

Benchmarking Physician Performance: Reliability of Individual and Composite Measures

Sarah Hudson Scholle, MPH, DrPH; Joachim Roski, PhD, MPH;
John L. Adams, PhD; Daniel L. Dunn, PhD; Eve A. Kerr, MD, MPH;
Donna Pillittere Dugan, MS; and Roxanne E. Jensen, BA

Measuring physician performance is becoming commonplace as health plans and purchasers look for ways to drive quality improvement and to increase physicians' accountability and rewards for achieving quality goals. A recent study¹ reported that, among 89% of health maintenance organization plans using physician-oriented pay-for-performance programs, more than one-third measured and rewarded quality at the individual physician level. In addition, public and private purchasers are demanding more information about America's physicians and hospitals to aid in value-based purchasing and selection of health plans and providers.²

However, concerns remain regarding the validity and reliability of such physician performance profiles. Several factors are needed to support fair and accurate comparisons among physicians. These include evidence-based quality measures, complete and accurate data sources, and standardized methods of data collection. Physician-level reliability of a quality measure is another key consideration in this measurement. Physician-level reliability refers to the ability of a quality measure to distinguish an individual physician's performance from the performance of physicians overall. Good physician-level reliability requires the following 2 factors: (1) a sufficient number of patients eligible for a given quality measure and (2) performance variation across physicians on that quality measure.³⁻⁵ The greater the number of a physician's patients who are eligible for a quality measure, the more precise the estimate of the physician's performance. When performance variation for a given quality measure across physicians is limited, the likelihood that a physician's performance is statistically significantly different from that of his or her peers is also decreased. Hofer and colleagues⁶ showed that not controlling for a quality measure's physician-level reliability significantly misrepresented performance differences across physicians. However, adjusting performance profiles in such a manner is not commonplace across the healthcare industry.

Ensuring that measurement results are valid and reliable is important when purchasers and plans (and potentially consumers) use the data to make decisions about which physicians get financial rewards or other benefits. The stakes are particularly high when profiling results are used for public reporting or eligibility for participation in a health plan network. Paying attention to the validity

Objective: To examine the reliability of quality measures to assess physician performance, which are increasingly used as the basis for quality improvement efforts, contracting decisions, and financial incentives, despite concerns about the methodological challenges.

Study Design: Evaluation of health plan administrative claims and enrollment data.

Methods: The study used administrative data from 9 health plans representing more than 11 million patients. The number of quality events (patients eligible for a quality measure), mean performance, and reliability estimates were calculated for 27 quality measures. Composite scores for preventive, chronic, acute, and overall care were calculated as the weighted mean of the standardized scores. Reliability was estimated by calculating the physician-to-physician variance divided by the sum of the physician-to-physician variance plus the measurement variance, and 0.70 was considered adequate.

Results: Ten quality measures had reliability estimates above 0.70 at a minimum of 50 quality events. For other quality measures, reliability was low even when physicians had 50 quality events. The largest proportion of physicians who could be reliably evaluated on a single quality measure was 8% for colorectal cancer screening and 2% for nephropathy screening among patients with diabetes mellitus. More physicians could be reliably evaluated using composite scores ($\leq 17\%$ for preventive care, $>7\%$ for chronic care, and 15%-20% for an overall composite).

Conclusions: In typical health plan administrative data, most physicians do not have adequate numbers of quality events to support reliable quality measurement. The reliability of quality measures should be taken into account when quality information is used for public reporting and accountability. Efforts to improve data available for physician profiling are also needed.

(*Am J Manag Care.* 2008;14(12):829-838)

**For author information and disclosures,
see end of text.**

In this issue

Take-away Points / p837

www.ajmc.com

Full text and PDF

Web exclusive

eAppendices

and reliability of data will help to ensure that these decisions are based on real differences in performance among physicians rather than any shortcomings of the measurement.

Although performance results based on limited sample sizes could be adjusted for the reliability of individual measures,^{7,9} the creation of composite scores may also be a useful way to increase the reliability of physicians' performance scores.¹⁰ Little is known about the extent to which constructing composite scores mitigates the limitations of sample size and reliability, while continuing to provide useful and understandable information.¹¹

To date, there have been few reports regarding the reliability of physician-level performance scores associated with commonly used practices and methods in the healthcare industry. To begin to address this deficiency, this study relied on a large data set that combined patient-level administrative data from 9 large health plans to compute performance for primary care physicians (PCPs) using 27 commonly measured quality indicators. This data set is typical of data sources often used by individual health plans to profile physician performance. Specifically, we examined for each quality measure and composite score the proportion of PCPs who could be evaluated given different minimum sample size criteria and the physician-level reliability under those minimum sample size criteria. Our primary research questions were the following: (1) What is the physician-level reliability of commonly used performance measures calculated exclusively based on administrative data? (2) Can more physicians be reliably evaluated using a composite score?

METHODS

Data Sources

This study used administrative data from the Ingenix Impact Pro database.¹² Deidentified claims and enrollment data for individuals enrolled in 9 health plans from 9 separate geographic regions for 2003 and 2004 were available for this study. Each of these plans had at least 250,000 members and accounted for 15% to 50% of managed care enrollees in their markets (Table 1). In all, these plans covered more than 11 million unique members and many physicians and employer groups. The members included in these organizations were primarily enrolled in commercial health maintenance organization, preferred provider organization, and point-of-service health plan product designs, with fewer individuals enrolled in Medicare risk products. Pharmacy benefit status, an indicator of the general availability of pharmacy data to support measurement, ranged from 51% to 80% of the enrolled populations for each plan. Although the study population was drawn from multiple geographic census regions, most indi-

viduals were located in the northeast United States. The data were deidentified to protect patient, physician, and organization confidentiality. This study was reviewed and determined to be exempt by Chesapeake Research Review, Inc (Columbia, MD).

Because the Impact Pro database may not include complete data on all services (eg, pharmacy, laboratory, or mental health services) needed for calculating some performance measures, we conducted specific analyses to assess the completeness of the data available for the study. Using only administrative data sources, we compared performance rates based on Impact Pro data with performance data reported to the National Committee for Quality Assurance (NCQA) through the Healthcare Effectiveness Data and Information Set (HEDIS) reporting. If we found more than a 5-percent-point difference between the plan's reported rate to the NCQA and the rate in the Impact Pro database, the data were excluded for that quality measure.

Selection of Quality Measures

Twenty-seven quality measures often used to assess care effectiveness were calculated using study data following HEDIS specifications.¹³ The quality measures were identified from an environmental scan of existing physician quality measures and prioritization by an NCQA expert panel on physician profiling. The quality measure set primarily includes quality measures that have been endorsed by the AQA Alliance and the National Quality Forum. Only quality measures that could be obtained through administrative claims data were included because we were emulating efforts to profile physicians based on data commonly available and used by health plans. Quality measures for diabetes care, cervical cancer screening, and colorectal cancer screening are specified for hybrid data collection for HEDIS (ie, using medical records data to supplement claims). Relying exclusively on administrative data for performance calculations for these quality measures may not accurately reflect performance.¹⁴ The selected quality measures describe preventive, chronic, and acute care activities and were considered appropriate for supporting comparisons of PCPs. (See eAppendix Table 1, available at www.ajmc.com.)

Identification and Attribution of Quality Events to Physicians

We identified individual physicians using the unique physician identifiers used by health plans. Because we did not have a way to link a physician's claims in one data set to that physician's claims in another health plan's data set, we did not pool patients for the same physician across health plans. Most of the 9 health plans in our study did not operate in the same

■ **Table 1.** Characteristics of Plans Included in the Study

Plan	No. of Members	Female Sex, %	Age of Members, y, %			% With Pharmacy Benefit
			0-17	18-64	≥65	
A	250,000-750,000	51	29	69	2	57
B	≥750,000	51	33	65	2	76
C	250,000-750,000	52	25	74	1	80
D	250,000-750,000	50	27	71	2	51
E	≥750,000	52	30	65	5	51
F	≥750,000	53	25	69	6	71
G	250,000-750,000	54	44	51	5	70
H	250,000-750,000	52	30	66	4	64
I	≥750,000	51	25	72	2	77

healthcare markets, so pooling would have a limited effect. Primary care physicians, including family physicians, general internists, and general pediatricians, were identified based on the specialty designated in the credentialing records of the participating health plans.

The 9 health plans included in this study generally did not require patients to designate a PCP. Therefore, we developed algorithms based on patient care patterns to attribute a patient's care to 1 or more physicians. We required at least 1 claim for an outpatient visit during the measurement period to attribute a patient to a PCP for inclusion in the study. This means that some patients who were eligible for quality measures may not have been attributed (eg, a woman eligible for mammography would not be attributed if she did not have a qualifying visit during the year). Outpatient visits were defined based on coding conventions established through HEDIS to identify preventive and ambulatory health services.¹³

For a physician to be considered responsible for a quality event (defined as an event for which a patient is eligible for a quality measure), the patient had to have a visit with the physician during a period when the physician would have an opportunity to meet the quality indicator. We chose this less stringent approach to maximize the number of quality events assigned to each physician. Any PCP rendering 1 or more visits for a patient during the eligibility period was considered responsible for the quality measure. A specific patient may be eligible for multiple quality measures and contribute multiple quality events for the responsible physicians. Likewise, more than 1 physician could be responsible for a specific quality event.

Statistical Analysis

Individual Measures. For each quality measure, a physician's rate of performance was computed as the number of

patients meeting indicator requirements divided by the number of eligible patients attributed to that physician. In preliminary analyses, we observed high variation in performance rates among physicians with fewer than 10 attributed quality events for a quality measure, so we excluded these physicians from the analysis. We determined the number of physicians meeting 4 different volume thresholds (≥ 10 , ≥ 20 , ≥ 30 , and ≥ 50 quality events) for each quality measure. We also report the mean number of quality events per physician within each volume category and the mean performance rates for each category. The volume categories are not mutually exclusive; a physician with 25 quality events for a quality measure will be included in the volume categories for 10 or more and for 20 or more.

To determine how well quality measures capture differences in physician performance, we calculated a reliability estimate for each quality measure.¹⁵ The methods are summarized herein (see **eAppendix. Reliability Formula**, available at www.ajmc.com). Reliability estimates were calculated for physicians within each plan because we wanted the variability between physicians to reflect the natural variation within a plan and not to be unduly affected by patient or other factors that might differ across plans. We selected 2 plans (plans C and I in Table 1) for inclusion in our presented findings to illustrate potential differences across plans in reliability results for different performance measures. The estimated reliability for a quality measure was computed as the ratio of the physician-to-physician variance divided by the sum of the physician-to-physician variance plus the measurement variance. This reliability estimate varies between 0 and 1, and a reliability of 0.70 or higher is typically considered acceptable for psychometric purposes.¹⁶ Reliability estimates were computed separately for each of the volume categories based on the mean number of quality events for physicians within

that category and on the minimum number of quality events. For example, in the category for 30 or more quality events, all physicians have at least 30 quality events, but some have many more than 30, which tends to elevate the mean reliability. The reliability at the minimum for that volume category reflects the reliability for physicians who had no more than 30 quality events. The reliability estimate at the mean number of quality events for the quality measure and volume category reflects typical experience of physicians in this population. The reliability at the minimum number of quality events per physician shows the “worst case” reliability for the volume category. The reliability estimate at the minimum level of quality events also facilitates comparison of reliability estimates across quality measures because the mean number of quality events varies by quality measures. To promote comparability across volume categories for a quality measure, the physician-to-physician variance for the group of physicians with 10 or more quality events was used for computing reliability for all groups. The Spearman-Brown prophecy formula was used to produce the reliability estimate at each volume level.¹⁷ We also present the distribution of reliability estimates by the number of quality events per physician.

Composite Measures. We grouped the performance measures into 3 composite scores for preventive, chronic, and acute, as well as an overall composite comprising all 27 quality measures. Column 1 of **Table 2** gives the mean number of quality events per quality measure per PCP. Calculating composites at the physician level is a concern because the distribution of quality events may vary across physicians. Some quality measures are harder to meet than others, and if some physicians are eligible for disproportionately high numbers of those quality measures, they will score lower than other physicians. To address these issues, we constructed a composite score for each physician based on the standardized mean rate for the relevant quality measures for that physician’s patient population. By standardizing the quality measures before averaging, we took into account differences in quality measure means and standard deviations that might otherwise distort comparisons across physicians. This method allows for the comparison of performance among physicians on the same scale even if their patient mixes and applicable quality measures were different. For each physician, we calculated a mean of the standardized quality measure rates weighted by the number of quality events that the physician has for the quality measure. In computing a composite score for a physician, only individual indicators with 4 or more quality events were used. With fewer than 4 quality events, a standardized rate is constrained to be 0, 1, or 0.5 and had an unacceptably strong effect on the composite results.

To create the reliability estimate for the composite scores using the method already described, we computed the physician-to-physician variance using a hierarchical modeling approach. For the aforementioned reasons, the reliability estimates were computed at the plan level and within each composite score at different volume thresholds (≥ 10 , ≥ 20 , ≥ 30 , and ≥ 50 quality events). Using the physician-to-physician variance for each plan and volume group for each composite, a reliability was computed for each physician. A sample of up to 900 physicians was used for each composite for each plan. We report the median of the reliability estimates across this sample of physicians.

RESULTS

Individual Measures

Table 2 summarizes the results of the analyses of selected individual performance measures (see **eAppendix Table 2**, available at www.ajmc.com, for full results). The leftmost columns in the table provide information for all the plans on the number and percentage of physicians with quality events for each quality measure and the performance rates, stratified by volume categories. For example, 23,985 physicians (25% of all PCPs with ≥ 1 quality event in the data set) had 10 or more patients eligible for cervical cancer screening attributed to them.

Performance rates were fairly consistent across the volume categories (Table 2). For example, the mean performance on asthma medication management varied from 79% for physicians with 50 or more quality events to 82% for physicians with 10 or more quality events.

Reliability estimates showed considerable variation across quality measures and, as expected, generally improved with an increase in the number of quality events available for characterizing physician performance. With a minimum of 10 quality events, reliability estimates for cervical cancer screening were 0.20 in plan C and 0.23 in plan I. Even with a minimum of 50 quality events, reliability for this quality measure is below the recommended level (0.48 in plan C and 0.57 in plan I). However, among physicians who had at least 10 observations, the mean number of observations per physician was high. For example, the reliability estimate for cervical cancer screening in plan C was 0.70 among physicians with 10 or more quality events because the mean number of quality events per physician was 94. For colorectal cancer screening and Chlamydia screening, the reliability estimates based on the minimum thresholds suggest that at least 50 quality events are needed to gain acceptable reliability. For colorectal cancer screening, a threshold of 50 quality events would ensure a minimum reliability of 0.76 for the physician with exactly 50 quality events,

Reliability in Physician Profiling

Table 2. Results for Individual Measures

Quality Measure ^a	All Plans					Plan C		Plan I	
	Volume Category (No. of Attributed Quality Events) ^b	No. of PCPs Meeting Threshold for Volume Category	% Of PCPs Meeting Threshold for Volume Category ^c	Mean No. of Quality Events Among PCPs Meeting Threshold for Volume Category	Mean Performance Rate for Volume Category, %	Reliability at Mean No. of Quality Events for Volume Category ^d	Reliability at Minimum No. of Quality Events for Volume Category ^e	Reliability at Mean No. of Quality Events for Volume Category ^d	Reliability at Minimum No. of Quality Events for Volume Category ^e
Preventive Care									
Cervical cancer screening	≥10	23,985	25	62	67	0.70	0.20	0.74	0.23
	≥20	17,459	18	79	67	0.75	0.34	0.76	0.37
	≥30	13,898	14	93	67	0.77	0.43	0.77	0.47
	≥50	9326	10	120	68	0.80	0.57	0.80	0.60
Chlamydia screening in women	≥10	2003	2	19	33	0.51	0.34	0.58	0.42
	≥20	626	1	31	34	0.63	0.51	0.68	0.58
	≥30	234	0	43	35	0.70	0.60	0.74	0.67
	≥50	50	0	67	38	0.79	0.72	0.79	0.76
Colorectal cancer screening	≥10	20,680	21	54	42	0.86	0.39	0.74	0.28
	≥20	14,987	15	69	43	0.87	0.56	0.76	0.43
	≥30	11,718	12	82	44	0.88	0.66	0.78	0.53
	≥50	7502	8	106	45	0.89	0.76	0.81	0.65
Chronic Care									
Use of appropriate medications for people with asthma	≥10	1053	1	15	82	0.25	0.17	0.04	0.03
	≥20	164	0	27	82	0.36	0.29	0.07	0.05
	≥30	40	0	40	80	0.45	0.36	0.09	0.07
	≥50	6	0	57	79	0.52	0.49	—	—
Comprehensive diabetes care									
Low-density lipoprotein cholesterol testing	≥10	3578	4	21	84	0.57	0.37	0.44	0.28
	≥20	1286	1	33	86	0.69	0.56	0.54	0.43
	≥30	499	1	48	86	0.75	0.65	0.61	0.52
	≥50	142	0	76	88	0.84	0.75	0.68	0.63
Medical attention for nephropathy	≥10	5796	6	19	47	0.81	0.64	0.72	0.56
	≥20	1765	2	33	50	0.86	0.78	0.80	0.72
	≥30	664	1	48	54	0.90	0.84	0.84	0.79
	≥50	178	0	80	59	0.94	0.90	0.89	0.86
Annual monitoring for patients taking persistent medications (angiotensin-converting enzyme inhibitor)	≥10	7949	8	24	76	0.66	0.37	0.52	0.29
	≥20	3602	4	36	78	0.71	0.54	0.60	0.45
	≥30	1825	2	47	78	0.75	0.64	0.65	0.55
	≥50	530	1	71	79	0.81	0.75	0.72	0.66
Acute Care									
Appropriate treatment for children with upper respiratory tract infections	≥10	5914	6	43	83	0.73	0.36	0.82	0.59
	≥20	3806	4	60	84	0.78	0.54	0.88	0.76
	≥30	2758	3	73	85	0.81	0.64	0.91	0.83
	≥50	1669	2	96	85	0.85	0.76	0.93	0.89
Inappropriate antibiotic treatment for adults with acute bronchitis ^f	≥10	2126	2	19	27	0.63	0.45	0.45	0.28
	≥20	594	1	32	26	0.74	0.63	0.59	0.45
	≥30	219	0	47	24	0.81	0.72	0.68	0.58
	≥50	65	0	72	25	0.86	0.80	0.76	0.71

PCP indicates primary care physician.

^aThe mean is based on the group of physicians who had at least 1 quality event attributed to them for this measure.

^bThe volume categories are not mutually exclusive; a physician with 25 quality events will be included in the volume categories of 10 or more and 20 or more.

^cThe proportion of physicians who meet the volume category requirements for the quality measure out of all physicians with at least 1 quality event on any quality measure (n = 97,268).

^dThe reliability estimate for the mean number of quality events among physicians who meet the volume category requirements for the composite. Estimates above the recommended level are boldfaced.

^eThe reliability estimate for the minimum number of quality events for the volume category. Estimates above the recommended level are boldfaced.

^fLower performance is better for this measure.

but because the mean number of quality events per physician with at least 50 quality events is 129, the reliability estimate for all physicians in this group is 0.89 (based on results from plan C).

Diabetes care and medication monitoring quality measures had the highest reliability estimates among the chronic disease quality measures. For example, the reliability estimate for diabetes low-density lipoprotein cholesterol testing was 0.37 for a minimum of 10 quality events and 0.75 for 50 quality events; for diabetes nephropathy testing, the results were 0.64 for 10 quality events and 0.90 for 50 quality events (both sets of estimates are from plan C). Using the mean number of quality events at each of these thresholds yielded higher reliability estimates.

The following 10 quality measures have reliability estimates above 0.70 when a minimum of 50 quality events is used for the calculation in at least 1 plan: Chlamydia screening, colorectal cancer screening, low-density lipoprotein cholesterol screening, glycosylated hemoglobin testing and nephropathy testing for patients with diabetes mellitus (DM), appropriate use of antibiotics for children with upper respiratory tract infections and adults with bronchitis, and monitoring of patients taking angiotensin-converting enzyme inhibitors, diuretics, and statins. Of these quality measures, only the following 4 achieve this level of reliability with a minimum of 30 quality events in at least 1 plan: nephropathy testing for patients with DM, medication monitoring for statins, and appropriate use of antibiotics in children with upper respiratory tract infections

■ **Table 3.** Results for Composite Measures

Volume Category (No. of Attributed Quality Events) ^a	No. of PCPs Meeting Threshold for Volume Category	% Of PCPs Meeting Threshold for Volume Category ^b	Mean No. of Quality Events Among PCPs Meeting Threshold for Volume Category	Reliability at Mean No. of Quality Events for Each Subscale or Overall Composite Score ^c		Observed Minimum No. of Quality Events Needed to Reach Reliability Estimate of 0.70 ^d	
				Plan C	Plan I	Plan C	Plan I
Preventive care							
≥10	29,802	31	115	0.89	0.88		
≥20	24,035	25	140	0.90	0.87	50	67
≥30	20,659	21	159	0.88	0.89		
≥50	16,502	17	189	0.88	0.87		
Chronic care							
≥10	15,865	16	63	0.91	0.82		
≥20	12,210	13	78	0.90	0.80	40	37
≥30	9811	10	91	0.91	0.79		
≥50	6589	7	116	0.90	0.79		
Acute care							
≥10	11,452	12	44	0.71	0.77		
≥20	7110	7	62	0.69	0.77	40	70
≥30	5000	5	78	0.75	0.82		
≥50	2951	3	106	0.83	0.87		
Overall care							
≥10	34,604	36	145	0.89	0.88		
≥20	28,422	29	173	0.90	0.87	70	62
≥30	24,825	26	195	0.88	0.89		
≥50	20,360	21	229	0.88	0.87		

PCP indicates primary care physician.

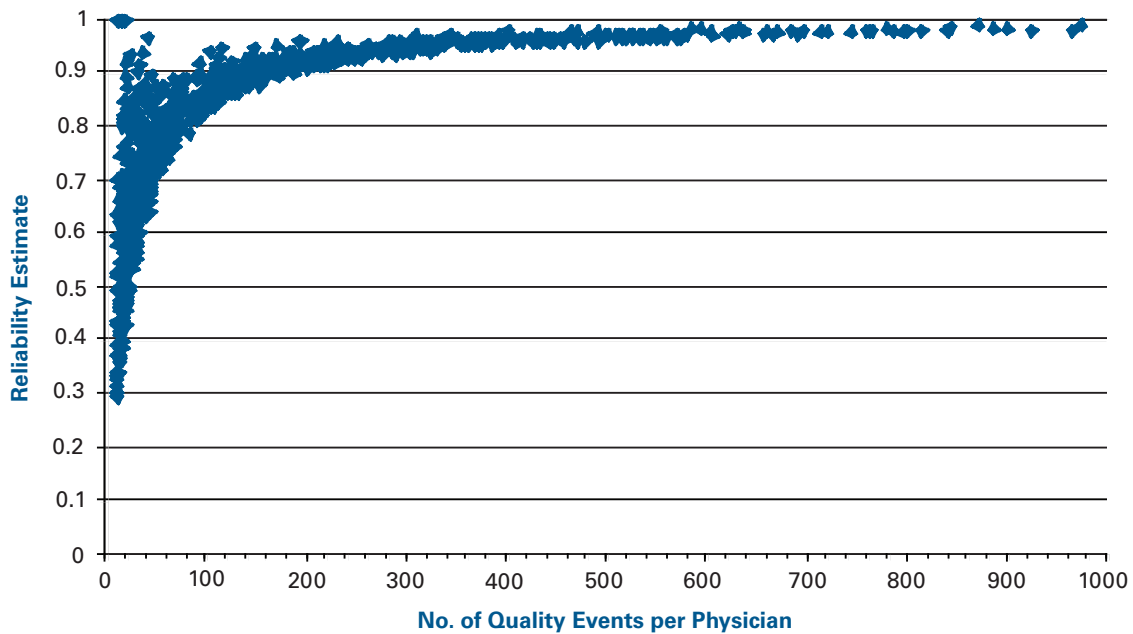
^aThe volume categories are not mutually exclusive; a physician with 25 quality events will be included in the volume categories of 10 or more and 20 or more.

^bThe proportion of physicians who meet the volume category requirements for the quality measure out of all physicians with at least 1 quality event on any quality measure (n = 97,268).

^cThe reliability estimate for the mean number of quality events among physicians who meet the volume category requirements for the composite. Estimates above the recommended level are boldfaced.

^dFrom the distribution of events (see Figure), the minimum number of events found to achieve a reliability estimate of at least 0.70.

■ **Figure.** Reliability of Preventive Care Composite Measure by the Number of Quality Events Per Physician (Plan C)



and in adults with bronchitis. Some quality measures attain a reliability estimate above 0.70 at the mean number of quality events when the threshold is 10 or 20 quality events; none of the quality measures achieve this level of reliability assuming the minimum level for those ranges.

Composite Measures

Table 3 gives the results for the composite measures. Under the least stringent threshold of 10 or more quality events, 36% of physicians received an overall composite score. The proportions of physicians who can be evaluated are higher using the composite scores than using individual items (Table 2), with 31% for the prevention composite versus 25% for cervical cancer screening (the prevention quality measure with the largest proportion of physicians evaluated), 16% for the chronic care composite versus 8% for medication monitoring of patients taking angiotensin-converting enzyme inhibitors, and 12% for the acute care composite versus 6% for appropriate antibiotic prescribing for children with upper respiratory tract infections.

The reliability estimates of the composite scores are presented as the mean number of quality events for the range and are high even for the smallest volume category, namely, physicians with 10 or more quality events (0.90 for preventive care, 0.79 for chronic care, and 0.88 for overall care). This is because even with a minimum threshold of 10 or more quality events, the mean number of quality events per physician is high. For example, using a threshold of 10 or more

quality events, the mean numbers of quality events per physician were 115 for the prevention composite and 63 for the chronic composite (Table 3).

The **Figure** shows the distribution of reliability estimates for the random sample of 900 physicians used to calculate the prevention composite measures of plan C. These estimates show that many physicians have large numbers of quality events available for calculating each of these composites. From this distribution, we can identify a minimum number of quality events needed to achieve a reliability estimate of 0.70 (Table 3). The minimum differs for the 2 plans included in this analysis. For example, 50 quality events are needed to achieve a 0.70 reliability estimate for the preventive care composite in plan C compared with 67 quality events for plan I. We used similar distributions to identify the minimum sample size needed for the other composites (**eAppendix Figures** available at www.ajmc.com). Fewer observations are needed for the chronic care composite (40 in plan C and 37 in plan I). The number varies dramatically by plan for the acute care composite, with 40 quality events needed in plan C and 70 quality events in plan I. For the overall composite, the minimum required is 70 quality events in plan C and 62 quality events in plan I.

DISCUSSION

Measuring and comparing physician performance using administrative data in a reliable way are challenging when

data are limited to only those patients who are represented by data in a single health plan. We found that a small fraction of physicians could be reliably profiled on the most common quality measures when data are limited to a single health plan's administrative data. Using a minimum reliability estimate of 0.70 as the threshold for "reliable measurement," the largest proportions of physicians who could be reliably evaluated on a single quality measure were 8% for colorectal cancer screening (based on plan C results for ≥ 50 quality events) and 2% for nephropathy screening for DM (based on plan C and plan I results for ≥ 20 quality events). Greater proportions of physicians can be reliably evaluated using composite measures ($\leq 17\%$ of physicians using a preventive care composite, $>7\%$ using a chronic care composite, and 15%-20% using the overall composite). Even in this large and varied database representing health plans with moderate-to-large market shares, most physicians could not be evaluated reliably on composites or on individual measures.

We focused on the minimum sample size needed for reliable measurement because physicians who are near the minimum threshold are at the greatest risk for misclassification. Only 10 of 27 quality measures evaluated in this study met the recommended reliability of 0.70 with a minimum of 50 quality events. Four quality measures met this standard at a minimum of 30 quality events. The numbers needed for composite measures ranged from 37 to 70 quality events.

A quality measure can have good reliability (1) because there is comparatively high physician-to-physician variance or (2) because there is not much "noise" or measurement error in the estimate of the individual physician performance, usually as a result of large sample sizes. When most physicians have high scores on a quality measure, reliability may be lower because there is less variation across physicians; therefore, the ability of that quality measure to discriminate performance decreases. For example, the reliability of the nephropathy screening quality measure for DM is higher than that of the glycosylated hemoglobin screening quality measure in large part because on average physicians performed a documented screen for nephropathy in about 50% of their eligible patients with DM, while about 90% of their patients received low-density lipoprotein cholesterol screening. The number of quality events also affects reliability; none of the behavioral health quality measures meet the recommended standard for reliability, in part because of the smaller numbers of quality events for these quality measures.

The desired level of reliability of measurement depends on how the data are used. For confidential performance feedback to physicians that is intended to spur quality improvement efforts, meeting a reliability threshold of 0.70 may not be necessary. However, if plans or purchasers use the information to

determine network eligibility or to create public report cards, a minimum threshold for reliability is needed to ensure fair comparisons and to protect against misclassification bias.

The trade-off in reliable measurement between individual measures and composite scores has to be weighed against the different information available and the different uses of the data by various stakeholders. Individual measures provide readily actionable information for quality improvement purposes for physicians; composite measures may be useful for high-level comparisons of physicians (eg, for report cards directed at purchasers or consumers) but are less directly useful for quality improvement.¹⁸

Composites may be unduly affected by particular quality measures or may combine unrelated quality measures, and the relationship of the composite score to an underlying construct of quality may be unclear. We attempted to address methodological issues that make it difficult to compare quality in one physician's practice with that in another by standardizing scores on each quality indicator before creating a summary score. Physicians vary in the types of patients they see, and quality measures for some conditions tend to be associated with higher (or lower) physician performance than others. The amount of variation in performance also differs from among quality measures. Taking a simple sum of quality indicators (adding up the number of quality events and the number of times the recommended actions were taken) would be straightforward, but this approach would give an advantage to physicians whose patients are mostly eligible for quality measures associated with generally high performance. Standardizing rates means that physicians are evaluated on how they compare with the mean performance on a quality measure. However, standardizing rates is problematic when physicians have few eligible quality events, so we excluded quality measures for which a physician had fewer than 4 quality events (which limited the number of physicians who could be evaluated). Furthermore, we weighted the quality measures based on the number of quality events that the physician had. This meant that the composites were strongly affected by performance on preventive care and medication management quality measures that apply to many patients; quality measures for chronic disease conditions (even the most common ones such as DM) had less affect on the overall composite. We chose this approach because it represents, from a volume perspective, the types of quality events that a physician has the opportunity to influence.

In addition, we calculated reliability at the plan level because we wanted the variability between physicians to reflect the natural variation within a plan and not to be unduly affected by patient or other factors that might differ across plans. The difference in reliability across plans suggests that plans

should consider reliability of quality measures within their own data.

Limitations

Limitations of this study included the limited number of quality measures studied, the reliance on administrative data, the lack of information on physician specialties other than primary care, the lack of direct information about a physician's relationship with the patient, our inability to link physicians across plans, and the lack of risk adjustment. These findings are based on only 27 quality measures from an administrative-only data set. All quality measures are well tested, and most are included in health plans' and employers' physician performance measurement programs and can be calculated based on administrative data. The metrics were selected with input from a stakeholder panel based on relevance, evidence, and feasibility.¹³

Administrative data were used because information from electronic medical records is unavailable on a large scale and because most physician-level measurement efforts around the country rely on administrative data, including enrollment records, medical claims, pharmacy claims, and laboratory claims (but not values) in the calculation of performance results. Therefore, these results represent what can be commonly accomplished using data available to most health plans. However, administrative data limit the types of clinical actions that can be profiled,¹⁹ and reliance on administrative data means that profiles generally depend on process measures rather than on intermediate outcomes, linked action measures, or outcomes such as health status, functional status, and mortality.

This study focuses on PCPs because of the many relevant performance measures and the total number of physicians represented in our database. Our ability to characterize care for specialists was limited by the few quality measures that could be reliably measured through administrative data. In the absence of clear information about whether any single physician was responsible for each quality event, a simple rule—essentially giving credits (or demerits) to any physicians who came into contact with the patient—was used to attribute patients (and their associated quality events) to a physician.

We were unable to link physicians across health plans because there was not a common identifier. To the extent that several of the plans included in this study have overlapping markets and may use the same physician networks, the results may underestimate the true number of denominator counts and the quality measure reliability. Therefore, our results

Take-away Points

When health plan administrative data are used to evaluate physician performance, most quality measures require at least 50 quality events per physician to gain a reliable estimate of physician performance (ie, to ensure that a quality measure is able to distinguish a physician's performance from average performance).

- Composite measures allow more physicians to be evaluated reliably but are less actionable for quality improvement.
- The physician-level reliability of quality measures should be considered when quality information is used for public reporting and accountability.
- Efforts to improve the quality and quantity of data available for physician profiling are also needed.

present a worst-case scenario but on the whole reflect what is generally happening in the marketplace because pooling of data is uncommon.

Furthermore, there was no attempt to adjust for differences in case mix or severity across physicians. All of the quality measures used in this study were process-of-care measures with refined eligibility criteria and exclusion criteria for cases in which recommended care may not be applicable. Risk adjustment is undoubtedly an issue for intermediate outcome measures and may be of particular concern even for process-of-care measures that depend on patient adherence (ie, medication management and breast cancer screening) for practices that serve disadvantaged populations.²⁰

Implications

Our results suggest that information about the reliability of quality measures should be calculated and presented to make quality information transparent at the physician or group level and should be taken into account in determining how measurement information will be used for financial rewards or incentives. Efforts to improve the depth and breadth of information available for assessing physician performance are also needed, along with research to address continuing methodological concerns. Large sample sizes per physician are required to achieve recommended reliability at the level of individual measures, and reliability varies across quality measures. Current reporting efforts based on available administrative data from a single plan or insurer may misrepresent performance of individual physicians if reliability standards are not considered. Efforts to encourage the aggregation of databases across health plans, government purchasers, and other entities are needed to maximize the numbers of quality events per physician that are available for characterizing physician performance.

Author Affiliations: National Committee for Quality Assurance (SHS, JR, DPD), Washington, DC; RAND Corporation (JLA), Santa Monica, CA; Ingenix (DLD), Waltham, MA; Ann Arbor Veterans Affairs Medical Center and University of Michigan (EAK), Ann Arbor; and Johns Hopkins Bloomberg School of Public Health (REJ), Baltimore, MD. Dr Roski is now with the Engelberg Center for Healthcare Reform at the Brookings Institution, Washington, DC.

Funding Source: This study was supported by a grant from the Commonwealth Fund and by grant 1 R13 HS016277 from the Agency for Healthcare Research and Quality. Dr Kerr was supported in part by grant DIB 98-001 from the Veterans Affairs Health Services Research and Development Quality Enhancement Research Initiative for Diabetes Mellitus (DIB #98-001) and by Michigan Diabetes Research and Training Center Grant P60DK-20572 from the National Institute of Diabetes and Digestive and Kidney Diseases.

Previous Presentations: Portions of this work were presented at the 2006 Annual Research Meeting of AcademyHealth; June 26, 2006; Seattle, WA; and at an invitation-only conference convened by the National Commission for Quality Assurance entitled "Benchmarking Physician Performance: Current Practice and Research Needs"; January 28, 2006; Rockville, MD.

Author Disclosure: The authors (SHS, JLA, DPD, REJ) report no relationship or financial interest with any entity that would pose a conflict of interest with the subject matter of this article. Dr Roski reports having served as a consultant for Children's Hospital Minneapolis, National Committee for Quality Assurance, and Impact Education, Inc, and has received grants from Robert Wood Johnson Foundation, Commonwealth Fund, and the Agency for Healthcare Research and Quality. Dr Dunn is an employee of Ingenix, a company that sells quality measurement products and services. However, none of those products are mentioned in this study. Dr Kerr reports serving as an unpaid consultant to the National Committee for Quality Assurance.

Authorship Information: Concept and design (SHS, JR, JLA, DLD, EAK, DPD); acquisition of data (SHS, DLD, DPD); analysis and interpretation of data (SHS, JR, JLA, DLD, EAK, REJ); drafting of the manuscript (SHS, JR, JLA, DLD, REJ); critical revision of the manuscript for important intellectual content (SHS, JR, JLA, EAK); statistical analysis (JLA, DLD); obtaining funding (SHS, JR); administrative, technical, or logistic support (DPD, REJ); and supervision (JR).

Address correspondence to: Sarah Hudson Scholle, MPH, DrPH, National Committee for Quality Assurance, 1100 13th St NW, Ste 1000, Washington, DC 20005. E-mail: scholle@ncqa.org.

REFERENCES

- Baker G, Carter B.** *Provider Pay-for-Performance Incentive Programs: 2004 National Study Results.* San Francisco, CA: Med-Vantage, Inc; 2005.
- Galvin R, Milstein A.** Large employers' new strategies in health care. *N Engl J Med.* 2002;347(12):939-942.
- Greenfield S, Kaplan SH, Kahn R, Ninomiya J, Griffith JL.** Profiling care provided by different groups of physicians: effects of patient case-mix (bias) and physician-level clustering on quality assessment results. *Ann Intern Med.* 2002;136(2):111-121.
- Tucker JL III.** The theory and methodology of provider profiling. *Int J Health Care Qual Assur Inc Leadersh Health Serv.* 2000;13(6-7):316-321.
- Krein SL, Hofer TP, Kerr EA, Hayward RA.** Whom should we profile? examining diabetes care practice variation among primary care providers, provider groups, and health care facilities. *Health Serv Res.* 2002;37(5):1159-1189.
- Hofer TP, Hayward RA, Greenfield S, Wagner EH, Kaplan SH, Manning WG.** The unreliability of individual physician "report cards" for assessing the costs and quality of care of a chronic disease. *JAMA.* 1999;281(22):2098-2105.
- Landon BE, Normand SL, Blumenthal D, Daley J.** Physician clinical performance assessment: prospects and barriers. *JAMA.* 2003;290(9):1183-1189.
- Huang IC, Diette GB, Dominic F, Frangakis C, Wu AW.** Variations of physician group profiling indicators for asthma care. *Am J Manag Care.* 2005;11(1):38-44.
- Huang IC, Diette GB, Dominic F, Frangakis C, Wu AW.** Application of a propensity score approach for risk adjustment in profiling multiple physician groups on asthma care. *Health Serv Res.* 2005;40(1):253-278.
- Caldis T.** Composite health plan quality scales. *Health Care Financ Rev.* 2007;28(3):95-108.
- Felt-Lisk S, Lavin B, Gold M.** *Aggregate Quality Measures for the National Healthcare Quality Report: Summary of Technical Advisory Panel Meetings May/June 2005.* Draft July 1, 2005. Report to the Agency for Healthcare Quality and Research under contract 03 R0030801 D/SIBN 874-1.
- Ingenix.com Web site.** Impact Pro. 2008. <http://www.ingenix.com/Products/Employers/HealthandProductivity/EvidenceBasedHealthEMP/IngenixImpactProPAYUA/>. Accessed October 31, 2008.
- National Committee for Quality Assurance.** *HEDIS 2007 Technical Specifications for Physician Measurement.* Washington, DC: National Committee for Quality Assurance; 2007.
- Pawlsion LG, Scholle SH, Powers A.** Comparison of administrative-only versus administrative plus chart review data for reporting HEDIS hybrid measures. *Am J Manag Care.* 2007;13(10):553-558.
- Fleiss JL, Levin B, Paik MC.** *Statistical Methods for Rates & Proportions.* Indianapolis, IN: Wiley-Interscience; 2003.
- Nunnally J, Bernstein I.** *Psychometric Theory.* 3rd ed. New York, NY: McGraw-Hill; 1994.
- Shrout PE, Fleiss JL.** Intraclass correlations: uses in assessing rater reliability. *Psychol Bull.* 1979;86(2):420-428.
- Kerr EA, Hofer TP, Hayward RA, et al.** Quality by any other name? a comparison of three profiling systems for assessing health care quality. *Health Serv Res.* 2007;42(5):2070-2087.
- Kerr EA, Krein SL, Vijan S, Hofer TP, Hayward RA.** Avoiding pitfalls in chronic disease quality measurement: the case for the next generation of technical quality measures. *Am J Manag Care.* 2001;7(11):1033-1043.
- Casalino LP, Elster A, Eisenberg A, Lewis E, Montgomery J, Ramos D.** Will pay-for-performance and quality reporting affect health care disparities? *Health Aff (Millwood).* 2007;26(3):w405-w414. ■