ORIGINAL RESEARCH

Setting a Fair Performance Standard for Physicians' Quality of Patient Care

Brian J. Hess, PhD, Weifeng Weng, PhD, Lorna A. Lynn, MD, Eric S. Holmboe, MD, and Rebecca S. Lipner, PhD

American Board of Internal Medicine, Philadelphia, PA, USA.

BACKGROUND: Assessing physicians' clinical performance using statistically sound, evidence-based measures is challenging. Little research has focused on methodological approaches to setting performance standards to which physicians are being held accountable.

OBJECTIVE: Determine if a rigorous approach for setting an objective, credible standard of minimally-acceptable performance could be used for practicing physicians caring for diabetic patients.

DESIGN: Retrospective cohort study.

PARTICIPANTS: Nine hundred and fifty-seven physicians from the United States with time-limited certification in internal medicine or a subspecialty.

MAIN MEASURES: The ABIM Diabetes Practice Improvement Module was used to collect data on ten clinical and two patient experience measures. A panel of eight internists/subspecialists representing essential perspectives of clinical practice applied an adaptation of the Angoff method to judge how physicians who provide minimally-acceptable care would perform on individual measures to establish performance thresholds. Panelists then rated each measure's relative importance and the Dunn–Rankin method was applied to establish scoring weights for the composite measure. Physician characteristics were used to support the standard-setting outcome.

KEY RESULTS: Physicians abstracted 20,131 patient charts and 18,974 patient surveys were completed. The panel established reasonable performance thresholds and importance weights, yielding a standard of 48.51 (out of 100 possible points) on the composite measure with high classification accuracy (0.98). The 38 (4%) outlier physicians who did not meet the standard had lower ratings of overall clinical competence and professional behavior/attitude from former residency program directors ($p=0.01$ and $p=0.006$, respectively), lower Internal Medicine certification and maintenance of certification examination scores ($p=0.005$ and $p<0.001$, respectively), and primarily worked as solo practitioners ($p=0.02$).

CONCLUSIONS: The standard-setting method yielded a credible, defensible performance standard for diabetes care based on informed judgment that resulted in a reasonable, reproducible outcome. Our method represents one approach to identifying outlier physicians for intervention to protect patients.

KEY WORDS: clinical performance assessment; standard setting; composite measures; diabetes care.

J Gen Intern Med 26(5):467–73

DOI: 10.1007/s11606-010-1572-x

Electronic supplementary material *The online version of this article (doi:10.1007/s11606-010-1572-x) contains supplementary material, which is available to authorized users.*

Received April 19, 2010
Revised September 3, 2010
Accepted October 28, 2010
Published online November 23, 2010

INTRODUCTION

The quality of care provided by physicians in clinical practice is an area of intense public interest. Given the rising burden of healthcare costs, both patients and healthcare purchasers want to know which physicians deliver high quality care. However, no "gold standard" exists for measuring practice performance[1] and attempts at classifying physicians are complicated by heterogeneous practice types and patient panels.[2,3] Methodological challenges to achieving a psychometrically sound physician-level performance assessment are well documented, especially for small office practices without electronic health records.[3] Widespread adoption of clinical performance assessment is likely only if it is meaningful to both patients and physicians, data are feasible to collect, measures are evidence-based and clinically important, and the assessment is psychometrically sound as demonstrated by its reliability and validity.[4,5]

Research has addressed some of these challenges by investigating the fidelity and reliability of composite measures of diabetes care aggregated from evidence-based clinical measures and from patient experience measures.[6,7] Composites are more reliable than individual measures because they reflect a pattern of physician behavior across patients.[6] For public accountability, a credible and fair performance standard, or benchmark, whose outcome is reasonable and defensible must be set. We showed a composite measure of diabetes care yielded high classification accuracy no matter where a standard was set along the score continuum.[8] We believe a standard set on a robust composite measure can be used to evaluate physician practice performance.

A wide variety of methods have been developed to set performance standards.[9] The Angoff method[10] a commonly used content-based procedure, asks experts to estimate how marginally-qualified examinees would perform on individual test questions in multiple-choice examinations.[11] For performance-

based assessments, work-centered approaches using standardized patients are more typical and rely on expert review of examinees' performance on measures of real performance.[12,13] There is little research, however, on standard-setting techniques applied to physician's performance in actual clinical practice.

We present an innovative, multifaceted, rigorous methodology for setting an objective and fair standard for acceptable diabetes care using a composite measure of physician practice performance derived from a set of evidence-based clinical and patient experience measures.

METHOD

Instrument

Since 1990, physicians certified by the American Board of Internal Medicine (ABIM) must recertify every ten years through the maintenance of certification (MOC) program. As part of MOC, physicians conduct a self-assessment of practice performance by completing one of 16 available ABIM Practice Improvement Modules[SM] (PIMs) that focus on improving care of patients with specific disease conditions (e.g., diabetes) or communication skills. PIMs are web-based, self-evaluation tools that use medical chart reviews, patient surveys, and a practice system survey to create a comprehensive performance assessment.[14] We used data from the Diabetes PIM to create an assessment of individual physicians' quality of diabetes care. Physicians were encouraged to abstract 25 patient charts and distribute 25 patient surveys using a retrospective or prospective sequential sample, or a systematic random sample, with a minimum of 10 charts and 10 surveys required. Eligible patients had Type 1 or Type 2 diabetes, were between 18-75 years old, and received care from the practice for at least 12 months (including at least one visit within the past 12 months), with diabetes care management decisions made primarily by that practice. To acknowledge difficulty scheduling appointments, a grace period of one to three months, depending on the recommended interval, was given to all periodic measures (e.g., retinal exams).

Physician Sample

We obtained data from a retrospective cohort of 957 physicians from the United States certified in internal medicine (IM) and/or one of its subspecialties who completed the Diabetes PIM between 2005 and 2007. Our sample of 957 is all of the physicians who completed the Diabetes PIM between 2005 and 2007, and is an 11% subsample of the 9,100 physicians enrolled in MOC who completed any of the 16 available PIMs in those years.

Performance Measures

Table 1 shows the measures from the Diabetes PIM used in the composite. The intermediate outcome and process measures, originally developed by the National Committee for Quality Assurance (NCQA) in partnership with the American Diabetes Association (ADA), use evidence-based guidelines updated annually that describe ideal care for diabetic patients. The A1C-at-goal performance level is based on the new ADA 2009 *Standards of Medical Care in Diabetes*[15] and is more lenient for some patients than earlier ADA standards. Performance on intermediate outcome measures was defined as the percent of a physician's patients that met the recommended level, based on a patient's most recent reading (not an average over prior readings). Performance for process measures was defined as the percent of a physician's patients that received the service. The two patient experience measures, created using specific patient survey questions, were included because they underscore the importance of patient-centered care.[16]

Table 1. Measures from the ABIM Diabetes PIM used to Assess Physician Performance

Measure	Performance level
Process measures	
Retinal exam	Completed
Nephropathy assessment	Completed
Foot exam	Completed
Smoking status & cessation advice / treatment	Completed
Intermediate outcome measures	
A1C poor control	> 9.0%
A1C at goal	< 8.0% or<7.0% * (based on the patient)
Blood pressure poor control	>=140/90
Blood pressure superior control	< 130/80
LDL poor control	>=130 mg/dl
LDL superior control	< 100 mg/dl
Patient experience measures †	
Overall diabetes care satisfaction (1 survey question)	Excellent or very good responses
Patient self-care support (7 survey questions combined)	Excellent or very good responses

A1C, Hemoglobin A1C; LDL, Low-density Lipoprotein; ABIM, American Board of Internal Medicine; PIM, Practice Improvement Module.
** The performance level for A1C at Goal is <8.0% for patients aged 65 and over, with coronary heart disease, cerebrovascular disease, peripheral artery disease, or end-stage renal disease, or significant loss of vision or blindness; the level is <7.0% for other patients.*
† Overall diabetes care satisfaction was defined as the percent of patients in a physician's panel who rated their overall diabetes care "excellent" or "very good" based on one question using a five-point Likert scale. Patient self-care support was defined as the percent of responses per physician that were "excellent" or "very good" across seven questions which included:
-showing understanding of living with diabetes
-encouraging questions and answering them clearly
-providing information on taking medications properly
-providing information on side effects of medications
-teaching foot care
-providing information on proper diabetic diet
-teaching home blood glucose monitoring
Patients who answered "not applicable" or did not answer these questions were excluded from the calculation of this measure.

Standard-Setting Procedure

For each measure, we established a minimum performance threshold for delivering acceptable diabetes care using an adaptation of the Angoff standard-setting method. A panel of eight physicians was selected to represent essential perspec-

tives of clinical practice. All panelists were certified in internal medicine for at least ten years; two were also certified in nephrology, one in endocrinology, and one in geriatric medicine. Four were enrolled in MOC; four were experts in quality improvement. The panel accepted certain limitations including (1) chart data were self-reported, physicians were volunteers, and no external audit was done; (2) differences among practices (e.g., practice size) would not be considered in the deliberations; and (3) no formal risk adjustment for patient case-mix differences was done. The panel deemed the set of performance measures in the composite adequate for assessing the quality of diabetes care. Then the panel listed in detail and agreed upon the characteristics of a hypothetical physician who would provide a minimally-acceptable level of care for diabetic patients (referred to as the "borderline" physician).

After the panel agreed upon a shared understanding of the hypothetical "borderline" physician, they began the process for determining minimum performance thresholds and point values (scoring weights) for each measure. This process (described in detail in Appendix A available online) required panel members to estimate how the "borderline" physician would perform on each measure. For example, each panelist answered "what percent of diabetic patients seen by a borderline physician would receive an annual retinal exam?" To assist in this task, statistics describing patient characteristics from the dataset were presented. After panelists shared their initial estimates, actual results on each measure based on our sample of 957 physicians, along with other available national performance data, were presented as a "reality check." Panelists were then permitted to change their estimates. Final estimates were averaged to represent the minimum performance threshold for each measure. After thresholds were identified, point values (scoring weights) for individual measures were determined using the Dunn–Rankin method[17] which required panelists to independently rate each measure's importance for delivering a minimally-acceptable level of diabetes care using an 11-point scale (0 = *Not at all important* to 10 = *Very important*).

Computing Performance Scores and the Standard

A physician's actual performance rate for each individual measure was multiplied by the assigned point value. Process measures were treated differently than the intermediate outcome and patient experience measures because physicians have more direct control over processes, and so the minimum performance thresholds were used as a lower bound for scoring. Thus, if the percent of a physician's patients receiving a process measure fell below the threshold, then the physician would earn zero points for that measure. Points earned for all measures were summed to yield a total score between 0 and 100 points.

To determine the standard for minimally-acceptable performance, the threshold for each measure was multiplied by the assigned point value (Table 2). For example, the threshold of 28.8% for the retinal exam was multiplied by 9 (i.e., 0.288 × 9= 2.59). The products for all measures were then summed to yield the minimum composite score or "standard" for acceptable diabetic patient care.

Table 2. Descriptive Statistics for Each Performance Measure and Computation of the Standard for Minimally-Acceptable Performance

Measure	Physician performance mean (SD)*	Threshold†	Points¶	Threshold X points
Process measures ‡				
Retinal exam	0.63 (0.29)	0.288	9	2.59
Nephropathy assessment	0.92 (0.12)	0.731	10	7.31
Foot exam	0.69 (0.33)	0.356	4	1.42
Smoking status & cessation advice / treatment	0.97 (0.07)	0.675	7	4.73
Intermediate Outcome Measures				
Not A1C poor control §	0.84 (0.18)	0.725	10	7.25
A1C at goal ‖	0.61 (0.20)	0.360	7	2.52
Not blood pressure poor control §	0.74 (0.16)	0.537	10	5.37
Blood pressure superior control	0.39 (0.17)	0.169	9	1.52
Not LDL poor control §	0.75 (0.19)	0.587	10	5.87
LDL superior control	0.54 (0.20)	0.238	8	1.91
Patient Experience Measures ‖				
Overall diabetes care satisfaction	0.75 (0.16)	0.463	7	3.24
Patient self-care support	0.79 (0.12)	0.531	9	4.78
Standard				Sum=48.51

A1C, Hemoglobin A1C; LDL, Low-density Lipoprotein.

* *The physician performance mean is the average proportion of patients meeting the measure across the sample of 957 physicians.*

† *Threshold is the minimally-acceptable performance rate determined by the panel via the standard-setting exercise.*

‡ *For all process measures, a physician must earn at least the threshold to be awarded any points.*

§ *For the Poor Control measures, points were awarded to physicians when their patients did not meet the performance level noted in Table 1. For example, if 80% of a physician's patients did not have poor control of their blood pressure, then the physician would earn 8 points for that measure (0.80×10=8).*

‖ *See footnote under Table 1 for a description of the A1C at goal measure and the patient experience measures.*

¶ *Points were determined using the Dunn–Rankin method, which required panelists to independently rate each measure's importance for delivering a minimally-acceptable level of diabetes care.*

Estimating Reliability and Classification Accuracy

Reliability refers to the amount of the reported measure that reflects true ability rather than measurement error. As described in Weng et al.[8] we estimated the reliability of the composite using the average observed patient sample size per physician (N=21.0 for chart review; N=19.8 for patient survey). We used a bootstrap sampling method to estimate the standard error (σ^2_{Error}) of measurement from the bootstrap samples for the composite. Reliabilities were estimated using the classical true score model, $\sigma^2_{Observed}=\sigma^2_{True}+\sigma^2_{Error}$,[18] and the reliability of the composite was obtained through Mosier's formula.[19]

In classical test theory, classification accuracy is a measure of the reproducibility of decisions made based on a person's score relative to a criterion (e.g., standard). It is a function of score reliability, the score distribution, the level of the standard, and the proportion of physicians that meet the standard[20]; the higher the accuracy, the fewer false classifications. Classification accuracy does not require a "gold standard" or multiple testing occasions, but rather it is a measure of how often a given physician would meet the standard if the physician sampled a different panel of patients. Therefore, we estimated the accuracy of the acceptable/not acceptable classifications over many different samples of patients using the bootstrap procedure,[8] generating multiple bootstrap samples (or replications) from our dataset, again based on the average patient sample size per physician observed. For each sample, we computed the number of classification decisions and then compared these decisions to the original sample. We calculated the proportion of accurate classifications over all replications for each physician, and then averaged these proportions across physicians to form the classification accuracy index. For more information about classification accuracy measures, see Clauser et al.[20]

Supporting the Standard-Setting Outcome

We used *t*-tests and chi-squared tests to examine differences in demographic characteristics between the physicians who completed the Diabetes PIM and those who completed other ABIM PIMs. We used *t*-tests to test the hypotheses that physicians below the standard would have lower ratings of overall clinical competence and professional behavior/attitude from their residency program directors, and lower first-attempt ABIM IM certification and MOC examination scores. A chi-squared test was used to determine if physicians below the standard tended to work primarily in solo practice.

Statistical analyses were conducted using SAS Version 9.1 software.[21] All data collection was HIPAA compliant; no patient identifying information was obtained and data were reported only in aggregate. Permission to use data for research purposes was granted by physicians upon enrollment in MOC.[22] Essex Institutional Review Board, Inc. approved this study.

Table 3. Demographic Information for the Study Sample and for the Population of Physicians Who Completed Any One of the ABIM PIMs

Characteristics	Study sample (N=957)	Completed any ABIM PIM (N=9100)	P-value*
Mean age (SD)	44.4 (6.18)	44.5 (6.26)	0.39
Female physicians	35%	29%	<.001
Mean % of time spent in an ambulatory setting (SD)	77% (20%)	61% (28%)	<.001
Subspecialty:			<.001
General internal medicine only	81%	50%	
Endocrinology	13%	3%	
Geriatric medicine	4%	4%	
Other	3%	44%	
Practice types:			<.001
Solo physician medical practice	26%	16%	
Group private practice or group/staff model HMO	48%	47%	
Academic faculty practice	6%	13%	
Hospital-owned office-based practice	10%	7%	
Other (e.g., military/ government, nursing homes)	9%	18%	
U.S. Region:			0.50
Northeast	22%	24%	
Midwest	22%	21%	
South	36%	35%	
West	19%	19%	
Outside U.S.	1%	1%	
Mean IM certification exam score – 1st attempt (SD) †	458 (103)	474 (96)	<.001

ABIM, American Board of Internal Medicine; HMO, Health Maintenance Organization; PIMs, Practice Improvement Modules; US, United States; IM, Internal Medicine.

** t-tests were conducted for age, percent of time spent in an office or ambulatory setting, and IM certification exam scores; chi-squared tests for the categorical variables.*

† Scores were statistically equated to be comparable over time and scaled to have a mean=500 and SD=100. Exam scores were available for 902 physicians in our study sample and for 8284 physicians who completed any one of the available ABIM PIMs.

RESULTS

Table 3 presents demographic information for the 957 physicians in our sample compared to the 9,100 physicians who completed any one of the 16 ABIM PIMs. As expected, the sample was comprised of mostly general internists because diabetes is a prevalent chronic disease treated by these physicians. Compared with the larger group, our sample spent more time in an ambulatory setting, contained a higher percentage of female physicians and physicians in solo practice, and scored slightly lower on the initial IM certification examination. We cannot compare our sample with older physicians with time-unlimited certificates and with those physicians not enrolled in MOC.

With regard to patient demographics, the mean number of medical charts abstracted per physician was 21.0 (SD=7.3); overall, physicians abstracted 20,131 charts. Mean patient age was 58.5 (SD=11.1), and 50% were male. The mean number of patient surveys per physician was 19.8 (SD=7.1) for a total of 18,974 patient surveys. The mean age of patients surveyed was 58.6 (SD=11.0); 49% were male. Patient age and gender were similar in the two data sources (medical chart and patient survey).

Mean performance rates (at the physician level) for intermediate outcomes measures were lower than for the process and patient experience measures (Table 2), consistent with findings from previous studies.[7] Based on the bootstrap sampling method, the composite achieved a respectable reliability of 0.91, meaning 91% of the measured performance reflects true ability, not random error. Table 2 presents minimum performance thresholds and number of points assigned to each measure resulting from the standard-setting exercise. The panel's judgment about a "borderline" physician's expected performance on each measure (i.e., thresholds) was well below the mean performance. The

variability (standard deviation) in panelists' estimates for an individual measure was about 10% on average. Based on the Dunn–Rankin method, panelists assigned 54% of the total possible points to intermediate outcome measures, 30% to process measures, and 16% to patient experience measures. The variability in panelists' importance ratings for an individual measure was small, ranging 2 points on average (on the 11-point scale).

The standard for minimally-acceptable performance calculated from the thresholds and point values was 48.51 out of 100 possible points (Table 2). The classification accuracy index for this standard was quite high at 0.98; with repeated sampling from a given physician's patient data the same classification result (acceptable or unacceptable) would occur 98% of the time. The mean composite score was 71.23 (SD=11.90). Appendix B (available online) contains a histogram of the distribution of composite measure scores (total points earned) from our sample of 957 physicians. Using 48.51 as the performance standard resulted in 38 physicians (4%) who were classified as providing unacceptable diabetes care. As a follow-up analysis, we found that physicians' total points earned and the standard did not significantly change if physicians were not required to meet the minimum thresholds to earn points for the process measures (results not shown).

Table 4 shows that the outlier group of 38 physicians had significantly lower overall clinical competence and professional behavior/attitude ratings from former residency program directors, had lower examination scores, and were more likely to work in solo practice.

Table 4. Characteristics of the Group of Physicians that Met the Performance Standard Compared to the Outlier Group of Physicians that Did Not

	Met the Standard (N=919)	Did Not Meet the Standard (N=38)			
Characteristics *	Mean (SD)	Mean (SD)	Test statistic ‖	P-value	Effect size ¶
Overall clinical competence ratings †	6.5 (1.2)	5.9 (1.2)	2.59	0.01	0.45
Professional behavior/ attitude ratings †	7.0 (1.3)	6.4 (1.3)	2.75	0.006	0.40
IM certification exam scores (1st attempt) ‡	459.6 (102.9)	410.9 (91.2)	2.79	0.005	0.49
MOC exam scores (1st attempt) ‡,§	514.8 (91.8)	443.2 (92.2)	4.38	<.001	0.78
Solo practice	29%	47%	5.49	0.02	0.08

IM, Internal Medicine; MOC, Maintenance of Certification.

* *Physician subspecialty, age, gender, birth/training country, or other practice characteristics were not significantly associated with the group of physicians that did not meet the standard.*

† *Program director ratings were based on a nine-point rating scale and were done at the end of the third year of residency training.*

‡ *Scores were statistically equated to be comparable over time and scaled to have a mean=500 and SD=100.*

§ *MOC exam scores were available for 749 physicians in our sample.*

‖ *t-tests were used to assess the significance of each difference, except for solo practice, in which case the chi-squared test was used.*

¶ *The Cohen's d standardized mean difference was used to measure the magnitude of the difference, except for solo practice, in which case the phi coefficient was used.*

DISCUSSION

Our study adds to the literature on performance measurement in that we evaluated *individual* physicians' diabetes care by creating a robust composite performance measure (using evidence-based clinical data combined with patient-experience data) and applying a rigorous standard-setting method on that composite. We show that performance measurement is strengthened when a composite is used instead of individual measures through increased reliability and classification accuracy which yields valid results. The inclusion of patient experience data captures the patient's "voice" in the assessment of physicians' care. Additionally, we use a credible standard-setting methodology to set an absolute standard based on informed judgment, carried out with due diligence, and supported by data and research.[23] As a standard-setting organization, ABIM must ensure that any decisions it makes about an individual physician based on performance must be reliable and valid.

During the standard-setting exercise, panelists incorporated their own patient-care experience into their decisions. We heard some spirited debate over the respective responsibilities of physician and patient, and the need for "grace periods" due to difficulties some patients have in keeping appointments. The method yielded a hurdle which most board-certified physicians easily met. Because the panel set a standard for *minimum acceptable* performance for previously certified physicians and classification accuracy was very high (98%), the result is defensible as a mechanism to identify outlier physicians in need of quality improvement. Furthermore, the process was conducted by peers, the performance data were for a condition of the physician's choosing and were self-collected (allowing for exclusion of patients for whom diabetes control was irrelevant, including the terminally ill), the measures used in the composite were weighted by importance to patient health, and the measurement of glucose control (A1C at goal) was adjusted to reflect characteristics of patients in a physician's sample.

The outlier physicians had distinct and predictable characteristics. They had lower examination scores, consistent with the literature on the relationship between certification examination performance and quality of care.[24–26] They were judged to have lower overall clinical competence and poorer professional behavior/attitudes at the end of residency training, and frequently worked as solo practitioners. Furthermore, they tended to perform consistently low on most measures, not compensating for poor performance on intermediate outcome measures through superior performance on process measures.

Our approach to setting standards on a composite performance measure serves a variety of purposes. A vital role for assessment organizations is to assure the public that physicians who care for them are delivering care of acceptable quality. Our methodology represents one approach to identifying outlier physicians for intervention to protect patients. For example, physicians who do not meet the performance standard might be required to complete focused training and/or self-assessment activities that lead to improved

practice performance to maintain certification. In the future, ABIM anticipates that the self-reported chart abstraction will be replaced with an automated process, such as a "middleware" data solution. With the physician's permission, performance measures would be abstracted directly from electronic medical health records, reducing the burden of data collection. In addition, our methodology could be used to set standards for other defined levels of care (e.g., excellent care). Information about a physician's performance could be used to support health care choices by patients and purchasers, or to reward physicians for the care they provide.

There are limitations to this research. First, rather than using a subjective rating scale, the relative importance of each measure might be better judged by its potential impact on quality-adjusted life years, as with the pathophysiologically-based *Archimedes* mathematical model for diabetes.[27] Second, the viability and generalizability of our method for setting practice-based performance standards for other disease conditions need to be explored. Third, at the time of the study we were not able to adequately account for patient case-mix and physician-level clustering in assessing physicians' quality of care.[28] However, in our study context, patient-case mix is less relevant for process measures and the A1C-at-goal measure by definition accounts for patient age and particular disease conditions. We believe that for the 38 outlier physicians, no level of patient case-mix adjustment could justify their poor performance, especially since physicians selected the patients. Fourth, physicians selected the Diabetes PIM to satisfy the practice performance requirement of MOC, limiting the generalizability of the results to other physicians caring for diabetic patients. Fifth, the sampling strategy and accuracy of chart data were not audited (there was no penalty for poor performance on the PIM); however previous work has confirmed the accuracy of physician-reported data in the Diabetes PIM.[29] It is possible that some physicians may not have adhered to the sampling instructions and their performance on individual measures might be inflated.

CONCLUSION

An effective means for assessing physician performance in clinical practice is needed to assure the public that their physician is delivering acceptable patient care in a particular area, like diabetes. Our approach to measuring performance and setting a standard for diabetes care provides a fair and objective means to identify physicians for a variety of purposes, such as accountability and recognition (pay-for-performance) programs. Ultimately, setting a standard across conditions is desirable, but sampling and other methodological challenges presently make this task quite complex.[30] We focused on a specific disease (diabetes) rather than assessing quality of care across conditions because diabetes is highly prevalent and exacts a tremendous toll on patients and the health care system. More research needs to be done before a defensible and meaningful standard could be used in a high-stakes assessment within MOC programs. Future research could examine the applicability of this standard-setting method in training programs, its generalizability across different specialties, and as a comprehensive assessment in the primary care setting.

Contributors: *We thank the standard-setting panel, and Dr. Gerald Arnold and Leslie Tucker for their help with the scoring strategy and manuscript preparation, respectively.*

Funders: *The ABIM Foundation funded this study.*

Prior Presentations: *An abstract of this study was presented at the annual AcademyHealth meeting on 28 June 2009.*

Conflict of Interest: *All authors are employed by the ABIM. Drs. Hess, Weng, and Lipner are co-inventors of a business method invention describing the application of the standard-setting method to practicing physicians. The invention is patent pending. Dr. Holmboe received honoraria for teaching about clinical assessment from the Uniformed Services University of the Health Sciences, the University of Kansas, and the Harvard-Macy Systems Assessment Course. Dr. Holmboe receives royalties for a textbook on assessment published by Mosby-Elsevier.*

Corresponding Author: *Brian J. Hess, PhD; American Board of Internal Medicine, 510 Walnut Street, Suite 1700, Philadelphia, PA 19106, USA (e-mail: bhess@abim.org).*

REFERENCES

1. **Miller TP, Brennan TA, Milstein A.** How can we make more progress in measuring physicians' performance to improve the value of care? Health Aff. 2009;28:1429–1437.
2. **Holmboe ES.** Assessment of the practicing physician: Challenges and opportunities. J Contin Educ Health Prof. 2008;28(Suppl 1):4–10.
3. **Landon BE, Normand S-LT.** Performance measurement in the small office practice: Challenges and potential solutions. Ann Intern Med. 2008;148:353–357.
4. **Landon BE, Normand S-LT, Blumenthal D, Daley J.** Physician clinical performance assessment: Prospects and barriers. JAMA. 2003; 290:1183–1189.
5. **Scholle SH, Pawlson LG, Solberg LI, et al.** Measuring practice systems for chronic illness care: Accuracy of self-reports from clinical personnel. Jt Comm J Qual Patient Saf. 2008;34:407–416.
6. **Kaplan SH, Griffith JL, Price LL, Pawlson LG, Greenfield S.** Improving the reliability of physician performance assessment: Identifying the "physician effect" on quality and creating composite measures. Med Care. 2009;47:378–387.
7. **Lipner RS, Weng W, Arnold GK, Duffy FD, Lynn LA, Holmboe ES.** A three-part model for measuring diabetes care in physician practice. Acad Med. 2007;82(Suppl 10):S48–S52.
8. **Weng W, Hess BJ, Lynn LA, Holmboe ES, Lipner RS.** Measuring physicians' performance in clinical practice: Reliability, classification accuracy, and validity. Eval Health Prof. 2010;33:302–320.
9. **Cizek GJ.** Setting Performance Standards: Concepts, Methods, and Perspectives. Mahwah, NJ: Lawrence Erlbaum; 2001.
10. **Angoff WH.** Scales, norms, and equivalent scores. In: **Thorndike RL, ed.** Educational Measurement. American Council on Education: Washington, DC; 1971:514–515.
11. **Downing SA, Tekian A, Yudkowsky R.** Procedures for establishing defensible absolute passing scores on performance examinations in health professions education. Teach Learn Med. 2006;18:50–57.
12. **Boulet JR, De Champlain AF, McKinley DW.** Setting defensible performance standards on OSCEs and standardized patient examinations. Med Teach. 2003;25:245–249.
13. **McKinley DW, Boulet JR, Hambleton RK.** A work-centered approach for setting passing scores on performance-based assessments. Eval Health Prof. 2005;28:349–369.
14. **Duffy FD, Lynn LA, Didura H, et al.** Self-assessment of practice performance: Development of the ABIM Practice Improvement Module (PIM[SM]). J Contin Educ Health Prof. 2008;28:38–46.
15. American Diabetes Association. Standards of medical care in diabetes-2009. Diab Care. 2009;32(Suppl 1):S13–S61.
16. **Rittenhouse DR, Shortell SM.** The patient-centered medical home. JAMA. 2009;301:2038–2040.

17. **Dunn–Rankin P.** Scaling Methods. Hillsdale, NY: Erlbaum; 1983.
18. **Reeves D, Campbell SM, Adams J, Shekelle PG, Kontopantelis E, Roland MO.** Combining multiple indicators of clinical quality: An evaluation of different analytic approaches. Med Care. 2007;45:489–496.
19. **Mosier C.** On the reliability of a weighted composite. Psychometrika. 1943;8:161–168.
20. **Clauser BE, Margolis MJ, Case SM.** Testing for licensure and certification in the professions. In: **Brennan RL, ed.** Educational Measurement. 4th ed. Westport, CT: Praeger Publishers; 2006:701–731.
21. SAS Institute. Statistical Analysis System. Version 9.1. Cary, NC: SAS Institute; 2002.
22. American Board of Internal Medicine. ABIM HIPAA Business Associate Agreement. Available at: http://www.abim.org/pdf/hipaa/hipaa_compliance.pdf. Accessed October 19, 2010.
23. **Norcini JJ, Shea JA.** The credibility and comparability of standards. Appl Meas Ed. 1997;10:39–59.
24. **Holmboe ES, Wang Y, Meehan TP, et al.** Association between maintenance of certification examination scores and quality of care for Medicare beneficiaries. Arch Intern Med. 2008;168:1396–1403.
25. **Norcini JJ, Lipner RS, Kimball HR.** Certifying examination performance and patient outcomes following acute myocardial infarction. Med Educ. 2002;36:853–859.
26. **Tamblyn R, Abrahamowicz M, Dauphinee WD, et al.** Association between licensure examination scores and practice in primary care. JAMA. 2002;288:3019–3026.
27. **Eddy DM, Schlessinger L.** Validation of the Archimedes diabetes model. Diab Care. 2003;26:3102–3110.
28. **Greenfield S, Kaplan SH, Kahn R, Nimomiya J, Griffith JL.** Profiling care provided by different groups of physicians: Effects of patient case-mix (bias) and physician-level clustering on quality assessment results. Ann Intern Med. 2002;136:111–121.
29. **Holmboe ES, Meehan TP, Lynn L, Doyle P, Sherwin T, Duffy FD.** Promoting physicians' self-assessment and quality improvement: The ABIM Diabetes Practice Improvement Module. J Contin Ed Health Prof. 2006;26:109–119.
30. **Holmboe ES, Weng W, Arnold GF, et al.** The comprehensive care project: Measuring physician performance in ambulatory practice. Health Serv Res. 2010;45:1912–1933.