

Physician performance assessment: prevention of cardiovascular disease

Rebecca S. Lipner · Weifeng Weng · Kelly J. Caverzagie · Brian J. Hess

Received: 16 August 2012 / Accepted: 23 January 2013 / Published online: 16 February 2013
© Springer Science+Business Media Dordrecht 2013

Abstract Given the rising burden of healthcare costs, both patients and healthcare purchasers are interested in discerning which physicians deliver quality care. We proposed a methodology to assess physician clinical performance in preventive cardiology care, and determined a benchmark for minimally acceptable performance. We used data on eight evidence-based clinical measures from 811 physicians that completed the American Board of Internal Medicine's Preventive Cardiology Practice Improvement ModuleSM to form an overall composite score for preventive cardiology care. An expert panel of nine internists/cardiologists skilled in preventive care for cardiovascular disease used an adaptation of the Angoff standard-setting method and the Dunn-Rankin method to create the composite and establish a standard. Physician characteristics were used to examine the validity of the inferences made from the composite scores. The mean composite score was 73.88 % (SD = 11.88 %). Reliability of the composite was high at 0.87. Specialized cardiologists had significantly lower composite scores ($P = 0.04$), while physicians who reported spending more time in primary, longitudinal, and preventive consultative care had significantly higher scores ($P = 0.01$), providing some evidence of score validity. The panel established a standard of 47.38 % on the composite measure with high classification accuracy (0.98). Only 2.7 % of the physicians performed below the standard for minimally acceptable preventive cardiovascular disease care. Of those, 64 % ($N = 14$) were not general cardiologists. Our study presents a psychometrically defensible methodology for assessing physician performance in preventive cardiology while also providing relative feedback with the hope of heightening physician awareness about deficits and improving patient care.

This study was presented at the annual meeting of the American Educational Research Association on April 14, 2012.

R. S. Lipner (✉) · W. Weng · B. J. Hess
American Board of Internal Medicine, 510 Walnut Street, Suite 1700, Philadelphia, PA 19106, USA
e-mail: rlipner@abim.org

K. J. Caverzagie
Division of General Internal Medicine, Department of Internal Medicine, University of Nebraska
Medical Center, 985185 Nebraska Medical Center, Omaha, NE 68198-5185, USA

Keywords Cardiovascular disease · Performance assessment · Prevention · Standard of care

Introduction

Given the rising burden of healthcare costs, both patients and healthcare purchasers are interested in discerning which physicians deliver high quality care as evidenced by the Patient Protection and Affordable Care Act of 2010 and the physician compare website (Centers for Medicare & Medicaid Services 2013). Widespread adoption of clinical performance assessment is likely only if it is meaningful, measures are evidence-based and clinically important, performance feedback is useful, data collection is feasible, and the assessment is psychometrically robust (Landon et al. 2003; Scholle et al. 2008). Research has addressed some of these challenges by investigating the validity and reliability of composite measures aggregated from evidence-based measures (Kaplan et al. 2009; Lipner et al. 2007; Weng et al. 2010; Hess et al. 2011). For public accountability, a credible and fair performance benchmark, whose outcome is reasonable and defensible, must be determined. Previous research has shown that a reasonable benchmark could be set for a composite measure of diabetes care using a unique methodology (Hess et al. 2011). A composite measure has been shown to present a more comprehensive picture of a physician's quality of care, is fairer, and more reliable than individual clinical measures (Weng et al. 2010).

A reliable and defensible composite measure for the quality of preventive care for cardiovascular disease, the leading cause of death in the US (Centers for Disease Control & Prevention 2013a), has not yet been developed. Because composite scores are more reliable than individual measures, physicians would be more likely to use the data to examine their performance relative to others and use the feedback to help improve their adherence to clinical practice guidelines for cardiovascular disease and stroke prevention (Ornstein et al. 2004).

The purpose of this study was to apply our previously developed methodology to examine performance of physicians in preventive cardiology care, evaluate the measurement characteristics of a practice performance composite score derived from a set of evidence-based clinical measures, and determine whether a reasonable benchmark for performance on this composite could be established. A secondary purpose was to determine the type of feedback to provide to physicians to help identify areas in which to improve the quality of care.

Methods

Instrument

We used data from the American Board of Internal Medicine (ABIM) Preventive Cardiology Practice Improvement ModuleSM (PIM) to create a composite of physicians' quality of preventive cardiology care. This PIM is a web-based, self-evaluation tool that uses medical chart reviews, patient surveys, and a practice system survey to create a comprehensive performance assessment. Since 1990, physicians certified by the ABIM must recertify every 10 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 PIMs that focus on improving care of patients with specific disease conditions (e.g., cardiovascular disease).

Physicians abstracted 25 patient charts using a retrospective or prospective sequential sample, or a random sample. Eligible patients were between 20 and 90 years old, received care from the practice for at least 12 months (at least one visit within the past 12 months) regarding the prevention of coronary heart disease. Due to common patient scheduling difficulties, a grace period of 1–3 months, depending on the recommended interval, was given to all periodic clinical measures (e.g., diabetes screening test).

Physician sample

We obtained data from a retrospective cohort of 811 physicians from the US certified in internal medicine (IM) and/or one of its subspecialties who elected to complete the Preventive Cardiology PIM between 2008 and 2011, and is a 4 % subsample of the 22,045 physicians enrolled in MOC who completed any of the 16 available PIMs in those years.

Performance measures

The evidence-based measures used to develop the composite include the following two intermediate outcome measures: blood pressure control and low-density lipoprotein (LDL) control. We also include the following six process measures: lipid testing, diabetes documentation or screening, 10 year risk calculation for coronary death or myocardial infarction, dietary and physical activity modifications prescribed, aspirin or other anti-platelet/anticoagulant therapy prescribed, smoking status and cessation support. Our focus on process and intermediate outcome measures is consistent with Donabedian's theoretical framework for assessing quality of patient care (Donabedian 1988). While practice structure is another component in Donabedian's framework, we did not include structural measures (e.g., use of electronic medical records) in the composite because their evidence base is still unclear (Duffy et al. 2008; Holmboe et al. 2010). The eight measures were specifically derived from the National Committee on Quality Assurance Health Employers Data Information Set, the AMA Physician Consortium for Performance Improvement measures, and from national guidelines established by the Joint Commission. [Appendix](#) presents a detailed description of each measure. Control of lipids and blood pressure and improvement in nutrition and fitness are key measures in mitigating cardiovascular disease. Performance on intermediate outcome measures was defined as the percent of a physician's patients that were at goal, based on a patient's most recent reading. Performance for process measures was defined as the percent of a physician's patients that received the service.

Composite and standard-setting methodology

The methodology used to develop the composite measure is previously described (Weng et al. 2010; Hess et al. 2011). For each of the eight measures, we established a threshold for delivering minimally-acceptable preventive cardiovascular care using an expert panel and an adaptation of the Angoff standard-setting method (Angoff 1971). The Angoff method is a well-established and accepted standard-setting approach with multiple-choice tests, and we sought to adapt it to bring additional rigor to the process of establishing a minimum

passing score for clinical measures of physician performance in practice. A panel of nine physicians was selected through a call for nominations to represent essential perspectives of clinical practice, specifically around cardiovascular care. All panelists were board certified in internal medicine, cardiology, or both. A majority spent at least 70 % of their time in clinical care delivery in relevant areas. As part of the process, the panel accepted the following limitations: (1) physicians were volunteers who self-reported their chart data without an external audit; (2) no adjustment would be applied for practice differences, like the existence of an electronic health record; and (3) although risk adjustments were made for co-morbidities in the measure calculations, no additional adjustment was done to control for patient case-mix (e.g., we did not control for the fact that race/ethnicity of a patient panel varies by physician).

The concept of a “borderline” physician—a physician who would provide a *minimally-acceptable level* of preventive cardiovascular disease care—was discussed. A shared understanding about the behavior of a hypothetical “borderline” physician (Table 1) was derived which included behaviors that were inconsistent and unreliable, denoting subpar patient care but not the worst care.

Thresholds and point values for *each* measure were then considered. Panelists estimated how the “borderline” physician would perform on each measure. For example, each panelist answered “what percent of preventive cardiovascular patients seen by a borderline physician would have their blood pressure at goal?” Statistics describing patient characteristics were presented to assist the panel; this included patients’ mean (SD) age, percentage of female patients, percentage of patients from different racial groups, percentage of patients with known coronary heart disease (CHD), the mean (SD) number of CHD risk factors, and the percentage of patients with certain factors that limited their self-care (e.g., cognitive impairment, other medical condition). After panelists shared their initial estimates with each other as a group, actual results for each measure based on our sample of 811 physicians, and other available national performance data, were presented as a “reality check.” Panelists had a discussion about their differences, and as in most standard-setting processes, were allowed to change their estimates; this was especially important if the

Table 1 Behavioral characteristics of the “Borderline” physician

Family history not taken or not up to date
No reliable follow up (e.g., abnormal lab results, blood pressure or weight (BMI))
Achieves <25 % of patients at LDL goal
<50 % of patients counseled on smoking cessation
Does not evaluate extended family risk of patients with cardiovascular risk
Overuse of testing
Poor record keeping
Gaps of knowledge in guidelines
Incomplete physical exam
Inconsistent medication choices
Not accessible to either patients or colleagues
Not appropriate referral (both excessive and delayed)
Fails to consistently engage patients in shared decision making
Fails to consistently include patients’ entire clinical care picture
Fails to consistently coordinate clinical care including social environment
Limited self-directed evidence-based learning

difference in initial estimates between any two experts was 20 % or greater. The panel's discussion often focused on the fairness of attributing measurement results to physician actions. Final estimates were averaged across the panelists to determine the threshold for each measure. The Dunn-Rankin method (Dunn-Rankin 1983)—a method of determining relative importance among the measures that counterbalances the order in which measures are presented to the panelist—was then applied to determine point values for individual measures. This required panelists to first independently rate each measure in terms of its importance to delivering a minimally-acceptable level of preventive care using an 11-point Likert scale with 0 = Not at all important and 10 = Very important. Then these ratings were used to calculate the point value for each measure, by incorporating both the average importance rating and maximum rating provided by each panelist so that the sum of the points ranged from 0 to 100.

To determine the standard for minimally-acceptable composite performance, the threshold for each measure was multiplied by the assigned point value (see Table 2). For example, the threshold of 32.10 % for blood pressure at goal was multiplied by 24 (i.e., $0.321 \times 24 = 7.71$). The products for all measures were then summed to yield the minimum composite score or “standard/benchmark” for minimally-acceptable patient care in preventive cardiology.

Table 2 Computation of the level for minimally-acceptable performance

Measure	Actual physician performance mean (SD) ^a	Threshold (%)	Points	Threshold × points
Process measures ^b				
Timing of lipid testing complies with guidelines	84 % (19 %)	52.50	9	4.73
Diabetes documentation or screening test	87 % (22 %)	51.90	9	4.67
Correct determination of 10 year risk for coronary death or MI	40 % (25 %)	16.30	3	0.49
Dietary and physical activity modifications appropriately prescribed	88 % (19 %)	60.60	10	6.06
Appropriate use of aspirin or other anti-platelet/anti-coagulant therapy	89 % (13 %)	63.80	8	5.10
Smoking status and cessation support	97 % (8 %)	82.10	11	9.03
Intermediate outcome measures				
Blood pressure at goal	64 % (15 %)	32.10	24	7.71
LDL cholesterol at goal	64 % (20 %)	36.90	26	9.59
Standard/benchmark				sum = 47.38

^a The physician performance mean is the average proportion of patients meeting the measure across the sample of 811 physicians

^b For all process measures, a physician must earn at least the threshold to be awarded any points

Computing a physician's performance score

A physician's actual performance rate for each measure was multiplied by its assigned point value. For example, if 50 % of a physician's patients had their blood pressure at goal, the physician would receive 12 points for the measure ($0.50 \times 24 = 12$). Process measures were computed slightly differently than intermediate outcome measures because physicians have more direct control over them. Products were computed, however, no points were awarded for process measures unless the threshold for the individual measure was met (e.g., no points were awarded for Timing of Lipid Testing unless the physician met the minimum threshold of 52.50 %). Points earned by the physician for all measures were summed to yield a composite score between 0 and 100 points.

Estimating reliability and classification accuracy

As previously described in Weng et al. (2010), we used the bootstrap sampling method to estimate the reliability of the individual clinical measures and the composite measure scores (Reeves et al. 2007), as well as the accuracy of classifications (Lee et al. 2009) based on different benchmarks set along the continuum of the composite measure scale. Reliability is important in that it measures how much of what you have measured is true and how much is based on errors of measurement. We estimated reliability using the bootstrap sampling method and an average observed patient sample size per physician of 25.50. The bootstrap sampling method is distinctly different from Cronbach's α which is typically used to assess the internal consistency reliability of measure scores. It is more like the generalizability coefficient because it takes into account the effect of nesting patients within physicians, but it also must consider the use of a lower bound for scoring the process measures and measures with different underlying distributions. Each of the bootstrap samples is obtained by random sampling with replacement from the original chart review. We estimated the standard error (σ_{Error}^2) of measurement through the bootstrap samples for individual clinical measures and the composite measure overall. The reliabilities were estimated using the classical true score model, $\sigma_{\text{Observed}}^2 = \sigma_{\text{True}}^2 + \sigma_{\text{Error}}^2$ (Brennan 2006). The Spearman-Brown formula was then applied to derive the intraclass correlation coefficient (ICC) based on the average patient sample size per physician observed (i.e., $N = 25.5$) so that we would be able to compare reliabilities across the measures. The reliability of the composite measure scores was obtained through Mosier's formula (Mosier 1943).

A measure of classification accuracy was also obtained to better understand whether the standard was able to make accurate pass/fail judgments about physicians and minimize false classifications. We used the same bootstrap samples (Lee et al. 2009) to measure classification accuracy (range 0–1). The higher the value, the fewer false classifications will be made. For a given benchmark or standard, the classification was first made for each physician using the original data. Second, a composite score was calculated for each physician's bootstrap sample and a classification was made using this calculation. If the classification decisions were the same for both the bootstrap sample and original sample, the decision was deemed accurate. Finally, the proportion of accurate classification over all samples for each physician was calculated. These proportions were averaged across physicians to form the overall classification accuracy. The overall classification accuracy was calculated for benchmarks along the composite measure scale continuum with 5 point increments. We also calculated the bootstrap 95 % confidence interval for each composite measure score.

Statistical analyses

We used *t* tests and Chi squared tests to examine differences in demographic characteristics between the physicians who completed the Preventive Cardiology PIM and those who completed other ABIM PIMs. We computed an intraclass coefficient to assess the inter-rater agreement in panelists' final estimates of the performance thresholds for the measures resulting from the Angoff exercise. We used multivariate regression analysis to determine if physician characteristics were meaningfully associated with the composite scores. We focused mainly on determining if the type of cardiologist (i.e., specialized versus a general cardiologist who would be more likely to be concerned with prevention) and amount of time that physicians provide direct care to patients were associated with the composite scores, controlling for other physician and patient characteristics. Standardized regression coefficients (β s) were used to measure the effect size for each predictor variable in the model. We also used physician characteristics to compare those physicians who did not meet the standard with those who did meet the standard. We used results from prior work on cognitive testing to determine appropriate feedback reports (Weng et al. 2010; Hess et al. 2011). Statistical analyses were conducted using SAS Version 9.1 software. All data collection was HIPAA compliant; no patient identifying information was obtained and data were reported only in aggregate.

Results

Table 3 presents demographic information for the 811 physicians in our sample compared to the 22,045 physicians who completed any one of the 16 available ABIM PIMs during the same time period. As expected, the sample was comprised of mostly cardiologists because the Preventive Cardiology PIM is most relevant to these physicians. Compared with the larger population, our sample spent slightly less time in an ambulatory setting, contained a lower percentage of female physicians, a higher percentage of physicians in group practice, and a slightly higher percentage from the northeast and southern US region. We cannot compare our sample with physicians not enrolled in MOC.

With regard to patient demographics, the mean number of medical charts abstracted per physician from the PIM was 25.50 (SD = 1.80); overall, physicians abstracted 20,656 charts. The mean age of the patients was 65.20 years (SD = 13.0); 41 % were female; and 70 % had coronary heart disease (CHD) or CHD risk equivalent.

Table 2 shows physicians' average performance on the eight measures. Physicians did not perform as well on intermediate outcomes as on the process measures. These findings are consistent with a previous study on diabetes measures (Lipner et al. 2007). The greatest room for improvement was on the correct determination of ten-year risk for coronary death or myocardial infarction (Mean = 40 %, SD = 25 %). The composite achieved a very respectable reliability of 0.87, meaning 87 % of the measured performance reflects true ability, not random error. The mean composite score was 73.88 (SD = 11.88) out of a possible 100 points.

Table 2 also presents minimum performance thresholds and number of points assigned to each measure resulting from the standard-setting exercise. The panel had discussion regarding the weighting of the points for each measure, and chose to assign a higher point value to those measures that have the greatest potential impact at recognizing the risk of and/or preventing cardiovascular disease (e.g., smoking status and cessation support). Despite having the greatest room for improvement, the panel assigned a relatively low

Table 3 Demographic information for the study sample and for the population of physicians who completed any one of the ABIM practice improvement modules

Characteristics	Study sample (N = 811)	Population (N = 22,045)	<i>P</i> value ^a
Mean age (SD)	47.6 (6.6)	46.0 (6.7)	<0.001
Female physicians (%)	15 %	31 %	<0.001
Mean % of time spent in an office or ambulatory Setting (SD)	52 % (25 %)	59 % (31 %)	<0.001
Mean equated IM certification exam score (SD) ^b	491 (89)	489 (89)	0.55
Subspecialties			<0.001
General internal medicine only (%)	20 %	44 %	
General cardiologists (%)	42 %	8 %	
Specialized cardiologists (%)	36 %	9 %	
Other subspecialists (%)	3 %	40 %	
US/Canada medical school trained (%)	62 %	61 %	0.55
Practice types			<0.001
Solo physician medical practice (%)	12 %	13 %	
Group private practice or group/staff HMO (%)	66 %	47 %	
Academic faculty practice (%)	9 %	16 %	
Hospital-owned office-based practice (%)	9 %	9 %	
Other (e.g. military/government, public clinics, nursing homes)	3 %	16 %	
Region			<0.001
Northeast	28 %	24 %	
Midwest	20 %	22 %	
South	36 %	33 %	
West	15 %	21 %	

^a *t* tests were used to compare the groups on age, % of time spent in an office or ambulatory setting, equated IM certification exam score; Chi square tests were used for the other demographic variables

^b Equated IM certification exam score were from the physicians' first attempt and were statistically equated to be comparable over time and scaled to have a mean of 500 and SD of 100

point value to correctly determining ten-year risk of coronary death or myocardial infarction because of the widely varied application of this calculation, particularly amongst practicing cardiologists whose expertise at understanding risk may extend beyond this straightforward screening method. As expected, the panel's judgment about a "borderline" physician's expected performance on each measure was well below the mean performance. For the final estimates, the average difference between panelists for an individual measure was about 10 %; the inter-rater agreement across the eight measures was 0.71, thus indicating consistency in the panelists' judgments. The final estimates were lower than the initial for two measures, and higher for the rest of the measures. On average, the absolute value of the difference between the initial and final estimates was 10 %. Based on the Dunn-Rankin method, the variability in panelists' importance ratings for an individual measure was small, ranging two points on average (on the 11-point scale).

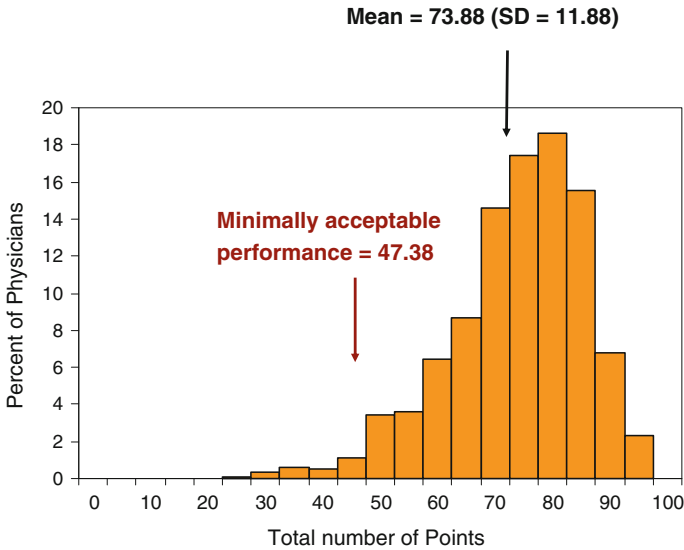


Fig. 1 Distribution of composite measure scores with the benchmark for minimally-acceptable performance (N = 811 Physicians)

The standard for minimally-acceptable performance calculated from the thresholds and point values was 47.38 out of a possible 100 points (Table 2). Figure 1 contains a histogram of the distribution of composite scores (total points earned) from our sample of 811 physicians along with the minimally acceptable performance level. Only 22 physicians (2.70 %) fell below the benchmark of 47.38 % and were classified as providing unacceptable preventive cardiovascular disease care. This is consistent with the standard observed for diabetes care (Hess et al. 2011). The classification accuracy index at 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. Of those who did not achieve the benchmark (N = 14), 64 % were not general cardiologists. Although these numbers are quite small they do reflect our expectations that general cardiologists would perform the best in this area, followed by the specialized cardiologists and general internists.

Table 4 presents the association between physician characteristics and the composite scores (excluding 116 physicians because they did not have equated IM certification exam scores since they took the exam prior to 1986 when equated scores were not available). The adjusted model R^2 was 0.08 which indicates that only 8 % of the variance in composite scores was explained by the set of physician characteristics; the effect of each individual characteristic, based on the β values, was modest. To get an idea of which characteristics in even small ways were related to performance we examined the individual variables. Controlling for other characteristics, specialized cardiologists had lower composite scores ($P = 0.027$). However, physicians who spent on average more time in primary, principal, or longitudinal consultative patient care had higher composite scores ($P = 0.021$). Physicians also tended to have higher composite scores if they scored higher on their initial IM certification exam ($P = 0.045$), were women ($P = 0.001$), had proportionally fewer

Table 4 Results of multivariate regression analysis associating physician and patient characteristics with composite scores (N = 695)^a

Explanatory variables	<i>b</i>	Standard error of <i>b</i>	β	<i>t</i> value	<i>P</i> value
Physician characteristics					
General cardiologists	-0.47	1.41	-0.02	-0.33	0.740
Specialized cardiologists	-3.68	1.66	-0.14	-2.21	0.027
Other subspecialists	-0.98	3.40	-0.01	-0.29	0.773
Equated IM certification exam score ^b	0.01	0.01	0.08	2.01	0.045
Age	0.14	0.09	0.06	1.64	0.101
Female	4.61	1.33	0.15	3.46	0.001
Solo practice	1.87	1.38	0.05	1.35	0.176
% Time in primary, principal, or longitudinal, consultative care	0.04	0.02	0.11	2.32	0.021
Patient characteristics^c					
Age	0.14	0.09	0.07	1.62	0.105
% Female	-0.15	0.03	-0.20	-4.97	<0.001
Patients' primary insurance sources (%)^d					
Medicare	-0.05	0.03	-0.10	-1.74	0.083
Medicaid	-0.15	0.06	-0.11	-2.80	0.005
Both medicare and medicaid	-0.04	0.03	-0.07	-1.30	0.196
Self-pay or uninsured	0.00	0.06	0.00	0.03	0.979
Other (e.g., Veterans administration, government)	-0.07	0.03	-0.14	-2.65	0.008

^a 116 observations were dropped due to missing data

^b Scores were from the physicians' first attempt and were statistically equated to be comparable over time and scaled to have a mean of 500 and SD of 100

^c Patient characteristics were averaged at the physician level

^d Omitted category is private insurance

female patients in their panel ($P < 0.001$), and proportionally more Medicaid insured patients in their panel ($P = 0.005$). Furthermore, we examined the characteristics of the 116 physicians that were not included in the regression analysis and determined that in general they had similar composite scores and characteristics as those in the study sample, except they were older, less likely to be female, and more likely to be subspecialists.

To help physicians identify areas of improvement, each physician received a feedback report on the PIM website about their performance on the individual measure and the composite measure. Specifically, we provided normative information that included their quartile ranking on each individual measure and the composite relative to their peers. We also included histograms for each measure showing the physician's performance relative to their peers and to the standard (Fig. 2). These feedback reports were based on results of prior cognitive testing with physicians and feedback reports (Weng et al. 2010; Hess et al. 2011). From this work, physicians helped us determine what figures were most useful to them for quality improvement and how performance relative to their peers as well as a benchmark provided incredibly valuable feedback that most were not getting from any other source.

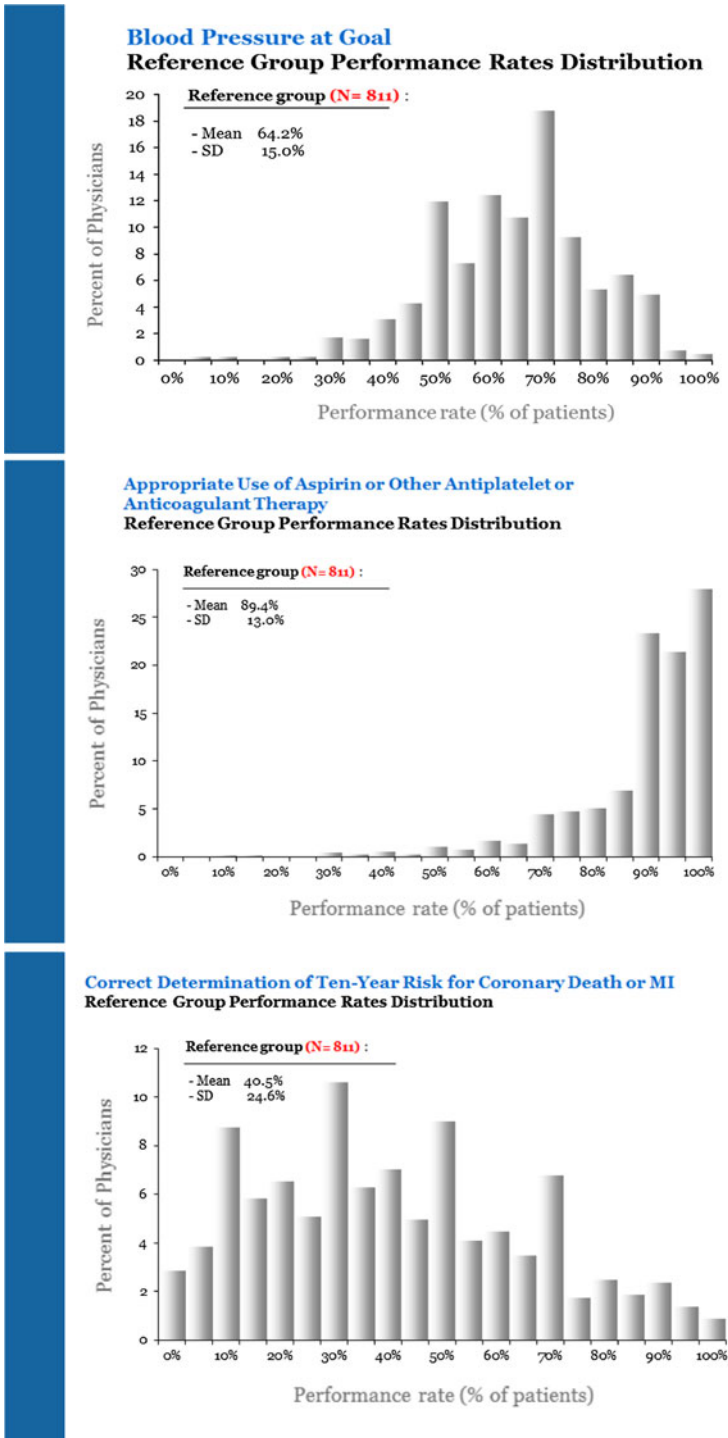


Fig. 2 Examples of the distribution of individual measures (N = 811 Physicians)

Discussion

Our study adds to the literature on clinical performance measurement. We evaluated *individual* physician's quality of preventive cardiovascular disease care by creating a robust composite score using clinical data, and adapted a rigorous standard-setting method to determine a benchmark for minimally acceptable performance. Consistent with Weng et al. (2010) the methodology yielded composite scores that although did not explain a lot of the performance difference, they were, on average, meaningfully associated with physician characteristics, providing some small amount evidence for the validity of these scores. That is, specialized cardiologists demonstrated lower quality preventive care than other physicians. This was expected because their practice is focused less on prevention and more on conducting procedural interventions, specialized testing, or the treatment of cardiovascular disease. Physicians who provided more primary, longitudinal, or preventive consultative care demonstrated higher quality care in preventive cardiology. Because they see patients more often, their patients may be more likely to comply with the physicians' recommendations. These physicians may also have a more systematic approach to managing patients with regards to preventive care (Holmboe et al. 2006). We also found that physicians who scored higher on their initial IM certification exam tended to have higher composite scores. This is consistent with Hess et al. (2012) demonstrating a similar modest association between physicians' cognitive skills, as measured by the ABIM certification exam and composite performance scores from the ABIM Diabetes PIM. Because six of our eight measures were processes of care, we found that female physicians tended to demonstrate higher quality care, consistent with other research that showed patients of female physicians had a greater likelihood of receiving preventive screening and counseling (Henderson and Weisman 2001). Furthermore, physicians who had proportionally fewer women and fewer Medicaid insured patients in their patient panels demonstrated higher quality care. One explanation for this may be that physicians tend to be less aggressive with female patients because of their lower risk of heart disease and different presentation of symptoms than males. The physician-level quality gap found for greater Medicaid patients is consistent with previous findings at the health plan level (National Committee for Quality Assurance 2012).

The standard/benchmark of performance classified physicians quite accurately (i.e., a classification index of 0.98). Most board-certified physicians easily met this standard. The panel acknowledged that the standard was not meant to signal good outcomes of care but rather initially identify outliers who were poor performers and to help physicians identify specific areas in need of improvement. This same methodology could be used to set benchmarks for other defined levels of preventive care (e.g., excellent preventive care). The panel understood that there would be no consequences at this time for physicians who did not meet the benchmark. However, eventually, responsibly-developed information about a physician's performance, particularly in primary care where the prevention of disease is important, could be used to support health care choices by patients and purchasers, or to reward physicians for the care they provide. Another purpose for this methodology is to provide feedback to physicians (Figs. 1, 2) not only on their performance relative to the benchmark but relative to their peers who have taken the PIM. Some evidence exists that when made aware of performance as compared to their peers, physicians are encouraged to improve due to this collegial form of competition (Ornstein et al. 2004; Centers for Disease Control & Prevention 2013b). Finally, the methodology could be used within residency and fellowship training programs, as well as in certification programs to assess performance in practice; however more research needs to be done on data

accuracy, measure specification, risk adjustment, and evidence-based guidelines before a defensible and meaningful standard could be used for this purpose.

This study has limitations. First, physicians voluntarily selected the ABIM Preventive Cardiology PIM to satisfy the practice performance requirement of MOC, thus these results may not generalize to all physicians practicing preventive cardiovascular care. Second, there was no audit process to ensure that instructions for sampling and accurately extracting data were followed. However, prior work reported the accuracy of physician reported data especially in situations where the purpose was quality improvement (Holmboe et al. 2006). Third, the relative importance of each measure was based on subjective ratings of nine experts but might be better defined by an evidence base when it becomes available. Fourth, the measure for correct determination of 10 year risk for coronary death or MI was controversial in that our experts differed on how this calculation is implemented in practice. Finally, no formal risk adjustment was done to control for differences in characteristics of patient panels among physicians. Most measures that we used, however, did incorporate some degree of patient differences especially with respect to co-morbidities in the calculation (e.g., the blood pressure measure set the goal as <130/80 mm Hg for those with chronic kidney disease or diabetes, and <140/90 mm Hg for other patients).

In conclusion, we have established an evidence-based model that is reliable, valid, and defensible for assessing physician performance in preventive cardiology. This model is backed by evidence-based guidelines, proven methodology, and empirical data. Performance standards have been previously set successfully using this model for the care of diabetes patients (Hess et al. 2011). To encourage improved patient care we provide physicians feedback relative to other physicians as well as to an absolute minimum acceptable performance level of care. Our future research will adjust the measures as they mature with the evidence base and apply this methodology to other areas of medicine and may incorporate other types of measures in the composite such as patient experience of care.

Acknowledgments We thank the standard-setting panel and Halyna Didura for her assistance in preparing the data.

Conflict of interest Drs. Hess, Weng, and Lipner are employees of the ABIM and co-inventors of a business method invention describing the application of the standard-setting method to practicing physicians. The invention is patent pending. Dr. Caverzagie has a small portion of his salary reimbursed by the ABIM for efforts related to PIM development and maintenance.

Appendix

See Table 5.

Table 5 Preventive cardiovascular disease: individual measure specifications

Measure title	Description	Numerator	Denominator	Rationale
Intermediate outcomes				
Blood pressure at goal	Percentage of patients in the sample whose most recent blood pressure reading was at goal	Number of patients in the sample whose most recent blood pressure reading was <130/80 mm Hg for those with chronic kidney disease or diabetes, OR <140/90 mm Hg for other patients	Number of patients in the sample	Randomized controlled trials conclusively demonstrate the benefit of lowering blood pressure to <140 systolic and <80 diastolic in patients. Epidemiologic studies show that the risk of cardiovascular disease begins at blood pressures of >115/75. Experts have therefore agreed that <130/80 is a reasonable target for blood pressure control in patients
Low-density lipoprotein (LDL cholesterol) at goal	Percentage of patients in the sample whose LDL cholesterol is considered to be at goal, based upon their coronary heart disease risk factors	Number of patients in the sample whose LDL cholesterol is at goal, based upon their risk factors for coronary heart disease	Number of patients in the sample	Continuing evidence shows that high total and LDL cholesterol levels are strongly related to coronary artery disease risk and that reductions in LDL levels are associated with reduced coronary disease risk
Processes of care				
Timing of lipid testing complies with guidelines	Percentage of patients in the sample whose timing of lipid testing complies with guidelines	Number of patients in the sample who had lipid testing performed (that includes serum triglycerides, total cholesterol, high-density lipoprotein (HDL), low-density lipoprotein (LDL)) and whose timing of testing complies with guidelines	Number of patients in the sample	Lipid abnormalities contribute to the risk of cardiovascular disease. Patients with prior myocardial infarction, other established coronary heart disease, or a coronary heart disease risk equivalent, including diabetes, should have had lipid testing done within two years of the most recent visit. All others should have had lipid testing done within 5 years of the most recent visit

Table 5 continued

Measure title	Description	Numerator	Denominator	Rationale
Diabetes documentation or screen test	Percentage of patients in the sample who had a screening test for type 2 diabetes or had a diagnosis of diabetes	Number of patients in the sample who either had diabetes diagnosis documented OR had a diabetes screening test if diabetes diagnosis was not documented	Number of patients in the sample	A large body of epidemiological and pathological data documents that diabetes is an independent risk factor for cardiovascular disease
Correct determination of 10 year risk for coronary death or myocardial infarction	Number of patients in the sample whose ten-year risk of coronary death or myocardial infarction is correctly assessed and documented	Number of patients in the sample whose ten-year risk of coronary death or myocardial infarction is correctly assessed and documented. This risk is calculated using prior history of coronary heart disease and other risk factor data entered as part of the chart review	Number of patients in the sample	Effective prevention of cardiovascular disease requires an adequate risk-factor assessment to categorize patients for the selection of appropriate therapeutic intervention. Framingham 10 year risk scores have face validity and provide excellent discrimination of high-risk (20 % or greater), intermediate-risk (10–20 %), and low-risk (<10 %) individuals
Dietary and physical activity modifications appropriately prescribed	Percentage of patients in the sample who received dietary and physical activity counseling	Number of patients in the sample who received a dietary prescription and has documentation of activity status for active patients or received counseling for physical activity	Number of patients in the sample	Multiple studies have shown that interventions targeting dietary patterns, weight reduction, and physical activity habits are effective in reducing cardiovascular disease risk factors. There are considerable published data to strongly support the benefits of physical activity and dietary changes as a means to decrease the morbidity and mortality of cardiovascular disease and stroke in adults

Table 5 continued

Measure title	Description	Numerator	Denominator	Rationale
Appropriate use of aspirin or other antiplatelet/anticoagulant therapy	Percentage of patients in the sample who are: (1) taking aspirin or other anticoagulant/antiplatelet therapy, or (2) under age 30, or (3) age 30 or older and who are documented to be at low risk	Number of patients in the sample who are: (1) taking aspirin or other anticoagulant/antiplatelet therapy, or (2) under age 30, or (3) age 30 or older and who are documented to be at low risk. Low-risk patients are those whose ten-year risk of developing coronary heart disease is <10 %	Number of patients in the sample	The net benefit of aspirin depends on the initial risks for stroke and gastrointestinal bleeding. Thus, decisions about aspirin therapy should consider the overall risk for stroke and gastrointestinal bleeding
Smoking status and cessation support	Percentage of patients in the sample whose current smoking status is documented in the chart. If they were smokers, were documented to have received smoking cessation counseling during the reporting period	Number of patients in the sample with documentation of smoking status AND for smokers, with documentation of smoking cessation counseling or treatment during the specified abstraction period (within 12 months of the visit date, with a 3 month grace period)	Number of patients in the sample	A number of large randomized clinical trials have demonstrated the efficacy and cost-effectiveness of smoking cessation counseling in changing smoking behavior and reducing tobacco use. The routine and thorough assessment of tobacco use is important as a means of preventing smoking or encouraging cessation

References

- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 514–515). Washington, DC: American Council on Education.
- Brennan, R. L. (2006). *Educational measurement* (4th ed.). Westport, CT: Praeger.
- Centers for Disease Control & Prevention. (2013a). *Million HeartsTM*. Retrieved January 1, 2013 from <http://millionhearts.hhs.gov/index.html>.
- Centers for Disease Control & Prevention. (2013b). *Strategies for Increasing Adult Vaccination Rates*. Retrieved January 1, 2013 from <http://www.cdc.gov/vaccines/recs/rate-strategies/adultstrat.htm>.
- Centers for Medicare & Medicaid Services. (2013). *Physician compare initiative*. Retrieved January 1, 2013 from www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/physician-compare-initiative/index.html.
- Donabedian, A. (1988). The quality of care: How can it be assessed? *Journal of the American Medical Association*, *260*, 1743–1748.
- Duffy, F. D., Lynn, L. A., Didura, H., Hess, B., Caverzagie, K., Grosso, L., et al. (2008). Self assessment of practice performance: Development of the ABIM practice improvement module (PIM). *Journal of Continuing Education in the Health Professions*, *28*, 39–46.
- Dunn-Rankin, P. (1983). *Scaling methods*. Hillsdale, NY: Erlbaum.
- Henderson, J. T., & Weisman, C. S. (2001). Physician gender effects on preventive screening and counseling: An analysis of male and female patients' health care experiences. *Medical Care*, *39*, 1281–1292.
- Hess, B. J., Weng, W., Holmboe, E. S., & Lipner, R. S. (2012). The association between physicians' cognitive skills and quality of diabetes care. *Academic Medicine*, *87*, 157–163.
- Hess, B. J., Weng, W., Lynn, L. A., Holmboe, E. S., & Lipner, R. S. (2011). Setting a fair performance standard for physicians' quality of patient care. *Journal of General Internal Medicine*, *26*, 467–473.
- Holmboe, E. S., Arnold, G. K., Weng, W., & Lipner, R. (2010). Current yardsticks may be inadequate for measuring quality improvements from the medical home. *Health Affairs*, *29*, 859–866.
- Holmboe, E. S., Meehan, T. P., Lynn, L., Doyle, P., Sherwin, T., & Duffy, F. D. (2006). Promoting physicians' self-assessment and quality improvement: The ABIM diabetes practice improvement module. *Journal of Continuing Education in the Health Professions*, *26*, 109–119.
- Kaplan, S. H., Griffith, J. L., Price, L. L., Pawlson, L. G., & Greenfield, S. (2009). Improving the reliability of physician performance assessment: Identifying the "physician effect" on quality and creating composite measures. *Medical Care*, *47*, 378–387.
- Landon, B. E., Normand, S. L., Blumenthal, D., & Daley, J. (2003). Physician clinical performance assessment: Prospects and barriers. *Journal of the American Medical Association*, *290*, 1183–1189.
- Lee, W.-C., Brennan, R. L., & Wan, L. (2009). Classification consistency and accuracy for complex assessments under the compound multinomial model. *Applied Psychological Measurement*, *33*, 374–390.
- Lipner, R. S., Weng, W., Arnold, G. K., Duffy, F. D., Lynn, L. A., & Holmboe, E. S. (2007). A three-part model for measuring diabetes care in physician practice. *Academic Medicine*, *82*(Suppl 10), S48–S52.
- Mosier, C. (1943). On the reliability of a weighted composite. *Psychometrika*, *8*, 161–168.
- National Committee for Quality Assurance. (2012). *Focus on Obesity and on Medicare Plan Improvement: The State of Healthcare Quality 2012*. Retrieved January 1, 2013 from <http://www.ncqa.org/Portals/0/State%20of%20Health%20Care/2012/SOHC%20Report%20Web.pdf>.
- Ornstein, S., Jenkins, R. G., Nietert, P. J., Feifer, C., Roylance, L. F., Nemeth, L., et al. (2004). A multi method quality improvement intervention to improve preventive cardiovascular care. *Annals of Internal Medicine*, *141*, 523–532.
- Reeves, D., Campbell, S. M., Adams, J., Shekelle, P. G., Kontopantelis, E., & Roland, M. O. (2007). Combining multiple indicators of clinical quality: An evaluation of different analytic approaches. *Medical Care*, *45*, 489–496.
- Scholle, S. H., Pawlson, L. G., Solberg, L. I., Shih, S. C., Asche, S. E., Chou, A. F., et al. (2008). Measuring practice systems for chronic illness care: Accuracy of self-reports from clinical personnel. *Joint Commission Journal of Quality & Patient Safety*, *34*, 407–416.
- Weng, W., Hess, B. J., Lynn, L. A., Holmboe, E. S., & Lipner, R. S. (2010). Measuring physicians' performance in clinical practice: Reliability, classification accuracy, and validity. *Evaluation and the Health Professions*, *33*, 302–320.