DOI: 10.1111/j.1475-6773.2007.00730.x
RESEARCH ARTICLE

Performance Measurement

Quality by Any Other Name?: A Comparison of Three Profiling Systems for Assessing Health Care Quality

Eve A. Kerr, Timothy P. Hofer, Rodney A. Hayward, John L. Adams, Mary M. Hogan, Elizabeth A. McGlynn, and Steven M. Asch

Objective. Many performance measurement systems are designed to identify differences in the quality provided by health plans or facilities. However, we know little about whether different methods of performance measurement provide similar answers about the quality of care of health care organizations. To examine this question, we used three different measurement approaches to assess quality of care delivered in veteran affairs (VA) facilities.

Data Sources/Study Setting. Medical records for 621 patients at 26 facilities in two VA regions.

Study Design. We examined agreements in quality conclusions using: focused explicit (38 measures for six conditions/prevention), global explicit (372 measures for 26 conditions/prevention), and structured implicit review physician-rated care (a single global rating of care for three chronic conditions and overall acute, chronic and preventive care). Trained nurse abstractors and physicians reviewed all medical records. Correlations between scores from the three systems were adjusted for measurement error in each using multilevel regression models.

Results. Intercorrelations of scores were generally moderate to high across all three systems, and rose with adjustment for measurement error. Site-level correlations for prevention and diabetes care were particularly high. For example, adjusted for measurement error at the site level, prevention quality was correlated at 0.89 between the implicit and global systems, 0.67 between implicit and focused, and 0.73 between global and focused systems.

Conclusions. We found moderate to high agreement in quality scores across the three profiling systems for most clinical areas, indicating that all three were measuring a similar construct called "quality." Adjusting for measurement error substantially enhanced our ability to identify this underlying construct.

Key Words. Quality of care, performance profiling, quality monitoring

Although the number of performance measures that are used to monitor and compare the quality of health care organizations continues to proliferate, we know little about the extent to which results from the many current methods to measure quality agree with one another. With the costs associated with quality monitoring increasing, and pay-for-performance initiatives becoming more broadly promoted, healthcare organizations and providers are rightly asking how much of their performance profiles depends on the choice of quality measures or profiling systems used. As important is the broader question of whether these different quality measures are tapping into a unified construct representing quality of care. Fundamentally, the strategy of contracting with or seeking care from certain providers or organizations based on quality requires a coherent and measurable construct that represents the quality of care. Yet, most current profiles of care quality are based on a limited set of feasible measures that may or may not represent the broader construct of health care quality.

For over three decades, experts have argued whether it is possible to capture such a health care quality "construct" by using measures that assess the quality of technical care. Donabedian (1985) asserted that methods that measure care quality can generally be categorized into two approaches: those involving implicit expert (physician) judgments of individual cases (implicit review) and those involving explicit objective standards (referred to as explicit review, although the measures are themselves developed by expert judgments about the literature). A number of studies have attempted to validate quality constructs by examining whether the process of care for discrete conditions as judged by explicit review agreed with implicit judgments of the same care. As long ago as the 1970s, Brook examined agreements in process quality using implicit and explicit review systems for three conditions.

Address correspondence to Eve A. Kerr, M.D., M.P.H., Center for Practice Management and Outcomes Research, Veterans Affairs Ann Arbor Healthcare System, Ann Arbor, MI and the Department of Internal Medicine, University of Michigan Medical School, Ann Arbor, MI. Timothy P. Hofer, M.D., M.S., and Rodney A. Hayward, M.D., are with the Center for Practice Management and Outcomes Research, Veterans Affairs Ann Arbor Healthcare System, Ann Arbor, MI and also with the Department of Internal Medicine, University of Michigan Medical School, Ann Arbor, MI. John L. Adams, Ph.D., and Elizabeth A. McGlynn, Ph.D., are with RAND, Santa Monica, CA. Mary M. Hogan, Ph.D., R.N., is with the Center for Practice Management and Outcomes Research, Veterans Affairs Ann Arbor Healthcare System, Ann Arbor, MI. Steven M. Asch, M.D., M.P.H., is with the Veterans Affairs Greater Los Angeles Health Care System, Los Angeles, CA; and also with Department of Medicine, Geffen School of Medicine at UCLA, Los Angeles, CA and with RAND, Santa Monica, CA.

He found that, while all explicit criteria were infrequently met for any condition, care was rarely judged as adequate by physicians when 50 percent or fewer of explicit criteria were met (Brook and Appel 1973; Donabedian 1985). Subsequently a number of studies have found both evidence of agreement between explicit and implicit measures (Greenfield et al. 1978; Hulka et al. 1979; Rubenstein et al. 1990; Ashton et al. 1999) as well as lack of agreement (Weingart et al. 2002).

However, these studies never compared more than two independent measurement systems and in assessing the level of agreement did not account for the imprecision that often plagues measurement of quality, as it also does many other areas of clinical science. Further, the implicit reviews in the above studies were often conducted by the same physicians who developed the explicit standards used in the explicit reviews, thus diminishing the independence of the two measurement systems (Brook and Appel 1973; Hulka et al. 1979). The studies were also conducted largely in hospitalized patients with a limited set of conditions—none examined care longitudinally or across the continuum of inpatient care and outpatient care (Greenfield et al. 1978; Rubenstein et al. 1990).

Furthermore, explicit measurement systems have improved substantially over the intervening years so there are now two major types. The first type, focused explicit review systems, rely upon explicit assessments of quality for a limited set of supported and feasible measures to evaluate care (National Committee for Quality Assurance [NCQA] 2005). Examples of focused systems include the Health Plan Employer Data and Information Set (HEDIS) (NCQA 2005) and the Department of Veterans Affairs' (VA) chart-review based performance measurement system, the External Peer Review Program (EPRP) (Kizer et al. 2000; Jha et al. 2003; Perlin, Kolodner, and Roswell 2004). While limited in scope, focused systems have the advantages of being concrete, easily understood and generally actionable at the level of the individual measure. Second, at the other end of the spectrum are more global explicit approaches for quality assessment, such as RAND's QA Tools system (McGlynn et al. 2003; Kerr et al. 2004). Global measurement systems use a broader set of quality measures for a much larger number of conditions, and results are reported as summary scores (e.g., for chronic disease or preventive care), thereby making them less effective at informing quality improvement for any individual measure. However, global systems could potentially spark broader systematic change, are more difficult to game, and could be particularly useful for informing contracting decisions and pay-for-performance initiatives because they appear better to represent overall care.

The method of implicit review for quality assessment has also been refined. Structured physician implicit review has been used as a standard tool on many landmark research studies on the quality of care and on medical errors (Rubenstein et al. 1990; Brennan et al. 1991; Rubin et al. 1992). While the main criticism of implicit peer review has been the relatively low inter-rater reliability (Rubin et al. 1992; Ashton et al. 1999), the method has been shown to have moderately high reliability in some cases (Rubenstein et al. 1990; Hofer et al. 2004). Further, more recently developed statistical methods can adjust for these moderate levels of reliability, thus allowing more reliable profiling using this method (Hofer et al. 2004).

Therefore, as demands intensify that health care organizations increase the level of performance monitoring and forge ahead into the world of pay-for-performance, it is critical to ask whether the tools employed to assess quality are capable of supplying the required information about performance. First, is there evidence of the validity of a measurable construct representing overall quality as reflected in agreement between different measurement systems (explicit global, explicit focused, and implicit) across outpatient and inpatient settings? Second, does agreement across the measurement systems differ by discrete medical areas or conditions? Finally, can the assessments of quality be improved by employing newer statistical methods that provide more precise quality estimates? Specifically, does accounting for the measurement error in each of these independent quality measurement systems improve agreement between the systems? To answer these questions, we evaluated technical process quality among 621 patients in 26 VA health care facilities using three different measurement systems: focused explicit, global explicit, and structured implicit review.

METHODS

Specific Instrument Selection, Development, and Testing

The three different measurement systems used in this study all evaluated the processes of medical care rather than the outcomes (Appendix 1). The focused and global explicit instruments were selected on the basis of their comparable scope and widespread use in operational or research settings (McGlynn et al. 2003; Perlin, Kolodner, and Roswell 2004). The focused explicit review instrument modified from the 2001 version of the VA EPRP instrument, measured 38 processes of care, and the global explicit review instrument, modified from RAND QA Tools, measured up to 372 processes of care (Table 1).

Table 1: Description of the Three Quality Assessment Systems

	Structured Implicit		*Focused Explicit*		*Global Explicit*	
Dimension	*No. of Domains**	*Mean Score (SD)*†	*No. of Measures*	*Mean Score (SD)*‡	*No. of Indicators*	*Mean Score (SD)*‡
Prevention	7	80 (11)	14	65 (12)	37	65 (6)
Specific chronic conditions						
Chronic obstructive pulmonary disease	7	75 (9)	5	62 (16)	20	69 (0)
Diabetes	7	60 (13)	8	76 (6)	13	71 (3)
Hypertension	5	66 (16)	3	94 (4)	26	97 (0)
Ischemic heart disease	0	—	5	—	37	—
Heart failure	0	—	3	—	36	—
Alcohol	0	—	0	—	5	—
Asthma	0	—	0	—	25	—
Atrial fibrillation	0	—	0	—	10	—
Benign prostatic hypertrophy	0	—	0	—	4	—
Cancer pain/palliation	0	—	0	—	3	—
Cerebrovascular disease	0	—	0	—	10	—
Colorectal cancer	0	—	0	—	12	—
Depression	0	—	0	—	14	—
Dyspepsia/peptic ulcer disease	0	—	0	—	8	—
Hyperlipidemia	0	—	0	—	7	—
Osteoarthritis	0	—	0	—	3	—
Prostate cancer	0	—	0	—	6	—
All chronic (total)	5	63 (16)	24	81 (6)	239	74 (1)
Acute (total)	4	83 (1)	NA	NA	100	58 (0)
Overall (total)	5	65 (15)	38	73 (9)	372	69 (4)

*Specific domains of care for structured implicit review included assessment of the initial presentation of the condition (if it occurred during this period), the assessment and monitoring of the course of the condition, the treatment of signs or symptoms, exacerbations or complications, and follow-up and the overall quality for the condition or set of conditions. For prevention, ratings were provided for screening, immunizations, prevention-related counseling, and overall preventive care.

†Ratings for implicit review were based on a six-point scale ("very good," "good," "adequate," "borderline poor," "poor," and "very poor"). In order to present a score for implicit review comparable to the score calculated for the explicit instruments, we estimated the probability for each patient or facility of receiving an overall rating of adequate or better.

‡Facility mean explicit scores for both the global and explicit systems represent the proportion of eligible processes of care that were received by each patient, aggregated at the level of the site of care ($N = 26$).

A structured implicit review instrument was developed for this project (Hofer et al. 2004) based on methods employed in previously published studies (Rubenstein et al. 1990; Hayward, McMahon, and Bernard 1993; Hayward et al. 1993) (see Appendix 1). The structured implicit review instrument

required trained physician reviewers to identify and rate process quality for three specific chronic conditions contained in both explicit systems (hypertension, diabetes, and chronic obstructive pulmonary disease [COPD]), other chronic conditions, acute conditions, preventive care, and overall care. Physicians were instructed to base their assessments on whether appropriate processes of care were delivered; that is, whether the care provided care was likely to enhance outcomes. Care was rated within domains (e.g., assessment, treatment, follow-up, and overall care) using a six-point scale ("very good," "good," "adequate," "borderline poor," "poor," "very poor").

Inter-Rater Reliability

For the explicit instruments the inter-rater reliability of chart abstraction, as documented in prior studies and reports, is generally 0.9 or higher (McGlynn et al. 2003). For the structured implicit reviews our reliabilities were comparable or higher to those obtained in prior studies and ranged from 0.2 to 0.5 (Hofer et al. 2004).

Sampling

We constructed a stratified random sample from 26 sites of care (medical centers and clinics) in 12 health care systems located within two VA regional networks, one in the Midwest and one in the West. We sampled 621 patients (out of 106,576) who had at least two primary care outpatient visits during each of the two study years (10/1/97 to 9/30/99), over-sampling diabetes and COPD to obtain a minimum of 175 cases for each condition while minimizing the design effect.

Abstraction

Two separate teams of four professional nurse reviewers experienced in using each instrument completed the focused and global explicit reviews, after receiving project specific training. Implicit reviews were conducted by 12 board-certified internal medicine physicians. The average time per review was 35 minutes for the focused explicit review, 164 minutes for the global explicit review, and 101 minutes for the implicit review.

Score Construction

From the abstracted data we constructed an unadjusted score for the explicit quality measures as the proportion of eligible processes of care that were received by each patient. These patient-level scores were then aggregated to

produce an average score at the level of the site of care (an average of patient averages). We present the scores as proportions ranging from 0 to 100 percent (Table 1).

We also constructed scores for both the implicit review and the two explicit review methods that took into account the major sources of measurement error inherent in each instrument. Specifically, we used multilevel ordinal and logistic regression models to construct scores at the patient and site level that were adjusted for measurement error. For the implicit review the major source of measurement error is between independent reviews of a patient record, or the inter-rater reliability. We estimated the level of reliability of the implicit reviews by conducting multiple reviews within a sample of patients (Hofer et al. 2004). By using these repeated measures at the level of the independent review (the lowest level in the multilevel model), we were able to estimate and remove this source of measurement error. In addition, we simultaneously adjusted for differences in the overall mean rating for each rater, thus removing any differences in how strict or lenient particular raters were in assessing quality.

For the explicit reviews, the major source of measurement error is the variation between the number of items that comprise the score and the fact that some patients were eligible for only a small number of indicators. For example, a score in which one out of two processes is met (1/2 or 50 percent) is less precisely estimated than a score in which seven out of 14 are met (7/14 or 50 percent). By adjusting for this source of error in the multilevel models, we produced better estimates of the true underlying score measured by an explicit instrument (in terms of mean squared error) (Bryk and Raudenbush 1992; Goldstein 1995; Snijders and Bosker 1999). Again we also adjusted for any differences in scoring across nurse abstractors. See Appendix 2 for further details of these analytical methods.

The resultant explicit and implicit quality scores were transformed onto a probability scale for purposes of presentation. In order to present a "pass rate" for implicit review comparable to the pass rate calculated for the explicit instruments, we estimated the probability for each patient or facility of receiving an overall rating of adequate or better.

Comparison of Scores

For comparison with prior studies that have compared explicit and implicit review (Rubenstein et al. 1990) we first present a table of the implicit review ratings (collapsed into four categories), and the mean explicit scores received

by patients falling into each implicit rating (Table 2). These mean scores are aggregated across all patients and sites, unadjusted for measurement error, and compared with implicit review ratings using a traditional analysis of variance. We next present correlations between the different quality measurement methods unadjusted and adjusted for measurement error (as described above), at the patient and site levels. These correlations are weighted to represent the sample design in the selection of the patient records for review. All adjusted analyses were performed using *Stata* using the GLAM program (Generalized Linear Latent and Mixed Models) (Rabe-Hesketh, Skrondal, and Pickles 2005).

RESULTS

Study Sample

Most patients in the sample were men (97 percent) and their mean age was 62.4 (SD = 11.7). For the 2-year time period from October 1, 1997 to September 30, 1999, 25.3 percent of patients were hospitalized at least once, with a mean of 0.56 hospitalizations per person (SD = 1.5). Patients had a median of nine primary care, cardiology, endocrine, or pulmonary outpatient visits during the 2 years.

Score Comparisons

Table 2 shows that for the unadjusted patient-level scores there is a clear monotonic relationship for most conditions, so that when quality was rated more highly by the implicit system it was also rated more highly by the explicit systems. For example, when implicit review scored preventive care as poor or very poor, the focused explicit review and global explicit review ratings were 33.2 and 42.6, respectively, but were higher at 72.0 and 73.6, respectively ($p < .001$), when the implicit rating was good or very good. In a few cases the monotonic relationship was less apparent (e.g., for COPD in the global explicit system) or absent (e.g., for hypertension in the focused explicit system).

Tables 3a and 3b show the correlations between the three systems unadjusted and adjusted for measurement error in the explicit and implicit scores. The unadjusted correlations between measurement methods for most of patient-level scores (e.g., correlation between individual patient's overall quality score using the implicit tool compared with using the global tool) generally fall into the small to medium effect size range (0.1–0.4) (Cohen 1988; Murphy and Myors 1998). Further, we find that the correlations between the

Table 2: Mean Explicit Quality Scores by Implicit Review System Rating Categories*

	Explicit Review Scores (SD) by Implicit Review Categories				
Condition and Explicit System Type	*Very Poor/ Poor*	*Borderline Poor*	*Adequate*	*Good/ Very Good*	*p-Value*
Hypertension, $N = 500$	$N = 57$	$N = 130$	$N = 168$	$N = 145$	
Focused explicit score	89.2 (22.7)	87.5 (23.1)	89.1 (21.4)	89.6 (21.0)	.91
Global explicit score	81.7 (35.6)	89.8 (22.6)	92.9 (23.6)	94.3 (18.9)	.014
COPD, $N = 167$	$N = 13$	$N = 44$	$N = 71$	$N = 39$	
Focused explicit score	36.7 (39.9)	45.3 (30.8)	58.0 (33.6)	65.4 (36.5)	.032
Global explicit score	67.4 (40.9)	63.9 (37.9)	67.9 (36.4)	74.0 (30.5)	.72
Diabetes, $N = 258$	$N = 34$	$N = 80$	$N = 68$	$N = 76$	
Focused explicit score	62.7 (26.3)	72.5 (21.5)	74.7 (22.4)	81.3 (16.8)	.001
Global explicit score	62.4 (22.1)	65.8 (21.9)	65.6 (22.4)	76.1 (19.0)	.003
Prevention, $N = 621$	$N = 23$	$N = 119$	$N = 231$	$N = 248$	
Focused explicit score	33.2 (21.9)	51.4 (23.4)	62.6 (24.9)	72.1 (23.9)	$< .001$
Global explicit score	42.6 (21.8)	53.3 (18.5)	64.9 (19.4)	73.6 (17.3)	$< .001$
All chronic, $N = 620$	$N = 77$	$N = 158$	$N = 199$	$N = 186$	
Focused explicit score	70.3 (26.4)	77.1 (22.7)	78.0 (24.5)	83.2 (21.7)	.001
Global explicit score	68.5 (21.1)	70.7 (21.8)	72.3 (24.6)	71.7 (24.0)	.67
Overall, $N = 621$	$N = 70$	$N = 164$	$N = 211$	$N = 176$	
Focused explicit score	60.4 (23.1)	67.2 (23.1)	68.5 (22.9)	74.6 (19.2)	$< .001$
Global explicit score	61.9 (18.2)	64.1 (16.4)	68.4 (14.9)	70.0 (15.0)	$< .001$

*For each condition or clinical area, mean quality explicit system scores were calculated for patients whose care was rated by implicit reviewers as very poor/poor, borderline poor, adequate, or good/very good. These patient-level scores are unadjusted for **measurement error** and mean scores across implicit rating categories are compared using traditional analysis of variance.

adjusted patient-level scores are about 20–50 percent higher than the correlations between the unadjusted patient-level scores. The unadjusted and adjusted site-level correlations (e.g., correlation between a site's overall quality score using the implicit tool compared with using the global review) are in turn larger than the correlations at the patient level across all three systems, and most represent large effect sizes for the three summary scores (overall, all chronic, and prevention) between the global and implicit systems (Table 3a). Site-level prevention scores showed particularly high levels of convergence with correlations in the 0.7–0.9 range (see Table 3a). Any measurement of quality may contain components that are attributable to how an organization as a whole functions to provide care and other components that are specific to a particular patient and his/her interaction with the provider. The relatively higher correlations at the site-level suggest that there is more agreement

Table 3a: Correlations between Each Set of Measurement Systems for Overall, All Chronic, and Preventive Care

	*Correlation between Each Set of Review Systems by Condition Category**								
	Overall			*All Chronic*			*Prevention*		
Level of Correlation	*Implicit /Global*	*Implicit /Focused*	*Global /Focused*	*Implicit /Global*	*Implicit /Focused*	*Global /Focused*	*Implicit /Global*	*Implicit /Focused*	*Global /Focused*
Patient, unadjusted	**0.19**	**0.17**	**0.37**	0.02	**0.16**	0.16	**0.45**	**0.37**	**0.47**
Patient, adjusted	**0.26**	**0.22**	**0.51**	0.12	**0.24**	**0.22**	**0.75**	**0.57**	0.68
Site, unadjusted	0.29	0.15	**0.67**	0.41	0.36	0.17	**0.83**	**0.67**	**0.73**
Site, adjusted	0.48	0.29	**0.75**	**0.59**	0.49	0.27	**0.89**	**0.67**	**0.73**

*Correlation coefficients are derived from the multilevel models that are adjusted for the **major sources of measurement error in the explicit and implicit measurement systems**. Patient-level correlations include effect of patient and site. Correlations in bold are significant at $<.05$ when a Bonferroni correction for multiple comparisons is applied.

between the quality measurement systems about the performance on the components of quality at the organization level than for the patient-specific components of quality.

For the disease specific scores (hypertension, COPD and diabetes—Table 3b), the diabetes score correlations seem to be both substantial and

Table 3b: Correlations between Each Set of Measurement Systems for Hypertension, Diabetes, and Chronic Obstructive Pulmonary Disease (COPD)

	*Correlation between Each Set of Instruments by Condition Category**								
	Hypertension			*Diabetes*			*COPD*		
Level of Correlation	*Implicit /Global*	*Implicit /Focused*	*Global /Focused*	*Implicit /Global*	*Implicit /Focused*	*Global /Focused*	*Implicit /Global*	*Implicit /Focused*	*Global /Focused*
Patient, unadjusted	0.13	0.02	0.09	**0.24**	0.22	**0.38**	0.10	**0.37**	0.12
Patient, adjusted	0.10	0.02	0.01	**0.45**	**0.43**	**0.56**	0.25	**0.46**	0.12
Site, unadjusted	0.44	0.06	0.04	0.47	0.53	0.49	**0.67**	**0.61**	**0.60**
Site, adjusted	—	0.02	—	**0.61**	**0.69**	**0.63**	—	0.55	—

*Correlation coefficients are derived from the multilevel models that are adjusted for the **major sources of measurement error in the explicit and implicit measurement systems**. Patient-level correlations include effect of patient and site. Correlations in bold are significant at $<.05$ when a Bonferroni correction for multiple comparisons is applied. As the global site-level scores for hypertension and COPD have no variance after adjustment for measurement error and reviewer effects, there are no correlations reported with the adjusted global site-level scores.

similar across the three methods of assessment. The COPD scores followed the general pattern as the other scores but the correlations were weaker and less consistent across the systems. Site-level scores obtained using the global tool for COPD and hypertension had no variance after adjustment for the uncertainty in the patient-level scores and the reviewer effects, therefore no correlations are reported for these scores (Table 3b). Further, the hypertension measures had fairly high amounts of variation for the scores at the patient level but little evidence of correlation between the scores obtained by the different measurement methods (small effect sizes and all correlations not significant). Although the global tool used for this study had 26 indicators for hypertension, 18 of the 26 indicators related only to newly diagnosed cases and 21 of the indicators applied to fewer than 5 percent of patients identified with hypertension. The median number of indicators applicable per subject was 2 (range 1–18). The focused tool had three indicators that applied to all of the patients (diet counseling, exercise counseling, and annual blood pressure measurement), and these indicators were met most of the time for all patients. The two conditions which had the lowest correlations between the methods (hypertension and COPD) were also those for which there were the fewest number of applicable indicators comprising the score.

CONCLUSION

To our knowledge, this is the largest and most comprehensive study examining agreement in care quality between such varied performance assessment methods. The correlations between the instruments, despite the differences in the measurement systems, support the conclusion that an underlying quality construct exists for most of the conditions and clinical areas that we examined. The global explicit system had particularly high convergence with the implicit system for summary scores and both explicit systems had moderately high convergence with implicit assessment of quality for all discrete conditions, except hypertension.

We also demonstrated that the correlations between quality assessment systems are substantially higher when we adjust for the measurement error inherent in each of the implicit and explicit systems. Many measurements in clinical science are imprecise, and we use analytical approaches such as sensitivity, prior probability, and positive predictive value to anchor meaning to results of such measurements (e.g., abnormal abdominal examinations, positive cardiac exercise stress tests, and prostate specific antigen values). Quality

measurements are similarly imprecise, and the imprecision increases for summary measures. Yet there is a compelling policy need for relatively simple summary measures of quality in order to implement pay for performance strategies to improve care. Therefore, methods that adjust for some of the imprecision in these measurements are essential and also helpful when trying to understand the relationships between different measures of quality.

Although there was a high degree of convergence across the measurement systems for summary measures of quality, there was also substantial variation in the amount of agreement seen between measurement approaches for quality ratings for specific conditions. For example, for preventive and diabetes care, the agreement across the three instruments was particularly high (site-level correlations ranging from 0.61 to 0.89). These content areas are distinguished by having relatively large amounts of published evidence and guidelines supporting specific interventions. Thus the two explicit tools share very similar process indicators and the physician reviewers have a better and more straightforward evidence-base upon which to form their judgments.

If availability of evidence contributed to the high correlations for diabetes and prevention, then why did we not see a similar trend for hypertension, which had both low agreement between measures and poor performance for the explicit scores? The hypertension explicit scores for both the global and focused methods were notable for having only two to three applicable indicators for most patients and very high pass rates (50 percent of the patients had a perfect score). This combination of few measurements per person and a ceiling effect is likely to have reduced the amount of usable information in the explicit scores to a point where they did not adequately measure the underlying construct in this population. Although a similar problem could have also led to low concordance for COPD measures, it seems more likely that the much smaller and more nebulous evidence base in COPD care resulted in a more inherent difficulty achieving convergence between physician assessments and different sets of explicit indicators (Hofer et al. 2004).

This study highlights several important factors about why different explicit systems and implicit review may agree with one another, even when the process measures used to assess quality in each system are quite different. For example, it is possible that explicit systems' process standards do not capture the bulk of important decision making, but providers who do well on aspects of care measured by the explicit tools also do well on aspects of care captured by implicit review but not by the explicit tools. On the other hand, implicit reviewers' judgment as to what defines a quality problem may have been

influenced (and narrowed) by widespread educational efforts focused around existing explicit measures. Finally, we must consider that the performance of explicit systems is dependent on the population. Some process measures may adequately assess quality of care among patients with mild but not severe disease, and agreement between different systems could therefore be dependent on the patient population considered.

Our study adds to the existing literature in this area in several important ways. First, we examined quality using three distinct measures of quality, two of which are in current use for operations or research purposes. Unlike many previous studies, each measurement system was developed independently, and used unique reviewers. Second, we used statistical methods not available at the time of many of the earlier studies to adjust for the measurement error in each of the systems, allowing us to produce more accurate scores and correlations. Third, we examined convergence at both the patient and site levels to inform our findings. This is important because profiling generally occurs at the site level and these methods may be better suited to evaluating quality for an entity rather than an individual patient. Finally, we examined agreement not only for discrete conditions but also for overall quality of care in inpatient and outpatient settings and across the continuum of care delivery (e.g., prevention, chronic care, overall).

Despite its methodological strengths, our study has some limitations. We reviewed care among persons who had chronic conditions, and who used care frequently, in a system that delivers high quality of care (Jha et al. 2003; Asch et al. 2004; Kerr et al. 2004). Our conclusions could differ if these tools were applied to healthier populations or in different health care settings. Second, all of the measures that comprised the explicit scores were weighted equally. If implicit reviewers take importance weights into account when rating care (i.e., by judging medication intensification for uncontrolled hypertension as more important than checking an urinalysis), then some of the lack of convergence may be due to equal weighting in the explicit systems. Failure to weight could also result in a systematic underestimate of differences in quality if standards that have a small potential impact on outcomes are more commonly met than standards that have a high potential impact. Although Ashton et al. (1999) found little effect of importance weighting on ratings of readiness-for-discharge scores, future research will need to examine whether weighting of indicators by importance improves the correlations between implicit and explicit systems for overall quality of care. This study assessed only technical aspects of care and was limited to factors usually noted in the medical record. Therefore, we were unable to assess quality of interpersonal care, such as

effective communication, which could also affect outcomes. While we found correlations in quality scores at the site level to be moderately high in many instances, our sampling frame did not allow us to examine correlations at individual physician levels. Our previous research suggests, however, that variance in scores may be lower at the physician level than at the site level. Thus, a larger number of records per physician may be necessary to produce a reliable score, and consideration of measurement error would be even more critical (Hofer et al. 1999; Krein et al. 2002).

Our findings also clarify some of the advantages and disadvantages of these three main approaches to measuring quality. A focused explicit tool produces scores that are easier to interpret and more easily guide targeted quality improvement efforts. Summary focused scores are correlated with overall implicit review, and our previous research has shown that organizations that employ them may experience some benefit in unmeasured but related areas of quality (Asch et al. 2004). But a focused set of indicators puts a spotlight on only a few processes of care, opening up the possibility of provider "gaming" and potential misallocation of resources if other processes are more closely related to outcomes. If the goal is to distinguish overall quality performance between facilities or to guide more systemic quality improvement efforts, then the global explicit and implicit approaches are designed to be more comprehensive and difficult to game. Our research shows that they are somewhat better correlated overall with each other than focused systems. The cost of measurement systems (which we did not directly measure) may tip the balance in deciding which system to employ. We found that the global system required substantially longer chart review times than the focused one, and the implicit review system required more expensive physician reviewers than the other two. How those findings will translate into measurement costs, especially as more of the needed information is available electronically, awaits future research.

Our findings also raise a note of caution. In contrast to the systems used in our study, many commercial profiling systems rely heavily on administrative rather than medical record data, and measures are often developed as much on the basis of feasibility as on validity and reliability. Previous research has shown that assessments of quality can be very different (and higher) when using only measures applicable to administrative data versus using a much broader set of clinically detailed measures (MacLean et al. 2006). By looking for quick and inexpensive solutions, the developers of profiling systems that rely only on currently available administrative data may be producing profiles, but we cannot know whether these profiles truly reflect "quality" (versus utilization, access, or another construct).

Instead, our findings lend credence to profiling approaches that use clinically detailed data, such as the approach used in VA, and possibly to expanding those approaches to cover more conditions in order to capture a larger slice of "overall" quality (as VA has already begun to do). Clinically detailed data need not always be retrieved from medical record review but could be extracted from correctly constructed electronic health records (Kerr et al. 2003; Kupersmith et al. 2007). Our results suggest that a measurable construct that represents health care "quality" does indeed exist and can be discerned with quality assessment systems and appears to be strongest in areas with a well-developed evidence base and widespread agreement on management and treatment. However, before we can reward or punish health care providers based on their "quality," we must understand whether the measures used to assess quality reflect reliable and meaningful estimates of providers' performance.

ACKNOWLEDGMENTS

The first two authors contributed equally to this manuscript. We gratefully acknowledge the role of the many nurse and physician medical record reviewers. We also thank the members of our study advisory committee, especially Dr. Thomas Craig, who gave us valuable advice on the conduct of the study. Results were presented, in part, at the VA HSR&D National Meeting, February 2005; and the SGIM National Meeting, May 2005.

This study was supported by the Department of Veterans Affairs Health Services Research and Development Service (IIR#98-103). Additional support was provided by the Michigan Diabetes Research and Training Center Grant P60DK-20572 from the NIDDK of the National Institutes of Health.

Disclosures: None.

Disclaimers: The opinions expressed are those of the authors and do not necessarily represent those of the funding agencies, RAND, or the Department of Veterans Affairs.

REFERENCES

Asch, S. M., E. A. McGlynn, M. M. Hogan, R. A. Hayward, P. Shekelle, L. Rubenstein, J. Keesey, J. Adams, and E. A. Kerr. 2004. "Comparison of Quality of Care for Patients in the Veterans Health Administration and Patients in a National Sample." *Annals of Internal Medicine* 141 (12): 938–45.

Ashton, C. M., D. H. Kuykendall, M. L. Johnson, and N. P. Wray. 1999. "An Empirical Assessment of the Validity of Explicit and Implicit Process-of-Care Criteria for Quality Assessment." *Medical Care* 37 (8): 798–808.

Brennan, T. A., L. L. Leape, N. M. Laird, L. Hebert, A. R. Localio, A. G. Lawthers, J. P. Newhouse, P. C. Weiler, and H. H. Hiatt. 1991. "Incidence of Adverse Events and Negligence in Hospitalized Patients." *New England Journal of Medicine* 324 (6): 370–6.

Brook, R. H., and F. A. Appel. 1973. "Quality-of-Care Assessment: Choosing a Method for Peer Review." *New England Journal of Medicine* 288 (25): 1323–9.

Bryk, A. S., and S. W. Raudenbush. 1992. *Hierarchical Linear Models: Applications and Data Analysis Methods.* Newbury Park, CA: Sage Publications.

Cohen, J. 1988. *Statistical Power Analysis of the Behavioral Sciences*, 2d Edition. Hillsdale, NJ: Lawrence Erlbaum Associates.

Donabedian, A. 1985. *The Methods and Findings of Quality Assessment and Monitoring: An Illustrated Analysis. Explorations in Quality Assessment and Monitoring.* Ann Arbor, MI: Health Administration Press.

Goldstein, H. 1995. *Multilevel Statistical Models*, 2d edition. New York: Halstead Press.

Greenfield, S., S. H. Kaplan, G. A. Goldbert, M. A. Nadler, and R. Deigh-Hewerston. 1978. "Physician Preference for Criteria Mapping in Medical Care Evaluation." *Journal of Family Practice* 6 (5): 1079–86.

Hayward, R. A., A. M. Bernard, J. S. Rosevear, J. E. Anderson, and L. F. McMahon Jr. 1993. "An Evaluation of Generic Screens for Poor Quality of Hospital Care on a General Medicine Service." *Medical Care* 31 (5): 394–402.

Hayward, R. A., L. F. McMahon, and A. M. Bernard. 1993. "Evaluating the Care of General Medicine Inpatients: How Good Is Implicit Review?" *Annals of Internal Medicine* 118 (7): 550–6.

Hofer, T. P., S. M. Asch, R. A. Hayward, L. V. Rubenstein, M. M. Hogan, J. Adams, and E. A. Kerr. 2004. "Profiling Quality of Care: Is There a Role for Peer Review?" *BMC Health Services Research* 4 (1): 9.

Hofer, T. P., R. A. Hayward, S. Greenfield, E. H. Wagner, S. H. Kaplan, and W. G. Manning. 1999. "The Unreliability of Individual Physician 'Report Cards' for Assessing the Costs and Quality of Care of a Chronic Disease." *Journal of the American Medical Association* 281 (22): 2098–105.

Hulka, B. S., F. J. Romm, G. R. Parkerson Jr., I. T. Russell, N. E. Clapp, and F. S. Johnson. 1979. "Peer Review in Ambulatory Care: Use of Explicit Criteria and Implicit Judgments." *Medical Care* 17 (3 suppl): i–vi, 1–73.

Jha, A. K., J. B. Perlin, K. W. Kizer, and R. A. Dudley. 2003. "Effect of the Transformation of the Veterans Affairs Health Care System on the Quality of Care." *New England Journal of Medicine* 348 (22): 2218–27.

Kerr, E. A., E. A. McGlynn, J. Adams, J. Keesey, and S. M. Asch. 2004. "Profiling the Quality of Care in Twelve Communities: Results from the CQI Study." *Health Affairs (Millwood)* 23 (3): 247–56.

Kerr, E. A., D. M. Smith, M. H. Hogan, T. P. Hofer, S. L. Krein, M. Berman, and R. A. Hayward. 2003. "Building a Better Quality Measure: Are Some Patients with 'Poor Quality' Actually Getting Good Care?" *Medical Care* 41 (10): 1173–82.

Kizer, K. W., J. G. Demakis, and J. R. Feussner. 2000. "*Reinventing VA* Health Care: Systematizing Quality Improvement and Quality Innovation." *Medical Care* 38 (6): I7–I16.

Krein, S. L., T. P. Hofer, E. A. Kerr, and R. A. Hayward. 2002. "Whom Should We Profile? Examining Diabetes Care Practice Variation among Primary Care Providers, Provider Groups, and Health Care Facilities." *Health Services Research* 37 (5): 1159–80.

Kupersmith, J., J. Francis, E. A. Kerr, S. Krien, L. Pogach, R. M. Kolodner, and J. B. Perlin. 2007. "Advancing Evidence-Based Care in Diabetes through Health Information Technology: Lessons from the Veterans Health Administration." *Health Affairs* 26 (2): W156–W158.

MacLean, C. H., R. Louie, P. Shekelle, C. P. Roth, D. Saliba, T. Higashi, J. Adams, J. T. Chang, C. J. Kamberg, D. H. Solomon, R. T. Young, and N. S. Wenger. 2006. "Comparison of Administrative Data and Medical Records to Measure the Quality of Medical Care Provided to Vulnerable Older Patients." *Medical Care* 44 (2): 141–8.

McGlynn, E. A., S. M. Asch, J. Adams, J. Keesey, J. Hicks, A. DeCristofaro, and E. A. Kerr. 2003. "The Quality of Health Care Delivered to Adults in the United States." *New England Journal of Medicine* 348 (26): 2635–45.

Murphy, K. R., and B. Myors. 1998. *Statistical Power Analysis.* Hillsdale, NJ: Lawrence Erlbaum Association.

National Committee for Quality Assurance (NCQA). 2005. *HEDIS 2006, Narrative: What's in It and Why It Matters*, 1. National Committee for Quality Assurance, pp. 1–88. Washington, DC: NCQA.

Perlin, J. B., R. M. Kolodner, and R. H. Roswell. 2004. "The Veterans Health Administration: Quality, Value, Accountability, and Information as Transforming Strategies for Patient-Centered Care." *American Journal of Managed Care* 10 (11 part 2): 828–36.

Rabe-Hesketh, S., A. Skrondal, and A. Pickles. 2005. "Maximum Likelihood Estimation of Limited and Discrete Dependant Variable Models with Nested Random Effects." *Journal of Econometrics* 128 (2): 301–23.

Rubenstein, L. V., K. L. Kahn, E. J. Reinsisch, M. J. Sherwood, W. H. Rogers, C. Kamberg, D. Draper, and R. H. Brook. 1990. "Changes in Quality of Care for Five Diseases Measured by Implicit Review 1981 to 1986." *Journal of the American Medical Association* 264 (15): 1974–9.

Rubin, H. R., W. H. Rogers, K. L. Kahn, L. V. Rubenstein, and R. H. Brook. 1992. "Watching the Doctor-Watchers: How Well Do Peer Review Organization Methods Detect Hospital Care Quality Problems?" *Journal of the American Medical Association* 267 (17): 2349–54.

Snijders, T. A. B., and R. J. Bosker. 1999. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modelings.* Thousand Oaks, CA: Sage Publications.

Weingart, S. N., R. B. Davis, R. H. Palmer, M. Cahalane, M. B. Hamel, K. Mukamal, R. S. Phillips, D. T. Davies Jr., and L. I. Iezzoni. 2002. "Discrepancies between Explicit and Implicit Review: Physician and Nurse Assessments of Complications and Quality." *Health Services Research* 37 (2): 483–98.

SUPPLEMENTARY MATERIAL

The following supplementary material for this article is available:

Appendix 1: Focused Explicit System.

Appendix 2: Further Explanation of Analyses.

This material is available as part of the online article from: http://www.blackwell-synergy.com/doi/abs/10.1111/j.1475-6773.2007.00730.x (this link will take you to the article abstract).

Please note: Blackwell Publishing is not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.