SPECIAL ARTICLE

# Physician Cost Profiling — Reliability and Risk of Misclassification

John L. Adams, Ph.D., Ateev Mehrotra, M.D., M.P.H., J. William Thomas, Ph.D., and Elizabeth A. McGlynn, Ph.D.


## ABSTRACT

### BACKGROUND

Insurance products with incentives for patients to choose physicians classified as offering lower-cost care on the basis of cost-profiling tools are increasingly common. However, no rigorous evaluation has been undertaken to determine whether these tools can accurately distinguish higher-cost physicians from lower-cost physicians.

### METHODS

We aggregated claims data for the years 2004 and 2005 from four health plans in Massachusetts. We used commercial software to construct clinically homogeneous episodes of care (e.g., treatment of diabetes, heart attack, or urinary tract infection), assigned each episode to a physician, and created a summary profile of resource use (i.e., cost) for each physician on the basis of all assigned episodes. We estimated the reliability (signal-to-noise ratio) of each physician's cost-profile score on a scale of 0 to 1, with 0 indicating that all differences in physicians' cost profiles are due to a lack of precision in the measure (noise) and 1 indicating that all differences are due to real variation in costs of services (signal). We used the reliability results to estimate the proportion of physicians in each specialty whose cost performance would be classified inaccurately in a two-tiered insurance product in which the physicians with cost profiles in the lowest quartile were labeled as "lower cost."

### RESULTS

Median reliabilities ranged from 0.05 for vascular surgery to 0.79 for gastroenterology and otolaryngology. Overall, 59% of physicians had cost-profile scores with reliabilities of less than 0.70, a commonly used marker of suboptimal reliability. Using our reliability results, we estimated that 22% of physicians would be misclassified in a two-tiered system.

### CONCLUSIONS

Current methods for profiling physicians with respect to costs of services may produce misleading results.


From RAND, Santa Monica, CA, and Pittsburgh (J.L.A., A.M., E.A.M.); the University of Pittsburgh School of Medicine, Pittsburgh (A.M.); and the University of Southern Maine, Portland (J.W.T.). Address reprint requests to Dr. Mehrotra at RAND, 4570 Fifth Ave., Suite 600, Pittsburgh, PA 15213, or at mehrotra@rand.org.

N Engl J Med 2010;362:1014-21.

Purchasers of health care are experimenting with a variety of approaches to control costs, several of which involve physicians, since they write the orders that drive spending.[1,2] Prior research suggests that if physicians adopted practices that made less intensive use of resources, health care spending would decrease.[3] Health plans are limiting the number of physicians who receive in-network contracts, offering patients differential copayments to encourage them to visit so-called high-performance physicians (i.e., those providing higher-quality, lower-cost services),[4,5] paying bonuses to physicians whose patterns of resource use are lower than average,[6] and publicly reporting the relative costs of physicians' services.[7] Legislation under consideration in the 111th Congress calls for the use of cost profiling in value-based purchasing strategies.

All 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.[8-12] Validity indicates how well a measure represents the phenomenon of interest, and reliability the proportion of variability in a measure that is due to real differences in performance. The use of episode-grouping tools is accepted as a valid means of constructing clinically homogeneous cost groups.[13,14] With respect to cost profiling, validity indicates whether the method of assigning episodes of care to physicians and creating summary scores accurately represents physicians' economic performance. We previously evaluated the convergent validity of different methods of assigning episodes to physicians[15]; to our knowledge, the reliability of physician cost profiling has not been previously addressed.

The reliability of cost profiles is determined by three factors: the number of observations (i.e., episodes of care), the variation among physicians in their use of resources to manage similar episodes, and random variation in the scores. For cost profiles, reliability is measured at the level of the individual physician because the factors 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 on the basis of the reliability of the physician's profile score.

We evaluated the reliability of current methods of physician cost profiling and analyzed 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 according to specialty.[5,16]

## Methods

### Data Sources and Populations

The data sources and methods used to construct cost profiles are summarized here and described in detail in the Supplementary Appendix, available with the full text of this article at NEJM.org.

Four insurance companies in Massachusetts provided us with all their commercial claims (professional, facility, pharmaceutical, and ancillary) for the calendar years 2004 and 2005, which represented 2.8 million people, or about 44% of the state's residents. We limited the analysis to adults who were at least 18 but less than 65 years old in 2004, who had been continuously enrolled in a plan for 2 years, and who had filed at least one claim (1.1 million persons).

We used a unique identifier from a statewide master directory of physicians created by Massachusetts Health Quality Partners to aggregate data across the four health plans at the physician level.[17] Physicians were included in the study if they provided direct patient care, contracted with one or more of the participating plans, were not in pediatric or geriatric specialties, and had filed at least one claim during the study period. Physicians were assigned to a single specialty on the basis of information from Massachusetts Health Quality Partners. Additional data on physician characteristics were obtained from the Massachusetts Board of Registration in Medicine.

### Construction of Physician Cost Profiles

The process of constructing cost profiles included four basic steps. The first involved grouping claims for services (e.g., office visits, laboratory tests, prescription medications, and other professional services) related to the management of a patient's condition into meaningful clinical categories called episodes. We used commercial software (Episode Treatment Groups, version 6.0, from Symmetry) to create nearly 600 different types of episodes, including preventive services

**Table 1. Elements of an Illustrative Yearlong Episode of Care for a Patient with Type 2 Diabetes.***

| Service | No. of Units | Average Price per Unit | Cost |
|---|---|---|---|
| | | $ | |
| Physician office visit with established patient† | 3 | 100 | 300 |
| Glycated hemoglobin | 2 | 25 | 50 |
| Oral hypoglycemic drug, 1-yr supply | 365 | 1 | 365 |
| Lipid profile | 1 | 35 | 35 |
| Ophthalmology evaluation for dilated-eye examination† | 1 | 250 | 250 |
| Endocrinology consultation† | 1 | 175 | 175 |
| Total observed cost of episode | | | 1,175 |

* The number of units and the costs in the table are illustrative and do not indicate the average reimbursement or the number of services observed in our analysis.
† This item would be included in the cost calculation for professional services.

and care for both chronic diseases and acute conditions. We also used this software to construct patient-specific risk scores based on the patient's mix of episodes, age, and sex. The risk score is used to adjust for differences in expected costs within episodes that reflect the complexity of the patient's condition.

The second step, determining episode costs, involved calculating the average allowed charge across the four health plans for each type of service in each episode (e.g., in Table 1, which lists the components of a yearlong episode of care for a patient with type 2 diabetes, the allowed charge for a glycated hemoglobin test is $25). To calculate the total cost of an episode, we multiplied the unit price for each service by the number of times the service was delivered and summed the costs (which came to $1,175 for the diabetes episode shown in the table). We refer to this total as the observed cost. The observed cost of an episode varies with the number of units of service delivered.

Most cost-profiling applications eliminate extreme values. We did this by setting all charges below the 2.5th percentile and above the 97.5th percentile of the distribution for each service to the values at those cut points, using a process known as Winsorizing.[18,19] We addressed extreme observed episode costs with Winsorizing, using the same cut points.

The third step in the process of constructing physician cost profiles involved assigning each episode to the physician who had the highest proportion of total professional costs and who had billed at least 30% of professional costs. In Table 1, this is the physician who provided three office visits ($300 total for this physician ÷ $725 total professional costs [including $250 for an ophthalmology evaluation and $175 for an endocrinology consultation] = 41%). We were able to assign 52% of episodes; those that could not be assigned to any physician were dropped from the analysis.

For the fourth step, construction of physician summary cost profiles, we calculated the average cost of each episode type assigned to physicians in each specialty (e.g., diabetes episodes assigned to internists) and adjusted the cost using the patient-specific risk score. We refer to this cost as the expected cost. A physician's cost profile is the sum of the observed costs for all assigned episodes divided by the sum of the expected costs for those episodes. The resulting summary cost-profile score is a continuous variable. A value of 1 indicates that a physician's costs are at the average level of costs for his peers, whereas values below or above 1 indicate that a physician's costs are lower or higher, respectively, than those of his peers.

### ANALYSIS OF RELIABILITY

Reliability ranges from 0 to 1; 0 means that all the variation in cost-profile scores is the result of measurement error, and 1 means that all the variation is the result of real differences in performance. High reliability does not mean that the physician's performance is good but rather that one can confidently classify that physician's performance relative to that of other physicians. We calculated reliability at the level of the individual physician using the following formula, where $\sigma^2$ indicates variance[20]:

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

The error variance is specific to a physician and is a function of the number of episodes assigned to the physician, the mix of episodes, and risk adjustment. A physician who had a high proportion of episode types characterized by large variations in cost would have a large physician-specific variation. The details of the standard error calculation are presented in the Supplementary Appendix.

We estimated the physician-to-physician variance ($\sigma^2_{physician\text{-}to\text{-}physician}$) for each specialty with a simple hierarchical linear model.[21] A two-level hierarchical linear model separates the observed variability in physicians' scores into two components: variability of scores among physicians (derived from the distribution of cost profiles within specialty) and variability of scores for individual physicians (derived from the variation in observed costs within an episode type). Physician-to-physician variance is larger in those specialties in which there is a wider distribution of cost-profile scores among the physicians. The physician-to-physician variance is combined with the physician-specific error variance to calculate the physician-specific reliability. We calculated the proportion of physicians whose cost-profile reliabilities were greater than or equal to two commonly used thresholds (0.70 and 0.90) to illustrate some implementation issues.[10,22-25]

### ANALYSIS OF MISCLASSIFICATION

We measured misclassification as the probability that the cost performance of a randomly selected physician in a specialty would be inaccurately categorized. Misclassification rates must be calculated in the context of a specific application. To make the potential problem concrete, we created a two-tiered classification system in which the physicians whose cost profiles were in the lowest 25% of the distribution were labeled as "lower cost." From the physician-specific cost-profile reliabilities calculated above, we estimated the probability of misclassification for each physician. We averaged the misclassification probabilities across all physicians in a specialty to derive the misclassification rates for that specialty. We estimated the proportion of physicians in each specialty who were labeled "lower cost" but were not lower cost, the proportion who were labeled "not lower cost" but were lower cost, and the overall misclassification rate.

### SENSITIVITY ANALYSES

We conducted a number of sensitivity analyses to test the effect of the methods for constructing cost profiles on reliability: one analysis did not have Winsorized extreme values, one used actual reimbursement costs, one involved separate cost profiles for each plan, one used different rules for assigning episodes to physicians, and one restricted profiling to physicians with at least 30 episodes of care for a given condition. We also examined the effect of using different methods of categorizing physicians' performance on misclassification. We used SAS software (version 9.1) for all data preparation and analyses.

## RESULTS

### STUDY SAMPLE

Among the 13,761 physicians in the sample, 12,789 (93%) were assigned at least one episode and were included in the study. The physicians were predominantly men who were board certified, had been trained in the United States, and had been in practice for more than 10 years (Table 2). The median score for summary cost profiles was 0.96, with an interquartile range of 0.80 to 1.17 (for details, see Fig. 3.2 in the Supplementary Appendix).

**Table 2. Characteristics of 12,223 Physicians.***

| Variable | No. (%) |
|---|---|
| Sex | |
| Female | 3,687 (30) |
| Male | 8,536 (70) |
| Board certification | |
| Yes | 11,250 (92) |
| No | 973 (8) |
| Medical school | |
| Domestic | 10,205 (83) |
| International | 2,018 (17) |
| Years in practice† | |
| <10 | 1,951 (16) |
| 10–19 | 3,784 (31) |
| 20–29 | 3,465 (28) |
| 30–39 | 2,099 (17) |
| 40–49 | 777 (6) |
| ≥50 yr | 147 (1) |
| Degree‡ | |
| D.O. | 267/12,210 (2) |
| M.D. | 11,943/12,210 (98) |

* Data are provided for the 12,223 physicians who could be linked to data from the Board of Registration — or 95.6% of the 12,789 physicians in the study population.
† The number of years in practice was defined as the time from the year of medical school graduation to January 1, 2005 (the midpoint of the study period).
‡ Data on type of medical degree were missing for 13 physicians.

### COST-PROFILE RELIABILITY

The results for 10 specialties are reported in this article; the results for 18 additional specialties are available in the Supplementary Appendix. Primary care physicians (i.e., those in family or general practice or internal medicine) made up 32% of the sample, were assigned 46% of attributed episodes, and accounted for 23% of attributed costs. The average number of assigned episodes ranged from 96 for vascular surgery to 383 for family practice. The physician-to-physician standard deviation (for which a higher number means greater variability in actual physician performance) ranged from 0.07 for vascular surgery to 0.36 for cardiology. The median standard error of the profile score (for which a higher number means less precision) ranged from 0.10 for gastroenterology and obstetrics–gynecology to 0.50 for pulmonology. The median reliability of physician cost profiles ranged from 0.05 for vascular surgery to 0.79 for otolaryngology (Table 3). Figure 1 shows that even among physicians with a large number of episodes (e.g., 100), reliability varies widely.

No consensus exists on the level of reliability that is adequate for physician cost-profiling applications. Table 4 shows the proportions of physicians in each specialty with cost-profile reliabilities of 0.70 or more and 0.90 or more. Overall, 41% of physicians had cost profiles with reliabilities greater than or equal to 0.70 (range across specialties, 0 to 62%), and 9% had reliabilities greater than or equal to 0.90 (range, 0 to 21%).

### MISCLASSIFICATION

The overall rate of misclassification ranged from 16% (gastroenterology and otolaryngology) to 36% (vascular surgery). Across the 10 specialties addressed here, the misclassification rate was 22% (Table 5). The proportion of physicians who were classified as lower cost but were not lower cost ranged from 29% (otolaryngology) to 67% (vascular surgery). The proportion of physicians who were not classified as lower cost but who actually were lower cost ranged from 10% (obstetrics–gynecology) to 22% (vascular surgery and internal medicine).

### SENSITIVITY ANALYSES

The results of the sensitivity analyses are presented in detail in the Supplementary Appendix and are summarized here. Retaining extreme unit and episode costs decreased median reliability for 11 specialties and increased median reliability for 7 specialties. Using actual reimbursements rather than average unit costs improved the median reliability for only three specialties, all of which were surgical. If the four health care

**Table 3. Median Reliability of Physician Cost-Profile Scores, According to Specialty.**

| Specialty* | No. of Physicians with at Least One Episode | Total Episodes Assigned† | Average No. of Episodes per Physician | Physician-to-Physician Standard Deviation‡ | Median Standard Error for Cost-Profile Score§ | Median Reliability¶ |
|---|---|---|---|---|---|---|
| | | *no. (%)* | | | | |
| Cardiology | 708 | 73,500 (2.4) | 104 | 0.36 | 0.31 | 0.58 |
| Endocrinology | 169 | 18,070 (0.6) | 107 | 0.16 | 0.21 | 0.37 |
| Family or general practice | 1065 | 408,174 (13.3) | 383 | 0.15 | 0.12 | 0.61 |
| Gastroenterology | 426 | 112,461 (3.7) | 264 | 0.19 | 0.10 | 0.79 |
| Internal medicine | 2979 | 1,008,861 (33.0) | 339 | 0.19 | 0.14 | 0.66 |
| Obstetrics–gynecology | 922 | 299,990 (9.8) | 325 | 0.17 | 0.10 | 0.74 |
| Orthopedic surgery | 580 | 71,156 (2.3) | 123 | 0.16 | 0.21 | 0.36 |
| Otolaryngology | 229 | 60,894 (2.0) | 266 | 0.21 | 0.11 | 0.79 |
| Pulmonary and critical care | 362 | 45,131 (1.5) | 125 | 0.25 | 0.50 | 0.20 |
| Vascular surgery | 72 | 6,879 (0.2) | 96 | 0.07 | 0.31 | 0.05 |

* Physicians were assigned to a single specialty.
† Percentages are based on all 28 specialties included in the study.
‡ The standard deviation is the square root of the physician-to-physician variance.
§ The median of the standard-error distribution for the cost profiles of individual physicians is shown.
¶ Reliability is an attribute of the individual physician's cost profile; the median of the reliabilities of the cost-profile score distribution is shown.

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 2. Requiring physicians to have at least 30 episodes to qualify for inclusion in profiling analyses increased the median reliability for 18 of the 28 specialties but substantially decreased the number of physicians that could be profiled (8689 vs. 12,789). We examined two alternative rules for episode assignment, both of which had lower reliabilities. We also evaluated two alternative profiling applications, both of which had higher rates of misclassification.

## DISCUSSION

We found that the median reliability of physician cost profiles, constructed to reflect typical approaches that insurers use, ranged from 0.05 for vascular surgery to 0.79 for gastroenterology and otolaryngology. Overall, the majority of physicians did not have cost profiles that met common thresholds of reliability. In an illustrative two-tiered classification system, one half of internists and two thirds of vascular surgeons were classified inaccurately as lower cost.

Sample size is one of three factors that determine reliability. We aggregated 2 years of data across four health plans that enrolled about 80% of commercially insured persons in Massachusetts to increase the number of potential episodes assigned to physicians. This strategy increased reliability for three of the four plans but reduced reliability slightly for the fourth. The lower reliability for the fourth plan in the aggregate data set, which resulted from higher physician-specific error estimates, might be seen as a reasonable compromise to make in order to achieve improved reliability in the other plans and to increase the potential for producing consistent scores across all plans.[4,25] Would adding more years of data increase reliability and decrease misclassification? We found that doubling the number of episodes for an average family physician would increase reliability for that physician from 0.61 to 0.76 and decrease his or her probability of misclassification from 17% to 15%. This modest change may not be acceptable because multiyear rolling averages make it difficult to rapidly detect improvements.

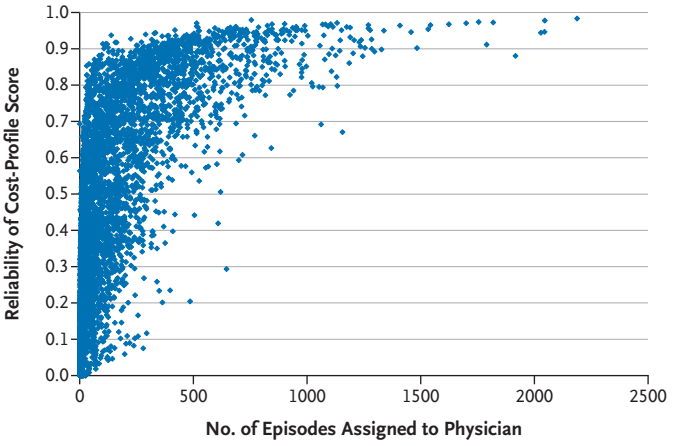


**Figure 1. Relationship between the Number of Episodes Assigned and the Reliability of a Physician's Cost-Profile Score.**
Each data point represents an individual physician.

**Table 4. Proportion of Physicians Whose Cost-Profile Reliability Meets or Exceeds Two Commonly Used Thresholds, According to Specialty.***

| Specialty | Physicians with Cost-Profile Reliability ≥0.70 | Physicians with Cost-Profile Reliability ≥0.90 |
|---|---|---|
| | *percent* | |
| Cardiology | 30 | 4 |
| Endocrinology | 22 | 2 |
| Family or general practice | 41 | 7 |
| Gastroenterology | 59 | 19 |
| Internal medicine | 47 | 13 |
| Obstetrics–gynecology | 57 | 6 |
| Orthopedic surgery | 6 | 0 |
| Otolaryngology | 62 | 21 |
| Pulmonary and critical care | 6 | 1 |
| Vascular surgery | 0 | 0 |
| Overall† | 41 | 9 |

* The reliability thresholds of 0.70 and 0.90 were selected on the basis of thresholds used in previously published literature.
† The numbers shown are for the 10 specialties listed in the table. When reliabilities were calculated across all 28 specialties included in the study, 35% of physicians had reliabilities of 0.70 or more and 9% had reliabilities of 0.90 or more.

On the basis of our findings, we recommend that users of physician cost profiles directly assess reliability instead of relying on proxies of minimum sample size.[7,16,26] This approach will present some implementation challenges. Users will

**Table 5. Misclassification in a Two-Tiered Classification System, According to Specialty.***

| Specialty | No. of Physicians | Physicians Misclassified as Lower Cost† | Physicians Misclassified as Not Lower Cost‡ | Overall Misclassification Rate |
|---|---|---|---|---|
| | | *percent* | | |
| Cardiology | 708 | 40 | 13 | 20 |
| Endocrinology | 169 | 50 | 19 | 25 |
| Family or general practice | 1065 | 39 | 16 | 21 |
| Gastroenterology | 426 | 32 | 11 | 16 |
| Internal medicine | 2979 | 50 | 22 | 25 |
| Obstetrics–gynecology | 922 | 36 | 10 | 17 |
| Orthopedic surgery | 580 | 50 | 17 | 25 |
| Otolaryngology | 229 | 29 | 13 | 16 |
| Pulmonary and critical care | 362 | 58 | 21 | 28 |
| Vascular surgery | 72 | 67 | 22 | 36 |
| Overall§ | 7512 | 43 | 18 | 22 |

* In this two-tiered system, physicians whose cost profiles were in the lowest 25% of the distribution were labeled "lower cost." The remaining 75% of physicians were labeled "not lower cost."
† This is the proportion of physicians who were classified as lower cost but were not lower cost.
‡ This is the proportion of physicians who were classified as not lower cost but who were lower cost.
§ The numbers shown are for the 10 specialties listed in the table. When percentages were calculated across 26 of the 28 specialties included in the study, 43% of physicians were misclassified as lower cost, and 17% were misclassified as not lower cost; overall, 22% of the physicians were misclassified. In two specialties, the reliability was 0, making misclassification impossible to calculate.

need to agree on a minimum acceptable reliability threshold, such as 0.70. Since only a minority of physicians had profiles that met the 0.70 threshold of reliability, users would have to decide how to classify physicians whose scores did not meet the threshold. Physicians with lower-reliability cost profiles could be classified in a lower tier or they could receive a designation indicating that there was not enough information to assess their performance. Since the surgical specialties in particular appear to have low reliability scores, providing incentives for patients to select lower-cost surgical specialists may have little effect in terms of reducing spending. However, physicians with median cost-profile reliabilities greater than or equal to 0.70 accounted for more than half of total costs across the plans, suggesting that opportunities for cost control still exist among physicians with more reliable scores.

The rates of misclassification for the one illustrative application that we examined were large enough to be cause for concern. Among the physicians who were classified as lower cost, 43% were not actually lower-cost performers, which suggests that there are serious threats to insurance plans' abilities to achieve cost-control objectives and to patients' expectations of receiving lower-cost care when they change physicians for that purpose.

Plans may want to consider how they could increase the reliability of cost profiles. Although sample size is a major contributor to reliability, we found that even substantial increases in sample sizes were not adequate to ensure reliability for many specialties. Adding public payers, particularly Medicare, could substantially increase the sample size for some specialties, but because the effects on physician-to-physician variation and on the error variance of the measure are uncertain, reliability might not improve. Episode mix will be difficult to change because it reflects the types of conditions typically managed by physicians in a given specialty. If current efforts to reduce variations in performance are successful, we can expect a decrease in reliability over time. The final option is to develop better measures of cost performance at the physician level. According to our analysis, this is the most promising avenue for further work.

Our study has some limitations in terms of its generalizability. We tested reliability with the use of data from a single state and had access

only to commercial claims. Although Massachusetts is unique in many ways, we believe that the pattern of results observed here is likely to be repeated in other data sets, but such testing should be performed. We tested only one commercially available software product for the purpose of constructing episodes; other tools may produce different results and should be evaluated.

These findings bring into question both the utility of cost-profiling tools for high-stakes uses, such as tiered health plan products, and the likelihood that their use will reduce health care spending. Consumers, physicians, and purchasers are all at risk of being misled by the results produced by these tools.

Supported by a contract from the Department of Labor (J-9-P-2-0033), a career development award from the National Center for Research Resources at the National Institutes of Health (05 KL2 RR024154-04, to Dr. Mehrotra), and a grant from the Robert Wood Johnson Foundation (to Dr. Thomas).

Dr. Thomas reports receiving consulting fees from the Integrated Healthcare Association, the American Medical Association, the American Board of Medical Specialties, and the Arkansas Medical Society; Dr. McGlynn, serving as a paid member of the American Board of Internal Medicine Foundation; and Drs. Adams, Mehrotra, and McGlynn, grant support from the American Medical Association, the Massachusetts Medical Society, the Physicians Advocacy Institute, and the Commonwealth Fund. Ingenix provided a free research license for the use of its commercial programs. No other potential conflict of interest relevant to this article was reported.

We thank Julie Lai for her programming work and Barbra Rabson and Jan Singer of Massachusetts Health Quality Partners, who facilitated our access to the data sets used in this study.

## REFERENCES

1. Sandy LG, Rattray MC, Thomas JW. Episode-based physician profiling: a guide to the perplexing. J Gen Intern Med 2008;23:1521-4.
2. Milstein A, Lee TH. Comparing physicians on efficiency. N Engl J Med 2007;357:2649-52.
3. 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;27:813-23.
4. Draper DA, Liebhaber A, Ginsburg PB. High-performance health plan networks: early experiences: Issue brief no. 111. Washington, DC: Center for Studying Health System Change, 2007. (Accessed February 22, 2010, at http://www.hschange.org/CONTENT/929/.)
5. 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;27:1160-6.
6. Sorbero MES, Damberg CL, Shaw R, et al. Assessment of pay-for-performance options for Medicare physician services: final report. RAND working paper. May 2006. (Accessed February 22, 2010, at http://www.keewu.com/IMG/pdf/18_Sorbrero_et_al__2006__RAND__options_for_P4P_for_physicians.pdf.)
7. Lake T, Colby M, Peterson S. Health plans' use of physician resource use and quality measures. Washington, DC: Medicae Payment Advisory Commission, 2007.
8. Institute of Medicine. Performance measurement: accelerating improvement. Washington, DC: National Academies Press, 2006.
9. 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;281:2098-105.
10. 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;21:13-21.
11. Hofer TP, Bernstein SJ, DeMonner S, Hayward RA. Discussion between reviewers does not improve reliability of peer review of hospital quality. Med Care 2000;38:152-61.
12. Hayward RA, Hofer TP. Estimating hospital deaths due to medical errors: preventability is in the eye of the reviewer. JAMA 2001;286:415-20.
13. Centers for Medicare & Medicaid Services. Medicare resource use measurement plan. (Accessed February 22, 2010, at http://www.cms.hhs.gov/QualityInitiativesGenInfo/downloads/ResourceUse_Roadmap_OEA_1-15_508.pdf.)
14. Measurement framework: evaluating efficiency across patient-focused episodes of care. Washington, DC: National Quality Forum, 2009.
15. Mehrotra A, Adams J, Thomas JW, McGlynn EA. Impact of different attribution rules on individual physician cost profiles. Ann Intern Med (in press).
16. The Leapfrog Group, Bridges to Excellence. Measuring provider efficiency, version 1.0: a collaborative multi-stakeholder effort. 2004. (Accessed February 22, 2010, at http://www.bridgestoexcellence.org/Documents/Measuring_Provider_Efficiency_Version1_12-31-20041.pdf.)
17. Friedberg MW, Coltin KL, Pearson SD, et al. Does affiliation of physician groups with one another produce higher quality primary care? J Gen Intern Med 2007;22:1385-92.
18. Tukey JW. The future of data analysis. Ann Math Stat 1962;33:1-67.
19. Thomas JW, Ward K. Economic profiling of physician specialists: use of outlier treatment and episode attribution rules. Inquiry 2006;43:271-82.
20. Fleiss J, Levin B, Paik M. Statistical methods for rates and proportions. 3rd ed. Hoboken, NJ: Wiley, 2003.
21. Raudenbush S, Bryk A. Hierarchical linear models: applications and data analysis methods. 2nd ed. Newbury Park, CA: Sage, 2002.
22. Hays RD, Revicki D. Reliability and validity (including responsiveness). In: Fayers P, Hays R, eds. Assessing quality of life in clinical trials: methods and practice. 2nd ed. New York: Oxford University Press, 2005:25-39.
23. McDowell I, Newell C. Measuring health, a guide to rating scales and questionnaires. 2nd ed. New York: Oxford University Press, 1996.
24. Scholle SH, Roski J, Adams JL, et al. Benchmarking physician performance: reliability of individual and composite measures. Am J Manag Care 2008;14:833-8.
25. Enhancing physician quality performance measurement and reporting through data aggregation: the Better Quality Information (BQI) to Improve Care for Medicare Beneficiaries Project. 2008. (Accessed February 22, 2010, at http://www.cms.hhs.gov/bqi/.)
26. Standards and guidelines for the certification of physician and hospital quality. Washington, DC: National Committee for Quality Assurance, 2008.