MAYO CLINIC

A Multifaceted Organizational Physician Assessment Program: Validity Evidence and Implications for the Use of Performance Data

Andrea N. Leep Hunderfund, MD, MHPE; Yoon Soo Park, PhD; Frederic W. Hafferty, PhD; Kelly M. Nowicki, MA; Steven I. Altchuler, PhD, MD; and Darcy A. Reed, MD, MPH

Abstract

Objective: To provide validity evidence for a multifaceted organizational program for assessing physician performance and evaluate the practical and psychometric consequences of 2 approaches to scoring (mean vs top box scores).

Participants and Methods: Participants included physicians with a predominantly outpatient practice in general internal medicine (n=95), neurology (n=99), and psychiatry (n=39) at Mayo Clinic from January 1, 2013, through December 31, 2014. Study measures included hire year, patient complaint and compliment rates, note-signing timeliness, cost per episode of care, and Likert-scaled surveys from patients, learners, and colleagues (scored using mean ratings and top box percentages).

Results: Physicians had a mean ± SD of 0.32±1.78 complaints and 0.12±0.76 compliments per 100 outpatient visits. Most notes were signed on time (mean ± SD, 96%±6.6%). Mean ± SD cost was 0.56±0.59 SDs above the institutional average. Mean ± SD scores were 3.77±0.25 on 4-point and 4.06±0.31 to 4.94±0.08 on 5-point Likert-scaled surveys. Mean ± SD top box scores ranged from 18.6%±16.8% to 90.7%±10.5%. Learner survey scores were positively associated with patient survey scores (r=0.26; P=.003) and negatively associated with years in practice (r=−0.20; P=.02).

Conclusion: This study provides validity evidence for 7 assessments commonly used by medical centers to measure physician performance and reports that top box scores amplify differences among high-performing physicians. These findings inform the most appropriate uses of physician performance data and provide practical guidance to organizations seeking to implement similar assessment programs or use existing performance data in more meaningful ways.

 ■ Mayo Clin Proc Inn Qual Out 2017;1(2):130-140

From the Department of Neurology (A.N.L.H.), Department of General Internal Medicine (F.W.H.), Department of Clinical Ethics (K.M.N.), Department of Psychiatry (S.I.A.), and Department of Primary Care Internal Medicine (D.A.R.), Mayo Clinic, Rochester, MN; and Medical Education, University of Illinois at Chicago (Y.S.P.).

As a self-regulating profession, medicine is accountable for ensuring that physicians are competent in performing their clinical roles and responsibilities,[1,2] and health care organizations play an important role in this process.[3,4] Organizations collect physician performance data for many reasons (eg, ensuring physician competency, supporting health care choices by consumers, improving care quality, or satisfying regulatory or accreditation requirements)[5] and can use performance data in various ways. For example, scores can be used to ensure that minimal performance expectations are met[3,6] or to drive continuous improvement.[7-9] Failure to meet performance expectations can lead directly to punitive consequences or can trigger additional investigations to determine whether a concern exists.[10-13] Likewise, scores can be used primarily as formative feedback[14,15] or for higher-stakes decisions (eg, promotion, employment, salary, privileging, and public transparency).[9,10,16-19]

This panoply of purposes complicates the collection, distribution, analysis, and interpretation of physician performance data. Without a rigorous examination of the validity of their physician assessment programs, organizations risk using physician performance data in ways that are inappropriate or potentially

detrimental.[20-22] Furthermore, the validity of commonly used physician performance measures may not be sufficient to support all intended purposes.

The use of physician performance data is further complicated by different approaches to scoring. For example, scores based on Likert-type ratings of performance can be reported as means (as often done for learner, multisource, or peer feedback surveys[1,23]) or as the percentage of optimal ratings, also known as top box scores (as often done for patient satisfaction surveys[24-26]). The way in which scores are calculated affects their validity (eg, mean scores better represent the distribution of ratings, while top box scores may be more readily understood),[27-29] yet this issue has not been extensively examined in the context of a multifaceted organizational physician performance assessment program.

For these reasons, we sought to (1) provide validity evidence for 7 different types of assessments commonly used to measure physician performance and (2) examine the practical and psychometric consequences of the 2 aforementioned approaches to scoring (mean vs top box scores).

PARTICIPANTS AND METHODS

This study was a retrospective analysis of deidentified physician clinical performance data collected via routine institutional practices and was considered exempt by the Mayo Clinic Institutional Review Board.

Study Participants and Setting

Study participants included all physicians with a predominantly outpatient practice in general internal medicine (GIM; n=95), neurology (n=99), and psychiatry (n=39) at Mayo Clinic in Rochester, Minnesota, from January 1, 2013, through December 31, 2014. Physicians within the 3 included specialties collectively completed more than 300,000 outpatient visits during the study time frame.

Measures

Physician performance measures included the following:

- Unsolicited patient complaints and compliments related to physician care, reported as the number of complaints or compliments per 100 outpatient visits.
- Percentage of notes that were signed on time according to institutional policy (eg, clinical notes must be signed within 30 days).
- Mean internal cost per episode of care (ie, cost to the institution of providing tests and consults within a discrete period), reported as a z score relative to the institutional mean. Internal costs reflect utilization (eg, physicians who order more or more costly tests and consultations have higher internal costs) and are unrelated to prices or charges to patients/insurers. Internal costs are attributed to the physician with the highest evaluation and management billing code on the first day of a patient's evaluation. An episode of care comprises the subsequent days over which tests and consultations are performed.
- Patient satisfaction survey provided by Avatar International LLC[30] (9 items rated using a 5-point Likert scale ranging from 1 = strongly disagree to 5 = strongly agree, 0 = not applicable).
- Learner feedback surveys, ie, evaluation forms completed by residents and fellows (subsets of items from a total pool of 22 items rated using a 5-point Likert scale: 1 = needs improvement, 2-4 = average, 5 = top 10%, 0 = not applicable; free-text comments required for ratings of 1 or 5).
- Multisource feedback (MSF) surveys for GIM (7 items rated using a 5-point Likert scale: 1 = needs improvement, 2-4 = meets expectations, 5 = exceeds expectations, 0 = not applicable; free-text comments required for ratings of 1 or 5) and psychiatry (5 items rated using a 4-point Likert scale ranging from 1 = strongly disagree to 4 = strongly agree).
- Peer feedback survey for neurology[31] (6 items rated using a 5-point Likert scale: 1 = never, 2 = rarely, 3 = occasionally, 4 = frequently, 5 = always, 0 = not applicable).

These data were collected for a variety of internal, accreditation, certification, and regulatory reasons, as is typical of physician performance data.[32-34] Scores were not linked to physician reimbursement or published publicly. The GIM and psychiatry MSF surveys

were completed by self-selected physicians, allied health professionals, and nonphysician coworkers. Neurology peer feedback surveys were completed by assigned physician and nurse practitioner colleagues. During the study time frame, specialties aimed to collect MSF surveys every 2 to 3 years (GIM), yearly (psychiatry), or twice yearly (neurology).

Because previous studies have reported that care quality may decline with increasing years in practice,[1,35-38] we also collected the hire year of each participant. To protect anonymity, hire year was reported as a categorical variable (with no fewer than 5 individuals per category), and other demographic data (age, sex, and academic rank) were not linked to performance measures.

Data Collection

An individual external to the study team compiled physician performance data from institutional databases and replaced identifiers with random subject IDs to allow data linkage across assessments. Only numeric ratings of performance were included due to the potential for written comments to contain identifying information. After all identifiers were replaced, the key linking identifiers with random subject IDs was destroyed. Only completely and permanently deidentified data were shared with the study team.

Score Calculation

For Likert-scaled assessments, we calculated mean scores and top box scores. To determine mean scores, we first calculated mean ratings for individual survey items, then a mean score across all items in each instrument. For top box scores, we calculated the percentage of ratings that received the highest possible rating. For all measures, separate scores were calculated for 2013 and 2014. To summarize overall performance, we calculated the mean score across both years.

Standard Setting

The Joint Commission requires health care organizations to periodically monitor physician performance via ongoing professional practice evaluation.[10-13] Organizations set performance thresholds, and failure to meet these thresholds triggers more detailed performance assessments using direct observation, medical record audits, etc (focused professional practice evaluation [FPPE]).[10-13] The rate at which FPPE is triggered is an important consideration for leaders, who must determine whether institutional resources are sufficient to accommodate the number of physicians requiring more detailed assessments.

To determine theoretical trigger rates for each assessment, we set normative cutoff scores at 1 and 2 SD from the mean, as recommended by others.[13,39,40] Specifically, cutoff scores were set 1 and 2 SD above the mean for patient complaints, below the mean for timeliness of note signing and feedback surveys, and above and below the mean for cost per episode of care.

Outcomes

As recommended in the *Standards for Educational and Psychological Testing* by the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education,[20] we sought validity evidence from (1) content, as measured by adequacy of content sampling (using the Value Compass,[14,41] which conceptualizes physician performance as clinical processes, clinical outcomes, patient satisfaction, and costs); (2) response process, as measured by score distributions and means for each assessment and scoring method, number of physicians assessed per year, and number of raters (and ratings) per physician per year; (3) internal structure, as measured by internal consistency reliability (consistency in measurement among survey items, reflecting the degree to which items measure a single construct) and item discrimination indices; (4) relations to other variables, as measured by associations among scores generated by the various assessments and associations between scores and hire year; and (5) consequences of testing, as measured by FPPE trigger rates.

Statistical Analyses

Descriptive summary statistics are reported as mean ± SD or frequency (percentage). Internal consistency reliability was estimated using Cronbach α, with reliability coefficients of 0.80 or greater considered sufficient for moderate- to high-stakes summative assessment.[42] Item discrimination indices were calculated

Mayo Clin Proc Inn Qual Out ■ September 2017;1(2):130-140 ■ http://dx.doi.org/10.1016/j.mayocpiqo.2017.05.005
www.mcpiqojournal.org

using item-rest correlations.[42] Associations among scores and between scores and hire year were measured using Pearson correlations. All tests were 2-sided, and $P<.05$ was considered statistically significant. Analyses were performed by one of us (Y.S.P.) using Stata 14 software (StataCorp LLC).

RESULTS

Demographic characteristics of study participants are shown in Table 1. The mean ± SD age of participants was 50.1±11.4 years. Sixty-five percent of participants (151 of 233) were men, and 35% (82 of 233) were women.

Validity Evidence From Content

Five assessments measured clinical processes either directly (timeliness of note signing) or indirectly (physician performance in clinical settings as rated by learners and colleagues) (Table 2). Three assessments measured patient satisfaction, and 1 measured costs. None measured clinical outcomes.

TABLE 1. Demographic Characteristics of the 233 Study Participants[a]

Characteristic	Physicians[b]
Age (y), mean ± SD[c,d]	50.1±11.4
Male sex (No. [%])[d]	151 (65)
Specialty (No. [%])	
General internal medicine	95 (41)
Neurology	99 (42)
Psychiatry	39 (17)
Academic rank (No. [%])[c]	
Professor	47 (20)
Associate professor	30 (13)
Assistant professor	114 (49)
Instructor	17 (7)
No rank	25 (11)
Hire year (No. [%])	
2010-2014	52 (22)
2005-2009	38 (16)
2000-2004	39 (17)
1995-1999	37 (16)
1990-1994	28 (12)
1985-1989	13 (5)
1980-1984	12 (5)
Before 1980	14 (6)

[a]Percentages may not sum to 100% due to rounding.
[b]Participating physicians were those identified by their department or division chair as having a predominantly outpatient clinical practice.
[c]Age and academic rank as of January 1, 2014 (the midpoint of the 2-year study time frame).
[d]Age and sex data were not linked to physician performance data to protect the anonymity of study participants.

Validity Evidence From Response Process

The mean ± SD rates of complaints and compliments per physician were 0.32±1.78 and 0.12±0.76 per 100 outpatient visits, respectively (Table 2). A high percentage of notes were signed on time (mean ± SD, 96.0%± 6.6%), and the mean ± SD internal cost per episode of care was 0.56±0.59 SD above the institutional mean. As shown in Table 2, mean scores were quite high for patient, learner, multisource, and peer feedback surveys and were skewed toward favorable ratings irrespective of the rating scale used. Top box scores showed more variation. The percentage of optimal ratings was less than 50% for learner and GIM MSF surveys (which required free-text comments to select the highest rating) and greater than 80% for patient, psychiatry multisource, and neurology peer feedback surveys (which did not have this requirement).

Assessments supported by institutional resources (patient complaints and compliments, timeliness of note signing, cost per episode of care, patient satisfaction surveys, and learner evaluations) were used to assess more physicians per year than assessments developed and deployed in individual divisions or departments (multisource and peer feedback surveys) (Table 3). Patient satisfaction surveys had a mean ± SD of 36±18 raters per physician per year; the other survey-based assessments averaged 7 or fewer raters per physician per year.

Validity Evidence From Internal Structure

Cronbach α values for patient, learner, multisource, and peer feedback tools were all greater than 0.83 (Table 3). Mean item discrimination indices ranged from 0.73 to 0.88, indicating that items were very effective at discriminating between high and low levels of performance.

Validity Evidence From Relations to Other Variables

Physicians who received higher mean scores on learner evaluations tended to also receive

TABLE 2. Physician Clinical Performance Assessments: Corresponding Content Domains and Scores[a]

Assessment (physicians, No.)[b]	Scale	Content domain[c]	Mean scores[d]		Top box scores[e]	
			Potential scores	Observed scores, mean ± SD	Potential scores	Observed scores, mean ± SD
Patient complaint rate[f] (n=226)	Complaints per 100 outpatient visits (No.)	Patient satisfaction	0.00+	0.32±1.78	NA	NA
Patient compliment rate[f] (n=226)	Compliments per 100 outpatient visits (No.)	Patient satisfaction	0.00+	0.12±0.76	NA	NA
Timeliness of note signing (n=231)	Clinical notes signed on time (%)	Clinical processes	0-100	96.0±6.6	NA	NA
Mean internal cost per episode of care[g] (n=210)	*z* Score relative to the institutional mean	Costs	−3 to +3 SD[h]	0.56±0.59	NA	NA
Patient satisfaction survey (n=201)	5-Point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree)	Patient satisfaction	1.00-5.00	4.73±0.27	0%-100%	85.8 (11.0)
Learner evaluations (n=141)	5-Point Likert scale ranging from 1 (needs improvement) to 5 (top 10%)[i]	Clinical processes	1.00-5.00	4.06±0.31	0%-100%	18.6 (16.8)
MSF, internal medicine (n=10)	5-Point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree)[i]	Clinical processes	1.00-5.00	4.41±0.49	0%-100%	45.0 (25.7)
Peer feedback, neurology (n=94)	5-Point Likert scale ranging from 1 (never) to 5 (always)	Clinical processes	1.00-5.00	4.94±0.08	0%-100%	90.7 (10.5)
MSF, psychiatry (n=36)	4-Point Likert scale ranging from 1 (strongly disagree) to 4 (strongly agree)	Clinical processes	1.00-4.00	3.77±0.25	0%-100%	81.5 (21.9)[j]

[a]MSF = multisource feedback; NA = not applicable.

[b]Of 233 physicians (95 internists, 99 neurologists, and 39 psychiatrists); assessment data are from 2013 and 2014 except for general internal medicine MSF data, which were collected only during 2014.

[c]Using the Value Compass as a conceptual framework.[41]

[d]For Likert-scaled assessments, means were calculated first at the level of individual survey items, then across all items on a given instrument. For all measures, separate scores were calculated for 2013 and 2014, then averaged to summarize overall performance.

[e]For Likert-scaled assessments, scores represent the percentage of optimal ratings (ie, the highest possible Likert scale rating) across all items for a given a physician over the course of a year; separate scores were calculated for 2013 and 2014, then averaged to summarize overall performance.

[f]Unsolicited complaints and compliments related to physician care.

[g]Cost represents the internal costs of providing care to a patient, reflects utilization (eg, physicians who order more [or more costly] tests and consultations have higher internal cost per episode of care), and is unrelated to prices or charges to patients/insurers. Internal costs are attributed to the physician with the highest evaluation and management billing code on the first day of a patient's evaluation, and the subsequent days or weeks over which tests and consultations are performed are considered an episode of care.

[h]Captures greater than 99% of normally distributed data.

[i]Entry of free-text comments was required for ratings of 1 or 5.

[j]Data are from 2014 only (n=30) because psychiatry MSF data from 2013 were stored in a way that precluded calculation of top box scores.

TABLE 3. Physician Clinical Performance Assessments: Response Process and Internal Structure Validity Evidence[a,b]

Assessment	Items (No.)	Response process (No.), mean ± SD			Internal structure	
		Physicians assessed per year[c]	Raters per physician per year	Ratings per physician per year	Cronbach α	Item discrimination index, mean ± SD[d]
Patient complaint rate	NA	217 (1)	NA	NA	NA	NA
Patient compliment rate	NA	217 (1)	NA	NA	NA	NA
Timeliness of note signing	NA	225 (1)	NA	NA	NA	NA
Mean internal cost per episode of care	NA	205 (3)	NA	NA	NA	NA
Patient satisfaction survey	9	191 (2)	36 (18)	314 (156)	0.97	0.88 (0.04)
Learner evaluations	22[e]	115 (21)	6 (2)	126 (47)	0.96	0.74 (0.12)
MSF (general internal medicine)	7	11[f]	4[f]	27[f]	0.89	0.88 (0.08)
Peer feedback (neurology)	6	92 (1)	7 (0)	58 (19)	0.83	0.78 (0.06)
MSF (psychiatry)	5	26 (8)	6 (2)	17 (13)	0.96	0.73 (0.11)

[a]MSF = multisource feedback; NA = not applicable.
[b]Assessment data are from 2013 and 2014 except for general internal medicine MSF data, which were collected only during 2014.
[c]Of 233 eligible physicians (although the number of physicians eligible for assessment by learner evaluations was likely <233 because not all physicians interact with residents and fellows). Specialties aimed to collect multisource or peer feedback for each physician every 2 to 3 years (general internal medicine, 95 physicians), every year (psychiatry, 39 physicians), or twice per year (neurology, 99 physicians).
[d]Item discrimination indices were calculated at the item level using item-rest correlation coefficients, then averaged across all items within a given assessment to generate a mean item discrimination index.
[e]Total pool of items; individual learner evaluation forms contained subsets of items.
[f]No standard deviation because only 2014 data were available.

higher scores on patient satisfaction surveys (r=0.26; P=.003), whereas neurologists with higher mean internal costs per episode of care tended to receive lower scores on the neurology peer feedback survey (r=−0.27; P=.008) (Table 4). The latter finding remained significant when top box scores were used (Supplemental Table 1, available online at http://www.mcpiqojournal.org). There were no other significant correlations.

Physicians with more years in practice at our organization tended to receive lower mean scores on learner evaluations (r=−0.20; P=.02). Otherwise, there were no significant associations between scores and hire year, irrespective of whether mean or top box scores were used (Supplemental Table 2, available online at http://www.mcpiqojournal.org).

Validity Evidence From Consequences

Table 5 shows trigger rates resulting from the normative cutoff scores. The trigger rate was highest for internal cost per episode of care and lowest for patient complaints. Trigger rates for top box scores were higher than for mean scores when cutoff scores were set at 1 SD from the mean.

DISCUSSION

This study provides validity evidence for 7 different assessments commonly used by medical centers to determine whether physician performance is meeting professional standards. It is also the first, to our knowledge, to analyze the effects of different approaches to scoring (mean vs top box scores). A careful examination of validity evidence and scoring procedures is of practical importance to organizational leaders because it provides guidance regarding the most appropriate uses of physician performance data.[20] It can also inform the efforts of those seeking to develop an organizational physician assessment program or use existing performance data in more meaningful ways.

An ideal physician assessment program would be capable of adequately measuring physician clinical performance. In keeping with previous studies,[14] we analyzed the content validity of our assessment program using the Value Compass,[41] which conceptualizes physician performance as clinical processes, clinical outcomes, patient satisfaction, and costs. In doing so, we identified only 1 direct measure of clinical processes and no measures of actual clinical outcomes. Although clinical outcome data often exist at

TABLE 4. Physician Clinical Performance Assessments: Correlation Matrix (Using Mean Scores)[a,b]

	Patient complaint rate	Patient compliment rate	Timeliness of note signing	Mean internal cost per episode of care	Patient satisfaction survey	Learner evaluations	MSF (GIM)	Peer feedback (neurology)	MSF (psychiatry)
Patient complaint rate	1.00								
Patient compliment rate	−0.01 (.91)	1.00							
Timeliness of note signing	0.06 (.38)	0.05 (.50)	1.00						
Mean internal cost per episode of care	−0.01 (.90)	0.02 (.75)	0.05 (.47)	1.00					
Patient satisfaction survey	−0.02 (.74)	0.04 (.56)	−0.12 (.09)	0.02 (.81)	1.00				
Learner evaluations	−0.10 (.25)	−0.04 (.60)	−0.09 (.29)	−0.16 (.07)	0.26 (.003)	1.00			
MSF (GIM)	−0.34 (.32)	NA[c]	−0.50 (.11)	−0.22 (.51)	0.42 (.19)	−0.93 (.24)	1.00		
Peer feedback (neurology)	0.08 (.46)	0.12 (.27)	−0.08 (.43)	−0.27 (.008)	0.12 (.25)	0.10 (.35)	NA	1.00	
MSF (psychiatry)	−0.03 (.86)	NA[c]	0.18 (.29)	0.11 (.53)	0.10 (.59)	−0.07 (.71)	NA	NA	1.00

[a]GIM = general internal medicine; MSF = multisource feedback; NA = not applicable.
[b]Data are given as correlation coefficients (*P* values); mean scores were calculated first at the item level, then across all items within a given instrument.
[c]Insufficient variability precluded calculation of a correlation coefficient.

the practice group and organization levels, they are difficult to attribute to individual physicians in a team-based care environment.[15,43-45] This complicates the interpretation of cost data (which are best interpreted in conjunction with measures of care quality[46-48]) and highlights the challenge of procuring outcome data that can be ascribed to individual physicians.

Multisource feedback is more readily available than clinical outcome data, but it can be logistically challenging to obtain feedback from a sufficient number of raters. We found that patient satisfaction surveys averaged 36 raters per physician per year, which met the recommended minimum of 30 to 50 patient raters.[43,49] However, the other survey-based assessments averaged 7 or fewer raters per physician per year, which failed to meet the recommended minimum of 8 to 12 raters for learner,[50] multisource,[1] and peer[51] feedback. This may reflect rater fatigue, disquietude over assessing colleagues, or inattention to survey invitations. Institutional support seems to play an important role, as efforts to obtain feedback were more successful when they were supported institutionally than when they were developed and deployed in individual divisions and departments. This is consistent with previous studies demonstrating the feasibility of MSF,[1] particularly when it is collected via a national process[52,53] or required for licensure.[4]

The survey-based assessments had excellent internal consistency reliability and desirable psychometric properties. However, mean scores tended to be quite high, with little variation based on hire year. Other studies of Likert-scaled physician assessments completed by patients,[1,4,54] learners,[50] coworkers,[1,4] and peers[1,4,51] have had similar findings. This may reflect inflated ratings of performance (eg, due to reluctance on the part of raters to assign low scores). Alternatively, it could indicate that practicing physicians are generally at the top of the learning curve with respect to performance, as might be expected given their career stage. Taken together, these findings suggest that survey-based assessments may be able to identify physicians who fail to meet accepted performance standards. However, skew toward higher ratings makes it difficult to use scores for

Mayo Clin Proc Inn Qual Out ■ September 2017;1(2):130-140 ■ http://dx.doi.org/10.1016/j.mayocpiqo.2017.05.005
www.mcpiqojournal.org

TABLE 5. Physician Clinical Performance Assessments: Consequences of Measurement[a,b]

		Threshold							
		1 SD from the mean[c]				2 SD from the mean[c]			
		Mean scores[d]		Top box scores[e]		Mean scores[d]		Top box scores[e]	
Assessment	Physicians (No.)	Cutoff score	Trigger rate (No. [%])	Cutoff score	Trigger rate (No. [%])	Cutoff score	Trigger rate (No. [%])	Cutoff score	Trigger rate (No. [%])
Patient complaint rate	226	>2.1	4 (2)	NA	NA	>3.9	3 (1)	NA	NA
Timeliness of note signing	231	<89.5%	24 (10)	NA	NA	<82.9%	10 (4)	NA	NA
Mean internal cost per episode of care	210	<−0.02 or >1.15	63 (30)	NA	NA	<−0.61 or >1.74	13 (6)	NA	NA
Patient satisfaction survey	201	<4.46	15 (7)	<74.8%	18 (9)	<4.19	4 (2)	<63.8%	0
Learner evaluations	141	<3.75	13 (9)	<1.8%	20 (15)	<3.43	4 (3)	<1%[f]	0
MSF (internal medicine)	10	<3.92	1 (10)	<19.3%	2 (18)	<3.43	1 (10)	<1%[f]	0
Peer feedback (neurology)	94	<4.86	13 (14)	<80.3%	13 (14)	<4.78	3 (3)	<69.8%	4 (4)
MSF (psychiatry)	36	<3.52	6 (17)	<59.6%[g]	11 (31)[g]	<3.26	1 (3)	37.7%[g]	8 (22)[g]

[a]MSF = multisource feedback; NA = not applicable.

[b]Assessment data are from 2013 and 2014 except for general internal medicine MSF data, which were collected only during 2014; cutoff scores were not applied to patient compliments.

[c]Hypothetical cutoff scores set at 1 or 2 SD above the mean for patient complaints; 1 or 2 SD below the mean for timeliness of note signing, patient satisfaction survey, learner evaluations, and multisource or peer feedback surveys; or 1 or 2 SD above and below the mean for mean internal costs per episode of care.

[d]For Likert-scaled assessments, means were calculated first at the level of individual survey items, then across all items on a given instrument. For all measures, separate scores were calculated for 2013 and 2014, then averaged to summarize overall performance.

[e]For Likert-scaled assessments, scores represent the percentage of optimal ratings (ie, the highest possible Likert scale rating) across all items for a given physician over the course of a year; separate scores were calculated for 2013 and 2014, then averaged to summarize overall performance.

[f]A cutoff score of less than 1% was used when 2 SD below the mean was a negative value.

[g]Data are from 2014 only (psychiatry MSF data from 2013 were stored in a way that precluded calculation of top box scores).

more aspirational purposes (eg, continuous improvement to increasingly higher levels of professional excellence) because this would require instruments capable of discriminating among high-performing individuals.

We found that scores on Likert-scaled assessments were more discriminating when they were reported as top box scores than when they were reported as means. Greater discrimination among physicians may be advantageous if the intended purpose of measurement is to inspire continuous improvement. However, amplifying differences among high performers also risks engendering demoralization, disregard for performance data, or attempts to "game" the assessment system[5,22,33,55-58] and may result in higher FPPE trigger rates. This is especially problematic for scores based on small sample sizes.[59] Thus, the method used to calculate scores should be selected carefully in light of potential consequences of testing, and organizations receiving top box scores (eg, from patient satisfaction survey vendors) should be mindful of these considerations when interpreting and distributing performance data.

Interestingly, physicians who were rated more highly by residents and fellows also tended to be rated more favorably by patients. This suggests that these assessments measure a similar construct (eg, interpersonal and communication skills, as suggested by others),[50,51,60-62] whereas the other assessments generally provide distinct perspectives on physician performance. These findings support the value of a multifaceted physician assessment program and underscore the importance of combining multiple approaches when attempting to measure something as complex as physician performance.[5,63-65] It may be useful, for example, to compile various sources of performance data into a dashboard, portfolio, or report card rather than distributing and reviewing it in a piecemeal manner.

Previous studies have found that physician scores on knowledge tests and various performance assessments decrease with increasing years in practice.[1,35-38] This finding provides some rationale for monitoring physician

performance over time.[38] However, we observed little score variation by hire year, which may reflect the known mitigating effects of a practice setting that allows for frequent interactions with colleagues.[66,67] The one exception was learner feedback scores, which declined with increasing years in practice. This could be due to erosion of teaching skills (or an increasing number of competing priorities) among physicians over time. Alternatively, learners may prefer faculty who are closer to them in career stage. Further studies are needed to better understand this association.

This study has several limitations. First, we analyzed physician performance data from 3 specialties at 1 organization with a salary-based physician reimbursement model. However, other medical centers collect similar data,[4,18,68] which supports the generalizability of these findings. Second, written comments from patients, learners, or colleagues may provide a rich source of feedback,[54,64,69,70] but we only examined validity evidence for numeric data. Third, we were not able to analyze associations between scores and age, sex, or academic rank, given concern for preserving anonymity. Fourth, previous studies have used legal or disciplinary action, adherence to standards of care (based on analyses of billing, medical record, or administrative data), medical record audits, or specialty board recertification examination failure rates to identify underperforming physicians,[39] but these data were not available for analysis. Finally, we used a normative approach to standard setting, which detects deviations from the average performance of a high-performing group. However, other standard setting approaches exist[5,6,42] and may be preferred depending on the intended use of scores.

CONCLUSION

Health care organizations face the formidable task of implementing physician assessment programs capable of simultaneously advancing institutional goals, meeting regulatory and accreditation requirements, and providing meaningful feedback to physicians. These findings suggest that individual physician performance data are most appropriately used in combination to detect deviations from expected standards, which can then be further investigated (eg, using FPPE) to determine whether a true concern exists. Although MSF is more readily available than clinical outcome data, obtaining a sufficient number of raters per physician can be challenging without institutional support. Top box scores are more discriminating than mean scores. However, amplifying differences among high performers may have unintended consequences and increase FPPE trigger rates.

ACKNOWLEDGMENTS

We acknowledge Paul S. Mueller, MD (chair of the Mayo Clinic Rochester Division of General Internal Medicine), Mark A. Frye, MD (chair of the Mayo Clinic Rochester Department of Psychiatry and Psychology), and Robert D. Brown, MD, MPH (director of the Mayo Clinic Program in Professionalism and Values and former chair of the Mayo Clinic Rochester Department of Neurology) for their support of this project and valuable feedback on the manuscript; Kuan Xing, MEd (University of Illinois at Chicago) for his assistance with statistical analyses; and Brianna Tranby, MA (administrative assistant, Mayo Clinic Rochester Department of Psychiatry and Psychology) for her assistance with data collection.

The content of this article reflects the views of the authors. The American Board of Psychiatry and Neurology was not involved in the study design, data collection/analysis/interpretation, writing of the report, or decision to submit this article for publication.

SUPPLEMENTAL ONLINE MATERIAL

Supplemental material can be found online at http://www.mcpiqojournal.org. Supplemental material attached to journal articles has not been edited, and the authors take responsibility for the accuracy of all data.

Abbreviations and Acronyms: **FPPE** = focused practice performance evaluation; **GIM** = general internal medicine; **MSF** = multisource feedback

Grant Support: This study was prepared with financial support from the American Board of Psychiatry and Neurology and the Mayo Clinic Program in Professionalism and Values.

Potential Competing Interests: Dr Altchuler has disclosed that he was a part-time Field Representative for The Joint

Commission during the time this work was being developed. The rest of the authors report no competing interests.

Correspondence: Address to Andrea N. Leep Hunderfund, MD, MHPE, Department of Neurology, Mayo Clinic, 200 First St SW, Rochester, MN 55905 (leep.andrea@mayo.edu).

REFERENCES

1. Donnon T, Al Ansari A, Al Alwai S, Violato C. The reliability, validity, and feasibility of multisource feedback physician assessment: a systematic review. *Acad Med*. 2014;89(3):511-516.
2. Cruess SR, Cruess RL. The medical profession and self-regulation: a current challenge. *Virtual Mentor*. 2005;7(4).
3. Chassin MR, Baker DW. Aiming higher to enhance professionalism: beyond accreditation and certification. *JAMA*. 2015; 313(18):1795-1796.
4. Wright C, Richards SH, Hill JJ, et al. Multisource feedback in evaluating the performance of doctors: the example of the UK General Medical Council patient and colleague questionnaires. *Acad Med*. 2012;87(12):1668-1678.
5. Landon BE, Normand SL, Blumenthal D, Daley J. Physician clinical performance assessment: prospects and barriers. *JAMA*. 2003;290(9):1183-1189.
6. Hess BJ, Weng W, Lynn LA, Holmboe ES, Lipner RS. Setting a fair performance standard for physicians' quality of patient care. *J Gen Intern Med*. 2011;26(5):467-473.
7. Irons MB, Nora LM. Maintenance of certification 2.0: strong start, continued evolution. *N Engl J Med*. 2015;372(2):104-106.
8. Nora LM, Wynia MK, Granatir T. Of the profession, by the profession, and for patients, families, and communities: ABMS board certification and medicine's professional self-regulation. *JAMA*. 2015;313(18):1805-1806.
9. Lee VS, Miller T, Daniels C, Paine M, Gresh B, Betz AL. Creating the exceptional patient experience in one academic health system. *Acad Med*. 2016;91(3):338-344.
10. Standards BoosterPaks™: a quality improvement tool. The Joint Commission website. https://www.jointcommission.org/standards_booster_paks. Published December 7, 2015. Accessed February 17, 2017.
11. Are you on board with the Joint Commission's FPPE/OPPE requirements? *Hosp Peer Rev*. 2009;34(12):137-141.
12. Hunt JL. Assessing physician competency: an update on the joint commission requirement for ongoing and focused professional practice evaluation. *Adv Anat Pathol*. 2012; 19(6):388-400.
13. Makary MA, Wick E, Freischlag JA. PPE, OPPE, and FPPE: complying with the new alphabet soup of credentialing. *Arch Surg*. 2011;146(6):642-644.
14. Veloski J, Boex JR, Grasberger MJ, Evans A, Wolfson DB. Systematic review of the literature on assessment, feedback, and physicians' clinical performance: BEME guide No. 7. *Med Teach*. 2006;28(2):117-128.
15. Crosson FJ. Physician professionalism in employed practice. *JAMA*. 2015;313(18):1817-1818.
16. Bland CJ, Wersal L, VanLoy W, Jacott W. Evaluating faculty performance: a systematically designed and assessed approach. *Acad Med*. 2002;77(1):15-30.
17. Leverence R, Nuttal R, Palmer R, et al. Using organizational philosophy to create a self-sustaining compensation plan without harming academic missions. *Acad Med*. [epub ahead of print].
18. Kairouz VF, Raad D, Fudyma J, Curtis AB, Schünemann HJ, Akl EA. Assessment of faculty productivity in academic departments of medicine in the United States: a national survey. *BMC Med Educ*. 2014;14:205.
19. Fung CH, Lim YW, Mattke S, Damberg C, Shekelle PG. Systematic review: the evidence that publishing patient care performance data improves quality of care. *Ann Intern Med*. 2008;148(2):111-123.
20. American Educational Research Association; American Psychological Association; National Council on Measurement in Education [AERA/APA/NCME]. *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association; 2014.
21. Pronovost PJ, Lilford R. Analysis & commentary: a road map for improving the performance of performance measures. *Health Aff (Millwood)*. 2011;30(4):569-573.
22. Berwick DM. Era 3 for medicine and health care. *JAMA*. 2016; 315(13):1329-1330.
23. Beckman TJ, Lee MC, Mandrekar JN. A comparison of clinical teaching evaluations by resident and peer physicians. *Med Teach*. 2004;26(4):321-325.
24. Press Ganey Improvement Portal® user guide. http://helpandtraining.pressganey.com/lib-docs/default-source/ip-training-resources/ImprovementPortal_UserGuide.pdf?sfvrsn=8. Accessed February 23, 2017.
25. Summary analyses: Hospital Consumer Assessment of Healthcare Providers and Systems website. http://hcahpsonline.org/SummaryAnalyses.aspx. Accessed February 23, 2017.
26. Otani K, Waterman B, Faulkner KM, Boslaugh S, Burroughs TE, Dunagan WC. Patient satisfaction: focusing on "excellent." *J Healthc Manag*. 2009;54(2):93-102.
27. Fullam F, Garman AN, Johnson TJ, Hedberg EC. The use of patient satisfaction surveys and alternative coding procedures to predict malpractice risk. *Med Care*. 2009;47(5):553-559.
28. Sauro J. Top-box scoring of rating scale data. https://MeasuringU.com/top-box. Published December 14, 2010. Accessed February 24, 2017.
29. Drain M. Quality improvement in primary care and the importance of patient perceptions. *J Ambul Care Manage*. 2001;24(2):30-46.
30. Avatar International LLC. http://www.avatarsolutions.com. Accessed January 1, 2015.
31. American Board of Psychiatry and Neurology Peer Feedback Form. https://www.abpn.com/wp-content/uploads/2015/01/ABPN-Peer-Feedback-Form.pdf. Accessed June 14, 2017.
32. Panzer RJ, Gitomer RS, Greene WH, Webster PR, Landry KR, Riccobono CA. Increasing demands for quality measurement. *JAMA*. 2013;310(18):1971-1980.
33. Miller TP, Brennan TA, Milstein A. How can we make more progress in measuring physicians' performance to improve the value of care? *Health Aff (Millwood)*. 2009;28(5):1429-1437.
34. Cohen AB, Sanders AE, Swain-Eng RJ, et al. Quality measures for neurologists: financial and practical implications. *Neurol Clin Pract*. 2013;3(1):44-51.
35. Choudhry NK, Fletcher RH, Soumerai SB. Systematic review: the relationship between clinical experience and quality of health care. *Ann Intern Med*. 2005;142(4):260-273.
36. Kohatsu ND, Gould D, Ross LK, Fox PJ. Characteristics associated with physician discipline: a case-control study. *Arch Intern Med*. 2004;164(6):653-658.
37. Khaliq AA, Dimassi H, Huang CY, Narine L, Smego RA Jr. Disciplinary action against physicians: who is likely to get discipline? *Am J Med*. 2005;118(7):773-777.
38. McDade WA. Report 5 of the Council on Medical Education (A-15): competency and the aging physician. http://www.cppph.org/cppph/wp-content/uploads/2016/02/AMA-Council-on-Medical-Education-Aging-Physician-Report-2015.pdf. Accessed March 6, 2017.
39. Williams BW. The prevalence and special educational requirements of dyscompetent physicians. *J Contin Educ Health Prof*. 2006;26(3):173-191.
40. The accountability moment: 10 principles for moving beyond OPPE to a rigorous, credible physician PI framework. Advisory Board website. https://www.advisory.com/Research/Physician-Executive-Council/Studies/2010/The-Accountability-Moment. Accessed March 2, 2017.

41. Nelson EC, Mohr JJ, Batalden PB, Plume SK. Improving health care, part 1: the clinical value compass. *Jt Comm J Qual Improv*. 1996;22(4):243-258.
42. Downing SM, Yudkowsky R, eds. *Assessment in Health Professions Education*. New York, NY: Routledge; 2009.
43. The challenges of measuring physician quality. Agency for Healthcare Research and Quality website. https://www.ahrq.gov/professionals/quality-patient-safety/talkingquality/create/physician/challenges.html. Accessed February 16, 2017.
44. Higgins A, Zeddies T, Pearson SD. Measuring the performance of individual physicians by collecting data from multiple health plans: the results of a two-state test. *Health Aff (Millwood)*. 2011;30(4):673-681.
45. Attribution: principles and approaches. National Quality Forum website. https://www.qualityforum.org/Publications/2016/12/Attribution_-_Principles_and_Approaches.aspx. Published December 2016. Accessed February 24, 2017.
46. Smoldt RK, Cortese DA. Pay-for-performance or pay for value? *Mayo Clin Proc*. 2007;82(2):210-213.
47. Lee VS. Redesigning metrics to integrate professionalism into the governance of health care. *JAMA*. 2015;313(18):1815-1816.
48. Porter ME. What is value in health care? *N Engl J Med*. 2010;363(26):2477-2481.
49. National Committee for Quality Assurance. *HEDIS 2007 Technical Specifications for Physician Measurement*. Washington, DC: National Committee for Quality Assurance; 2007.
50. Beckman TJ, Ghosh AK, Cook DA, Erwin PJ, Mandrekar JN. How reliable are assessments of clinical teaching? a review of the published instruments. *J Gen Intern Med*. 2004;19(9):971-979.
51. Hawkins RE, Lipner RS, Ham HP, Wagner R, Holmboe ES. American Board of Medical Specialties Maintenance of Certification: theory and evidence regarding the current framework. *J Contin Educ Health Prof*. 2013;33(suppl 1):S7-S19.
52. Multisource feedback: how does multisource feedback (MSF) work? Royal College of Physicians and Surgeons of Canada website. http://www.royalcollege.ca/rcsite/credentials-exams/exam-eligibility/assessment-imgs/per/multisource-feedback-e. Accessed March 3, 2017.
53. Hall W, Violato C, Lewkonia R, et al. Assessment of physician performance in Alberta: the physician achievement review. *CMAJ*. 1999;161(1):52-57.
54. Siegrist RB Jr. Patient satisfaction: history, myths, and misperceptions. *Virtual Mentor*. 2013;15(11):982-987.
55. Ofri D. Quality measures and the individual physician. *N Engl J Med*. 2010;363(7):606-607.
56. Epstein A. Performance reports on quality: prototypes, problems, and prospects. *N Engl J Med*. 1995;333(1):57-61.
57. Zgierska A, Rabago D, Miller MM. Impact of patient satisfaction ratings on physicians and clinical care. *Patient Prefer Adherence*. 2014;8:437-446.
58. Hayward RA, Kent DM. 6 EZ steps to improving your performance: (or how to make P4P pay 4U!). *JAMA*. 2008;300(3):255-256.
59. Bachman JW. The problem with patient satisfaction scores. *Fam Pract Manag*. 2016;23(1):23-27.
60. Cheng SH, Yang MC, Chiang TL. Patient satisfaction with and recommendation of a hospital: effects of interpersonal and technical aspects of hospital care. *Int J Qual Health Care*. 2003;15(4):345-355.
61. Manary MP, Boulding W, Staelin R, Glickman SW. The patient experience and health outcomes. *N Engl J Med*. 2013;368(3):201-203.
62. Hall MF, Press I. Keys to patient satisfaction in the emergency department: results of a multiple facility study. *Hosp Health Serv Adm*. 1996;41(4):515-532.
63. Violato C, Marini A, Toews J, Lockyer J, Fidler H. Feasibility and psychometric properties of using peers, consulting physicians, co-workers, and patients to assess physicians. *Acad Med*. 1997;72(10, suppl 1):S82-S84.
64. van der Vleuten CP, Schuwirth LW, Scheele F, Driessen EW, Hodges B. The assessment of professional competence: building blocks for theory development. *Best Pract Res Clin Obstet Gynaecol*. 2010;24(6):703-719.
65. Schuwirth LW, van der Vleuten CP. Programmatic assessment and Kane's validity perspective. *Med Educ*. 2012;46(1):38-48.
66. Norcini JJ, Lipner RS, Benson JA, Webster GD. An analysis of the knowledge base of practicing internists as measured by the 1980 recertification examination. *Ann Intern Med*. 1985;102(3):385-389.
67. Day SC, Norcini JJ, Webster GD, Viner ED, Chirico AM. The effect of changes in medical knowledge on examination performance at the time of recertification. *Res Med Educ*. 1988;27:139-144.
68. Khullar D, Kocher R, Conway P, Rajkumar R. How 10 leading health systems pay their doctors. *Healthc (Amst)*. 2015;3(2):60-62.
69. Burford B, Illing J, Kergon C, Morrow G, Livingston M. User perceptions of multi-source feedback tools for junior doctors. *Med Educ*. 2010;44(2):165-176.
70. Santuzzi NR, Brodnik MS, Rinehart-Thompson L, Klatt M. Patient satisfaction: how do qualitative comments relate to quantitative scores on a satisfaction survey? *Qual Manag Health Care*. 2009;18(1):3-18.