that using standardized prices produces higher reliabilities in our data sample.

For the third sensitivity analysis, we analyzed the health plans separately and found that three of the four plans had markedly lower reliabilities for all specialties. The fourth plan, health plan A, had higher median reliabilities for 15 specialties and lower median reliabilities for two. Conducting the analysis separately for health plan A moved only three specialties above the 0.7 median reliability threshold (internal medicine, from 0.66 to 0.73; family/general practice, from 0.61 to 0.70; and urology, from 0.54 to 0.70). One specialty, obstetrics and gynecology, moved from above the threshold (median of 0.74) to below the threshold (median of 0.62). The increased reliability was driven by lower physician-specific error estimates when using just one health plan's data. Thus, although health plan A loses some reliability from aggregation, the substantial gains in reliability for the other plans suggests that aggregation is the superior approach for most stakeholders.

For the fourth sensitivity analysis using different attribution rules, for the patient cost plurality attribution rule, the median reliabilities changed for 24 of 28 specialties and all were lower. For the episode visit plurality attribution rule, median reliabilities changed for 19 specialties, and in all but one case (neurological surgery), the median was lower with the alternative rule. Thus, we conclude that the approach to attribution that we used in Adams et al. (2010) produces the highest reliability.

In the fifth sensitivity analysis, we calculated reliability only among physicians with at least 30 episodes. Using this cutoff, not surprisingly, increased the median reliability more than five percentage points for 18 of the 28 specialties but substantially decreased the number of physicians who could be profiled (8,689 versus 12,789).

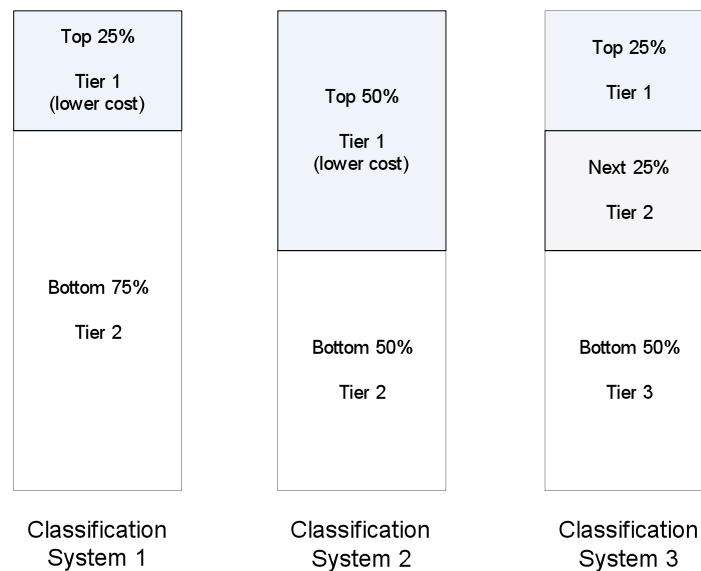
The sixth sensitivity analysis explored whether rates of misclassification would be different if alternative methods were used to identify low-cost physicians. We used 25 percent as the percentile cutoff in Adams et al. (2010), but a different percentile could be chosen. More elaborate systems with more than two categories are also possible. The key question is whether these alternative systems result in higher or lower misclassification probabilities than what we report in the article. To explore this question, we compared the overall misclassification rates for three systems, which are illustrated in Figure 6.1:

1. A two-tier system identifying the 25 percent of physicians with the lowest cost profiles. This is the system discussed in Adams et al. (2010).
2. A similar two-tier system that identifies the 50 percent of physicians with the lowest cost profiles.

3. A three-tier system that 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.

Table 6.2 presents overall misclassification rates by specialty for the three systems. The misclassification rates are higher in the two alternative systems. Classification system 2 (the lowest-50-percent cost system) has its cut point at the center of the distribution where there are many physicians with the potential to be misclassified. Classification system 3 (the three-category system) has even higher misclassification rates, because there are more ways to misclassify the physician.

Figure 6.1
Three Illustrative Classification Systems



NOTE: Classification system 1 is the focus of the analysis reported in Adams et al. (2010).

Table 6.2
Overall Misclassification Rates for Three Systems

Specialty	Misclassification Rate (%)		
	System 1: Lowest 25% Cost Profile	System 2: Lowest 50% Cost Profile	System 3: Two Low-Cost Tiers of 25% Each
Allergy and Immunology	15	17	32
Cardiology	20	24	42
Cardiothoracic Surgery	—	—	—
Dermatology	8	9	18
Emergency Medicine	22	27	47
Endocrinology	25	29	51
Family/General Practice	21	24	43

Table 6.2—Continued

Specialty	Misclassification Rate (%)		
	System 1: Lowest 25% Cost Profile	System 2: Lowest 50% Cost Profile	System 3: Two Low-Cost Tiers of 25% Each
Gastroenterology	16	19	35
General Surgery	23	28	47
Hematology/Oncology	28	33	58
Infectious Diseases	29	34	58
Internal Medicine	25	24	48
Nephrology	27	33	58
Neurological Surgery	26	33	55
Neurology	22	26	46
Obstetrics and Gynecology	17	20	37
Ophthalmology	16	18	33
Oral and Maxillofacial Surgery	23	29	50
Orthopedic Surgery	25	30	52
Otolaryngology	16	19	35
Physical Medicine and Rehabilitation	24	29	50
Plastic Surgery	21	27	47
Psychiatry	24	30	52
Pulmonary and Critical Care	28	34	58
Radiation Oncology	—	—	—
Rheumatology	18	21	40
Urology	20	25	45
Vascular Surgery	36	43	67

References

- Adams JL, Mehrotra A, Thomas JW, McGlynn EA. Physician cost profiling—Reliability and risk of misclassification, *N Engl J Med*. 2010;362(11).
- Brennan TA, Spettell CM, Fernandes J, Downey RL, Carrara LM. Do managed care plans' tiered networks lead to inequities in care for minority patients? *Health Aff (Millwood)*. 2008;27(4):1160–1166.
- Bridges to Excellence and Leapfrog Group. *Measuring Provider Efficiency Version 1.0: A Collaborative Multi-Stakeholder Effort*. December 31, 2004. As of May 13, 2009: http://www.bridgestoexcellence.org/Documents/Measuring_Provider_Efficiency_Version1_12-31-20041.pdf
- Centers for Medicare and Medicaid Services. *Medicare Resource Use Measurement Plan*. Baltimore, MD; undated. As of December 30, 2009: http://www.cms.hhs.gov/QualityInitiativesGenInfo/downloads/ResourceUse_Roadmap_OEA_1-15_508.pdf
- . Medicare Physician Group Practice Demonstration: Physicians Groups Improve Quality and Generate Savings Under Medicare Physician Pay for Performance Demonstration. Fact sheet. Baltimore, MD; 2007.
- Dang DK, Pont JM, Portnoy MA. Episode treatment groups: an illness classification and episode building system—part II. *Med Interface*. 1996;9(4):122–128.
- Draper DA, Liebhaber A, Ginsburg PB. *High-Performance Health Plan Networks: Early Experiences: Center for Health System Change*. Issue Brief 111. Washington, DC: Center for Studying Health System Change; 2007.
- Fleiss J, Levin B, Paik M. *Statistical Methods for Rates and Proportions*. Indianapolis, IN: Wiley-Interscience; 2003.
- Hayward RA, Hofer TP. Estimating hospital deaths due to medical errors: preventability is in the eye of the reviewer. *JAMA*. 2001;286(4):415–420.
- 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):152–161.
- 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):2098–2105 and comments.
- Institute of Medicine. *Performance Measurement: Accelerating Improvement*. Washington, DC; 2006.
- Lake T, Colby M, Peterson S. *Health Plans' Use of Physician Resource Use and Quality Measures*. Washington, DC: Mathematica Policy Research; October 24, 2007.
- MaCurdy T, Theobald N, Kerwin J, Ueda K. *Prototype Medicare Resource Utilization Report Based on Episode Groupers*. Burlingame, CA: Acumen; November 2008. As of May 13, 2009: <http://www.cms.hhs.gov/reports/downloads/MaCurdy2.pdf>
- Mehrotra A, Adams JL, Thomas JW, McGlynn EA. The impact of different attribution rules on individual physician cost profiles. *Ann Intern Med*. In press.
- Milstein A and Lee TH. Comparing physicians on efficiency. *N Engl J Med*. 2007;357(26):2649–2652.
- National Committee for Quality Assurance. *Standards and Guidelines for the Certification of Physician and Hospital Quality*. Washington, DC; 2008.

National Quality Forum. *Measurement Framework: Evaluating Efficiency Across Patient-Focused Episodes of Care*; 2009.

Pham HH, Schrag D, O'Malley AS, Wu B, Bach PB. Care patterns in Medicare and their implications for pay for performance. *N Engl J Med*. 2007;356(11):1130–1139.

Rattray MC, *Measuring Healthcare Resources Using Episodes of Care*. Seattle, WA: CareVariance; 2008. As of May 13, 2009:

http://www.carevariance.com/images/Measuring_Healthcare_Resources.pdf

Raudenbush S, Bryk A. *Hierarchical Linear Models: Applications and Data Analysis Methods*. 2nd ed. Newbury Park, CA: Sage; 2002.

Safran DG, Karp M, Coltin K, Chang H, Li A, Ogren J, Rogers WH. Measuring patients' experiences with individual primary care physicians: Results of a statewide demonstration project. *J Gen Intern Med*. 2006;21(1):13–21.

Sandy LG, Rattray MC, Thomas JW. Episode-based physician profiling: A guide to the perplexing. *J Gen Intern Med*. 2008;23(9):1521–1524.

Sirovich B, Gallagher PM, Wennberg DE, Fisher ES. Discretionary decision making by primary care physicians and the cost of U.S. health care. *Health Aff (Millwood)*. 2008;27(3):813–823.

Sorbero ME, Damberg CL, Shaw R, Teleki S, Lovejoy S, DeChristofaro A, Dembosky J, Schuster C. *Assessment of Pay-for-Performance Options for Medicare Physician Services: Final Report*. Washington, DC, U.S. Department of Health and Human Services; May 2006.

Thomas JW, Grazier KL, Ward K. Comparing accuracy of risk-adjustment methodologies used in economic profiling of physicians. *Inquiry*. 2004;41(2):218–231.

Thomas JW, Ward K. Economic profiling of physician specialists: Use of outlier treatment and episode attribution rules. *Inquiry*. 2006;43(3):271–282.