Comparison of Vignettes, Standardized Patients, and Chart Abstraction

A Prospective Validation Study of 3 Methods for Measuring Quality

John W. Peabody, MD, PhD
Jeff Luck, MBA, PhD
Peter Glassman, MBBS, MSc
Timothy R. Dresselhaus, MD, MPH
Martin Lee, PhD

Context Better health care quality is a universal goal, yet measuring quality has proven to be difficult and problematic. A central problem has been isolating physician practices from other effects of the health care system.

Objective To validate clinical vignettes as a method for measuring the competence of physicians and the quality of their actual practice.

Design Prospective trial conducted in 1997 comparing 3 methods for measuring the quality of care for 4 common outpatient conditions: (1) structured reports by standardized patients (SPs), trained actors who presented unannounced to physicians' clinics (the gold standard); (2) abstraction of medical records for those same visits; and (3) physicians' responses to clinical vignettes that exactly corresponded to the SPs' presentations.

Setting Outpatient primary care clinics at 2 Veterans Affairs medical centers.

Participants Ninety-eight (97%) of 101 general internal medicine staff physicians, faculty, and second- and third-year residents consented to be randomized for the study. From this group, 10 physicians at each site were randomly selected for inclusion.

Main Outcome Measures A total of 160 quality scores (8 cases × 20 physicians) were generated for each method using identical explicit criteria based on national guidelines and local expert panels. Scores were defined as the percentage of process criteria correctly met and were compared among the 3 methods.

Results The quality of care, as measured by all 3 methods, ranged from 76.2% (SPs) to 71.0% (vignettes) to 65.6% (chart abstraction). Measuring quality using vignettes consistently produced scores closer to the gold standard of SP scores than using chart abstraction. This pattern was robust when the scores were disaggregated by the 4 conditions ($P<.001$ to $<.05$), by case complexity ($P<.001$), by site ($P<.001$), and by level of physician training (P values from $<.001$ to $<.05$). The pattern persisted, although less dominantly, when we assessed the component domains of the clinical encounter—history, physical examination, diagnosis, and treatment. Vignettes were responsive to expected directions of variation in quality between sites and levels of training. The vignette responses did not appear to be sensitive to physicians' having seen an SP presenting with the same case.

Conclusions Our data indicate that quality of health care can be measured in an outpatient setting by using clinical vignettes. Vignettes appear to be a valid and comprehensive method that directly focuses on the process of care provided in actual clinical practice. Vignettes show promise as an inexpensive case-mix adjusted method for measuring the quality of care provided by a group of physicians.

JAMA. 2000;283:1715-1722 www.jama.com

ASSESSING QUALITY MUST ULTIMATELY rely on measures that are inexpensive, reliable, and able to adequately control for case-mix variation.[1-3] Health outcome measures, although a direct assessment of health status, also reflect a spectrum of confounding events such as comorbidities and the socioeconomic determinants of health—factors that are generally beyond the control of a physician's daily practice. As a result, process measures of quality are increasingly being used.[4,5] If linkages between the provision of care and better health status have been firmly established, there are substantiated benefits to measuring process over measuring outcomes.[6] Processes can be measured more frequently than outcomes (eg, a death or complication), do not require a lengthy interval to become manifest,[7] and are generally less expensive to monitor.[8,9]

The most common method for measuring process, which includes both the competence of the clinician and what the clinician actually does, is chart abstraction.[10-12] Chart abstraction primarily has been validated in the inpatient setting, where care tends to be exten-

For editorial comment see p 1740.

Author Affiliations are listed at the end of this article.
Corresponding Author and Reprints: John W. Peabody, MD, PhD, San Francisco Veterans Affairs Medical Center, c/o Institute for Global Health, University of California, San Francisco, Suite 508, 74 New Montgomery St, San Francisco, CA 94105 (e-mail: peabody@psg.ucsf.edu).

sively documented and clinical events are more temporally circumscribed.[13] As care has increasingly shifted to the outpatient setting, so has reliance on abstraction of outpatient charts to measure quality of care.[14] Despite increased use of chart abstraction, validity of outpatient process measures has been systematically evaluated in only a few studies,[15,16] and significant problems may exist with chart abstraction in this setting. For example, abstracted chart data may be subject to recording bias because of time constraints on outpatient visits. The usefulness of chart abstraction is further limited because a skilled (and costly) expert must collect the data.[17,18] Perhaps the most important limitation of chart abstraction is that adjustments for case-mix variation are insufficient, thereby limiting direct comparisons of quality of care across different sites or delivery systems.[19-21]

An alternative in the outpatient setting is to directly observe patient-provider interactions; this could be a gold standard if physicians were adequately masked to the measurement method. However, truly double-blind observations, where neither provider nor patient know they are being observed, are obviously not possible for ethical and logistical reasons. An extensive medical-education literature describes the successful use of standardized patients (SPs) as a practical gold standard[22-28] and reports that SPs can capture variation in clinical practice and reproducibly show how individual physician practices vary over time.[23,29,30] However, SPs require even more intrusion into a physician's practice than chart abstraction, and they cannot assess some aspects of physician observation.[31] They are expensive and incur the opportunity cost of time the physician does not spend with "real" patients.

Thus, alternative methods of measuring process are needed.[32] Vignettes or written case simulations have been widely used by educators, demographers, and health service researchers to measure processes in a wide range of practice settings.[33-35] Vignettes are easily administered, less costly, and can be used in all types of clinical practices.[36] Because they control for case mix, vignettes hold promise as a way to assess quality of care among different providers and between organizations that may (or may not) care for different populations of patients in different systems of care.[37-39] But despite the promise of vignettes and their growing use in a variety of settings, little work has been done to validate them.[40,41]

This study was performed to assess whether clinical vignettes are a valid method for measuring process of care compared with actual clinical practice. We used a prospective sample of a group of physicians to compare 3 measurement methods—clinical vignettes, chart abstraction (the standard method), and SPs (the gold standard). Quality scores were generated for 4 common outpatient conditions. The analyses directly compared all 3 methods, controlling for possible design effects of level of training, individual physician effects, site or location disparities, and case severity. We also evaluated quality scores for different domains of clinical care skills—history taking, physical examination, radiologic and laboratory testing, diagnostic accuracy, and clinical treatment or management.

METHODS

Physician Sample

The study was conducted at 2 general internal medicine primary care outpatient clinics located at the West Los Angeles and the San Diego Veterans Affairs medical centers, in California. All primary care staff physicians, faculty, and residents in these clinics except interns were eligible for the study. Ninety-eight of the 101 eligible providers (approximately 97%) consented to see SPs "sometime" during their regularly scheduled clinic hours over the course of the 12-month academic year. We randomly selected 10 physicians at each site to see SPs. All consenting physicians were asked to notify us if they suspected that a patient was an SP. The visits were completed over a 6-month period from February through July, 1997.

Measurement Methods

Each method measured the process of care for 4 common outpatient conditions: low back pain, diabetes mellitus, chronic obstructive pulmonary disease, and coronary artery disease (CAD). Two detailed clinical scenarios (cases) were developed for each of the 4 conditions, 1 simple and 1 complex, for a total of 8 cases. For each case, a physician both saw an SP and completed a vignette. The BOX contains detailed summaries of the simple and complex CAD cases. (Detailed descriptions of the vignettes and the scoring forms are available from the authors.)

Established protocols were followed for SP training and data collection. Educators running medical school SP programs trained the actors for each case. Only experienced actors from the SP teaching program were hired. They were trained to remember and record details of the clinical encounter. After training, the SPs were enrolled unannounced into the primary care clinics and scheduled for walk-in or new-patient visits. Their identities as SPs were not revealed to any of the outpatient staff or the examining physician. Realistic identities, necessary laboratory findings, and radiographs were all simulated. In all, 10 randomly chosen providers at each of the 2 sites saw 8 cases each, for a total of 160 visits. To match the vignettes as closely as possible, the SPs were carefully scripted not to volunteer any information other than the presenting problem.

The SPs completed checklists immediately after their visits. An SP quality score for each visit was generated directly from the checklist responses. Simultaneously, charts from SP visits were retrieved from the clinic. Data were abstracted from the charts by a trained nurse abstractor, generating 160 scores.

Several weeks after SPs had been seen in the clinic, vignettes were given to the same 20 physicians. The vignettes prompted open-ended responses to questions that were arranged in sections to re-create the sequence of a typical patient visit: the presenting prob-

lem, history, physical examination, radiologic or laboratory tests ordered, diagnosis, and treatment plan. Each section began with the presentation of new patient information gained from answers to questions in the previous section. After answering 1 section and moving on to the next, physicians could not return to a previous section to revise their answers. Thus, they could not use the new information to change (and improve) their previous answers. When the vignettes were completed, the responses were scored by the same expert nurse abstractor who performed the chart abstraction, generating another 160 scores. The abstractor, who was masked to physician identity, reviewed each vignette answer sheet and indicated on a scoring form those scoring items the physician had successfully completed. To evaluate whether there was a cuing effect (whether having seen an SP presenting with the same case might cue physicians to recognize features of the vignette), we also administered vignettes to 20 matched, randomly selected physicians who had not seen SPs, generating another 160 scores.

Coronary Artery Disease Scenarios

Case 1. A 65-year-old man, a new patient, comes to the clinic for follow-up of a myocardial infarction (MI) he had 3 months ago. In taking the history, the physician should ascertain that the patient is now free of pain and has no difficulty performing routine activities but continues to smoke although he has normal blood pressure. After the physician records what he or she intends to do in the physical examination, the findings are revealed by the vignettes or by the patient in response to physician questioning, and the physician then is asked what laboratory tests should be ordered (an electrocardiogram and cholesterol test), what the diagnosis is (uncomplicated MI), and how treatment should proceed. The physician should recognize that the MI is recent and associated with reversible risk factors and that the patient needs to be taking aspirin and a β-blocker.

Case 2. A 62-year-old new patient presents with roughly the same story—recent MI with similar risk factors—but in taking this history, the physician should learn that the patient has difficulty with routine activities and easily becomes short of breath since running out of his medication. On examination, the patient is to have slight tachypnea and slightly elevated blood pressure. When this information is revealed by the vignettes or by the patient in response to physician questioning, the physician is expected to order the same tests as in the first case plus a blood chemistry test and a chest radiograph (or schedule an echocardiogram). The electrocardiogram confirms that the patient has had an MI in the past. The physician should recognize that this is an MI complicated by mild heart failure. The physician should evaluate for potential risk factors (again) and prescribe aspirin and an angiotensin-converting enzyme inhibitor.

Scoring Criteria

We conceptualized quality as the comprehensive provision of services in a manner that leads to better outcomes for individuals and populations.[42] Thus, we identified candidate criteria to measure a full range of activities that potentially captured the process of outpatient primary care. Explicit quality criteria for each of the 8 cases were derived from national guidelines. We submitted the candidate criteria to local expert panels of academic and community physicians including both generalists and specialists for the conditions. Based on their recommendations and group consensus, we modified and finalized a master criteria list. Criteria for each case included both necessary care and some care that was either unnecessary or inappropriate for that condition.

Identical criteria were used in each method as explicit items on which to score provider responses for each of the 8 cases. Items felt by experts to be most critical were assigned a weight of 1.0.[43,44] Individual items that experts deemed less important, such as multiple physical examination items that were related to a single clinical construct, were grouped into categories, implicitly assigning them lower weights, typically 0.50 or 0.33. Scores were generated from SP responses to a closed-ended postinterview questionnaire that contained the explicit criteria for each case. Chart abstraction and vignette scores were based on scoring forms that contained the criteria and were completed by a trained nurse abstractor. The raw item scores for each method were aggregated into category scores for that method. These weighted scores, which averaged 21 categories per case, were then totaled and divided by the total possible score, generating a percentage correct ("quality") score for each physician-case combination. For the subanalysis, each scoring category was assigned to 1 of the 5 domains of the encounter—history taking, physical examination, test ordering, diagnosis, and treatment. Weekly team meetings were held to review criteria and ensure consistent application of scoring guidelines. Random audits enhanced the accuracy of the vignettes, the checklists, and the SPs' scoring. TABLE 1 lists a summary of scoring criteria for the CAD complex case.

Analyses

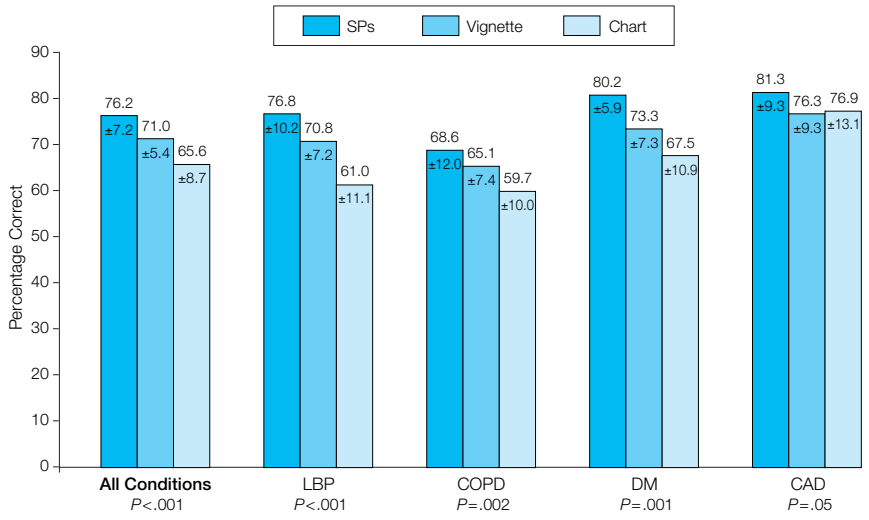
Scores for the 3 methods were compared using a 4-way (3-way nested, 1-way crossed) analysis of variance model. The factors were design effects (site and physician training level) and random effects (quality measurement method and provider). A site-method interaction term was also included. The statistical significance of the difference between means for the 3 methods was determined using an F test; where these differences were statistically significant, the significance of pairwise comparisons between methods was measured using the Student-Neuman-Keuls test. We used the same statistical methods to compare scores for individual conditions, acute and

Table 1. Scoring Categories for Complex Coronary Artery Disease Case 2

Domain	Assessment: Necessary	Assessment: Unnecessary/Inappropriate
History: ask patient about	Use of thrombolytics for myocardial infarction Previous invasive procedures performed Angina and other symptoms Selected risk factors or comorbidities Prevention Drug treatment Risk factors	Gastrointestinal or gastrourinary symptoms
Physical examination	Evaluate for signs of congestive heart failure Examine heart and lungs Evaluate for peripheral vascular disease	Neurological examination Abdominal examination
Tests ordered	Chemistry 7* Cholesterol Electrocardiogram Exercise tolerance test or cardiology referral Echocardiogram or chest radiograph	Holter monitor† Pulmonary referral†
Diagnosis	Large anterior myocardial infarction Symptoms of congestive heart failure	
Management		
Prescribe	Angiotensin-converting enzyme inhibitor Aspirin Diuretic	β-Blocker Cholesterol-lowering agent (before liver function test results)
Discuss	Preventive measures and counseling Follow-up visits General preventive care	Other medications Cardiac rehabilitation

*Chemistry 7 indicates laboratory tests for glucose, blood urea nitrogen, creatinine, potassium, sodium, chloride, and carbon dioxide.
†Not included in final scoring.

Figure. Three-Way Comparison of Standardized Patients, Vignettes, and Chart Abstraction



SPs indicates standardized patients; LBP, low back pain; COPD, chronic obstructive pulmonary disease; DM, diabetes mellitus; and CAD, coronary artery disease. *P* values are for 4-way analysis of variance comparing the 3 quality evaluation methods for specific conditions. Data presented as percentage ±SD correct.

chronic diseases, and 4 of the 5 domains of the clinical encounter: history taking, physical examination, diagnosis, and treatment. A 2-sample *t* test was used to evaluate the significance of the difference in mean vignette scores between providers who had seen SPs and those who had not.

RESULTS

Main Effect of Measurement Method

The 3-way comparison of the methods—SPs, vignettes, and chart abstraction—is shown in the FIGURE. Mean percentage scores are listed for all cases and for each of the 4 conditions.

The highest quality scores for all cases combined were from SPs (76.2%), followed by vignettes (71.0%), and chart abstraction (65.6%). When the overall scores were disaggregated by each of the 4 conditions, this pattern remained unchanged: vignette scores were consistently higher than scores obtained from chart abstraction and consistently produced scores closer to the SP gold standard than did chart abstraction when measured both in the aggregate and by individual condition. The differences among mean scores for the 3 methods were statistically significant in a 3-way comparison for all conditions except CAD ($P = .05$). The interaction effect expected to be strongest, site by method, was not significant ($P = .14$).

Case Effects

We performed subanalyses to assess whether vignettes were sensitive to case effects, defined as differences between methods across simple vs complex cases and acute vs chronic diseases. The results (TABLE 2) were similar to the overall and disease-specific findings above. For example, in the simple case, vignette quality scores (74.3%) were closer to the SP gold standard (76.9%) than was chart abstraction (63.9%) ($P<.001$). When we grouped the acute cases (low back pain and chronic obstructive pulmonary disease exacerbation) and compared them with the 2 more chronic disease conditions, sub-

analyses displayed a similar overall pattern: SP scores were higher than vignette scores, which were higher than chart abstraction scores.

Site and Provider Effects

We tested to see if the 3 methods in general, and vignettes specifically, would consistently reflect expected differences in quality scores due to design effects, defined as differences between sites and among provider training level. Site B consistently scored higher than site A, regardless of method, with a statistically significant ($P<.001$) difference between sites (TABLE 3). Within each site, we again found that vignette scores always approximated the SP scores better than the chart abstraction scores did. We also observed a difference in the scores between the 2. When we compared individual providers (not shown), there was variation in quality scores with all 3 methods ($P<.01$). When we stratified the analysis by physician training level—second- or third-year residents vs attending physicians—the pattern between methods was again robust. Third-year residents scored higher than attendings and typically (but not always) higher than second-year residents.

Domains

We measured 4 discrete skill domains of the outpatient visit: history taking, physical examination, diagnosis, and the treatment or plan. Patterns were similar to other subanalyses for the domains of history and diagnosis, but vignette scores were higher than SPs for the physical examination and lower than chart abstraction for treatment (TABLE 4).

Cuing and Detection Effects

We were concerned that there might be an ordering effect since we did not randomly administer the 3 methods (physicians who had completed vignettes might then look for similar patients in their clinic). Since this left open the possibility of cuing (answers to the vignettes would be higher because physicians had been prompted to think about this type of case after seeing an SP), we gave only vignettes to a second sample of 10 randomly chosen providers at each site. We compared the vignette scores of the second group with those of the physicians who had also seen the SPs. The difference in mean scores for the 2 groups was not statistically significant either overall ($P = .37$) or when the 4 cases were disaggregated.

We also surveyed physicians as to whether they believed they had seen any SPs. Of the 160 visits, only 5 SPs (3%) were detected (2 in site A, 3 in site B) comparing favorably with detection rates in similar SP studies.[45] Two false-positive detections were also reported (1%).

COMMENT

Valid measures of the competence and practice of physicians are basis of efforts to improve quality of care. However, competence and practice have been difficult to isolate from structural effects. Moreover, the cost of measuring quality across systems while controlling for case mix has further confounded efforts to improve physician practice. This study measured quality in an outpatient setting by using the common method of chart abstraction; a gold standard method of SPs; and clinical vignettes, which heretofore have not been rigorously validated.[10,46]

Despite widespread use of vignettes, there is uncertainty and controversy about whether vignettes reflect actual clinical practice or merely physician competence. Some investigators argue that vignettes only reflect what providers are competent or knowledgeable enough to do.[47,48] Other studies have found that vignettes predicted use of computed tomographic or magnetic

Table 2. Relative Ranking of Quality Evaluation Methods Across Case Effects

	Quality Evaluation Method, Percentage Correct*		
Case Effects	Standardized Patients	Vignettes	Chart Abstraction
Case complexity			
Simple	76.9 (9.7)	74.3 (5.6)	63.9 (10.4)
Complex	75.6 (6.5)	67.9 (7.2)	67.2 (8.5)
Disease state			
Acute	72.7 (9.1)	67.9 (6.6)	60.4 (9.0)
Chronic	80.7 (6.8)	74.8 (6.9)	72.3 (10.2)

*Data are presented as % (SD). P<.001 for all comparisons using 4-way analysis of variance comparing the 3 quality evaluation methods.

Table 3. Ranking of Quality Evaluation Methods Across Design Effects

	Quality Evaluation Method, Percentage Correct*			
Design Effect	Standardized Patients	Vignettes	Chart Abstraction	P Values†
Site				
A	72.2 (7.2)	68.7 (6.4)	60.2 (8.2)	<.001
B	80.2 (4.7)	73.2 (2.9)	71.0 (5.2)	<.001
Physician training level				
Second-year resident	76.5 (5.8)	70.5 (5.6)	67.7 (4.0)	.046
Third-year resident	78.1 (7.8)	71.4 (2.6)	65.9 (11.4)	<.001
Attending physician	73.7 (8.3)	71.0 (8.0)	62.8 (9.8)	.01

*Data are presented as % (SD).
†P values are for 4-way analysis of variance comparing the 3 quality evaluation methods.

Table 4. Relative Ranking of Quality Evaluation Methods Across Most Domains*

Domain	Method				
History	Standardized patients	>	Vignettes	>	Chart abstraction
Physical examination	Vignettes	>	Standardized patients	=	Chart abstraction
Diagnosis	Standardized patients	=	Vignettes	>	Chart abstraction
Treatment	Standardized patients	>	Chart abstraction	>	Vignettes

*Two-way comparisons using Student-Neuman-Keuls method. > indicates P≤.05; = indicates P>.05.

resonance imaging,[49] reflected variation in quality when vignettes with open-ended responses were used,[50] demonstrated poor history-taking skills,[47] or showed inadequate use of warfarin in atrial fibrillation.[51] This study advances these earlier studies in several ways: it used a comprehensive set of quality measures pertaining to all aspects of a clinical visit, quality was scored on explicit criteria based on national guidelines and expert panels, the vignettes had an open-ended response format, physicians were prospectively selected into the study, and vignettes were compared with 2 other quality measurement methods.

Our results suggest that vignettes may be a useful way to measure physician practice in an outpatient setting. Vignette scores appeared to reflect actual physician practice as recorded from SP visits, resulting in higher criterion validity, and consistently measured physician practice more accurately than did chart abstraction scores, resulting in better content validity. Vignettes also were more effective than chart abstraction at measuring variations in quality between the 2 study sites, yielding good face validity. We did not find a cuing effect for vignettes when physicians had already seen SPs.

We infer from these findings that low quality may be significantly determined by physician competence and not merely structural effects. If vignette scores had been much higher than SP scores, for example, it could be argued that practice deteriorated because of a structural effect such as the organization or delivery of care. When we initially designed the study, we hypothesized that we might find vignette scores (measures of competence) to be higher than those of SPs or chart abstraction. We reasoned that a social desirability bias in vignette responses and the vignettes' potential to emphasize knowledge over actual clinical practice would result in higher scores that overestimated the process of care.[52,53] However, we found that SP scores were consistently higher than vignette scores (which, in turn, were higher than chart abstraction), implying that practice is better than competence, at least for vignettes with open-ended responses. A clinically based explanation is that the dynamic nature of the patient-physician dialogue may cue the physician's thinking during the visit. The lower chart scores, we reasoned, are the effect of recording bias—everything that happens in the clinical encounter is not written down because of time constraints. In the future, modifying the vignettes or varying the SP presentation may help disentangle the direct effect of the patient encounter from the indirect simulation of the vignette.

The face validity of the study and the general variation in quality scores we observed deserve comment. Based on unquantified proxies such as competitiveness of the respective residency programs, we expected site B to score higher than site A. Vignettes were able to capture this effect. We also observed that third-year residents generally outperformed second-year residents and attending physicians. Perhaps this is not surprising—it is not unrealistic to believe that senior residents know more than junior residents and provide higher-quality care or exhibit a higher degree of assiduousness than faculty.

Despite their promise, vignettes are not a panacea for measuring quality. Our analyses of disaggregated data revealed a complex story. Vignettes appear to overestimate the quality of the physical examination and inconsistently assess the quality of the treatment plan. We surmise that the reason for the higher physical examination scores is that writing down an examination in the vignette has little "temporal cost," whereas carefully performing additional physical examination items on a patient in the clinic takes time away from other activities such as ordering tests. We believe that the chart may be more accurate than vignettes for recording treatment plans. The medical record is often used to convey treatment orders (eg, for a follow-up appointment or an imaging study). Structural problems may further degrade quality as measured by charts—for example, when orders that were correctly requested by the physician are lost or delayed.

We believe vignettes have an important niche in the overall measurement of quality but that their use should be carefully defined and further studied. Our study indicates that vignette scores are a valid overall measure of the process of care provided by groups of physicians for a range of common outpatient conditions. The measure appears to be responsive to real variations in quality among sites and robust for individual diseases. Such a measure could be useful to policymakers, purchasers, and managers as they seek to compare the quality of care in different settings or evaluate management and policy interventions.

In addition, vignettes are uniquely suited for comparative analyses because they better control for case-mix variation and reduce the impact of structural effects.[14,21,53] Properly adjusting aggregate measures of quality for variations in case mix and in patient populations is essential for valid comparisons of quality between health care systems.[54] Since vignettes in this study and elsewhere appear to be responsive to changes in quality, they make comparisons of quality across time possible.[39,40,55,56] Specifically, vignettes could be used to measure the impact of organization reforms or policy changes whose ostensible purpose is to improve the care that patients receive.

Finally, vignettes directly measure the process of care, which is where interventions can be targeted to improve overall health care quality. Structural features, it is sometimes argued, are major determinants of quality in certain circumstances, but they are difficult to measure or directly influence.[57] Vignettes appear to be most useful when the focus is on measuring the competence and even practices for a group of providers. The disadvantage of such a focus is that there may be other more important reasons for poor-quality care. Nevertheless, pro-

cess measures look directly at what services are provided, whether they are provided efficiently, and whether they lead to better health.

Identifying specific deficiencies in the process of care has implications on how clinical care might be improved. If, for example, specific limitations are identified for 1 condition, a disease-based approach might be used; if, instead, the deficiency is in ordering appropriate tests, training might shift to an analytic approach to diagnostic testing. Identified deficiencies in process could also be combined with population health issues, such as disease prevalence or management of underdiagnosed conditions.

The last point implies that when vignettes are used to measure process, they must be carefully constructed. The criteria should be linked to explicit outcomes or evidence-based guidelines, and the responses should be open-ended. As others have shown when measuring quality responses, eliminating disparity requires that methodical steps be taken to ensure that scoring criteria are evenly applied.[13,58,59]

As they are currently developed and validated, vignettes also have 2 important limitations that discourage their use to measure individual provider performance or individual quality criteria. First, the limited intermethod agreement, demonstrated by the domain variation from this and other studies, argues that vignettes should not be used to assess individual-level performance.[59-61] Second, it may be unwise to emphasize measurements of individual criteria or individual provider performance: poor performance on a single criterion may reflect a rare event and not indicate a pattern of poor quality; similarly, focusing on an individual provider fails to foster the type of relationships necessary to improve the care provided by a group of physicians and associated health care workers.[8]

If vignettes are to be used appropriately, more prospective evaluation of their strengths and weaknesses will be needed. Future studies are needed to extend the range of clinical conditions and practice settings. This study, for example, is limited to 4 outpatient conditions and new or walk-in patients. Another limitation is that, although we used 2 sites, they are both academically affiliated Veterans Affairs medical centers. It is possible that structural elements, such as organization of care or patient population characteristics, will affect the way providers answer vignettes.

Until these issues are formally addressed, caution is warranted before extending vignette responses beyond global-level performance assessments. While these results indicate that vignettes can measure actual clinical practice by a group of providers, they should not be used to ascertain the deficiencies in a single provider's ability to obtain a piece of information, perform a skill or task, or complete a treatment plan.

Vignettes are likely to prove less expensive than chart abstraction and are certainly less costly than training SPs. And if other studies substantiate our findings, vignettes hold promise as a method to measure quality in the outpatient setting while controlling for case mix and structural effects across sites. Ultimately, dependable quality measurement—which ensures that an intervention designed to improve practice actually does so—is central to health care reform.

Author Affiliations: San Francisco Veterans Affairs Medical Center and Institute for Global Health, University of California, San Francisco (Dr Peabody); RAND, Santa Monica (Dr Peabody); Veterans Affairs, Greater Los Angeles Healthcare System, West Los Angeles (Drs Peabody, Luck, and Glassman); University of California, Los Angeles, Schools of Medicine and Public Health, Los Angeles (Drs Luck, Peabody, and Lee); Veterans Affairs Center for the Study of Health Care Provider Behavior (Drs Peabody, Luck, Lee, and Glassman); and San Diego Veterans Affairs Medical Center, University of California, San Diego School of Medicine (Dr Dresselhaus).

Funding/Support: This research was supported by grant IIR 95-014B for Veterans Affairs Health Services Research and Development Service, Washington, DC. Dr Peabody is also the recipient of a Senior Research Associate Career Development Award from the Department of Veterans Affairs.

Acknowledgment: We thank Elizabeth O'Gara, University of California, Los Angeles, Barbara Arnaelsteen, RN, MPH, Veterans Affairs Greater Los Angeles Healthcare System (VAGLAHS), Cathy Verkaik, RN, San Diego Veterans Affairs Medical Center (SDVAMC), and Anita Richards, University of California, San Diego (UCSD), who coordinated the SP visits; and Amilcare Gentili, MD, Anthony Pineda, MPH, and Coletta Austin, RN, MSN, VAGLAHS, who assisted with the creation of the SP database; our colleagues, in particular, Lisa Rubenstein, MD, MPH, VAGLAHS and RAND, who was instrumental in discussing our findings; and Robert Brook, MD, ScD, RAND, Sam Bozzette, MD, PhD, SDVAMC and USCD, and Martha Gerrity, MD, Veterans Affairs Medical Center, Portland, Ore. We also thank Joanna Nelsen and Miriam Polon, RAND, who prepared the manuscript for publication, and Ming Ming Wang, MS, VAGLAHS, who assisted with the data. Finally, a special thanks to the actors for their fine performances.

REFERENCES

1. Palmer RH, Louis TA, Hsu LN, et al. A randomized controlled trial of quality assurance in sixteen ambulatory care practices. *Med Care*. 1985;23:751-770.
2. Goldman RL. The reliability of peer assessments of quality of care. *JAMA*. 1992;267:958-960.
3. Peabody JW, Luck J. How far down the managed care road? a comparison of primary care outpatient services in a Veterans Affairs medical center and a capitated multispecialty group practice. *Arch Intern Med*. 1998;158:2291-2299.
4. Brook RH, Lohr KN. Efficacy, effectiveness, variations, and quality boundary-crossing research. *Med Care*. 1985;23:710-722.
5. Peabody JW, Rahman MO, Gertler PJ, Mann J, Farley DO, Carter GM. *Policy and Health: Implications for Development in Asia*. Cambridge, Mass: Cambridge University Press; 1999.
6. US Congress, Office of Technology Assessment. *Identifying Health Technologies That Work: Searching for Evidence*. Washington, DC: US Government Printing Office; 1994. Publication OTA-H-608.
7. Donabedian A. *The Definition of Quality and Approaches to Its Assessment*. Ann Arbor, Mich: Health Administration Press; 1980.
8. Lawthers AG, Palmer RH, Edwards JE, Fowles J, Garnick DW, Weiner JP. Developing and evaluating performance measures for ambulatory care quality: a preliminary report of the DEMPAQ project. *Jt Comm J Qual Improv*. 1993;19:552-565.
9. Tamblyn R, Abrahamowicz M, Brailovsky C, et al. Association between licensing examination scores and resource use and quality of care in primary care practice. *JAMA*. 1998;280:989-996.
10. McDonald CJ, Overhage JM, Dexter P, Takesue BY, Dwyer DM. A framework for capturing clinical data sets from computerized sources. *Ann Intern Med*. 1997; 127:675-682.
11. Rubin HR, Rogers WH, Kahn KL, Rubenstein LV, Brook RH. Watching the doctor-watchers: how well do peer review organization methods detect hospital care quality problems? *JAMA*. 1992;267:2349-2354.
12. Gilbert EH, Lowenstein SR, Koziol-McLain J, Barta DC, Steiner J. Chart reviews in emergency medicine research: where are the methods? *Ann Emerg Med*. 1996;27:305-308.
13. Wu L, Ashton CM. Chart review: a need for reappraisal. *Eval Health Prof*. 1997;20:146-163.
14. Lawthers AG, Palmer RH, Banks N, Garnick DW, Fowles J, Weiner JP. Designing and using measures of quality based on physician office records. *J Ambulatory Care Manage*. 1995;18:56-72.
15. Rubenstein L, Mates S, Sidel VW. Quality-of-care assessment by process and outcome scoring: use of weighted algorithmic assessment criteria for evaluation of emergency room care of women with symptoms of urinary tract infection. *Ann Intern Med*. 1977; 86:617-625.
16. Luck J, Peabody JW, Dresselhaus TR, Lee M, Glassman P. How well does chart abstraction measure quality? a prospective comparison of quality between standardized patients and the medical record. *Am J Med*. In press.
17. Norman GR, Davis DA, Lamb S, et al. Competency assessment of primary care physicians as part of a peer review program. *JAMA*. 1993;270:1046-1051.

18. Ashton CM, Kuykendall DH, Johnson ML, et al. The association between the quality of inpatient care and early readmission. *Ann Intern Med*. 1995;122:415-421.
19. Kravitz RL, Greenfield S, Rogers W, et al. Differences in the mix of patients among medical specialties and systems of care: results from the Medical Outcomes Study. *JAMA*. 1992;267:1617-1623.
20. Salem-Schatz S, Moore G, Rucker M, Pearson SD. The case for case-mix adjustment in practice profiling: when good apples look bad. *JAMA*. 1994;272:871-874.
21. Rosen AK, Ash AS, McNiff KJ, Moskowitz MA. The importance of severity of illness adjustment in predicting adverse outcomes in the Medicare population. *J Clin Epidemiol*. 1995;48:631-643.
22. Badger LW, deGruy F, Hartman J, et al. Stability of standardized patients' performance in a study of clinical decision making. *Fam Med*. 1995;27:126-131.
23. Colliver JA, Vu NV, Marcy ML, Travis TA, Robbs RS. Effects of examinee gender, standardized-patient gender, and their interaction on standardized patients' ratings of examinees' interpersonal and communication skills. *Acad Med*. 1993;68:153-157.
24. Pieters HM, Touw-Otten FW, DeMelker RA. Simulated patients in assessing consultation skills of trainees in general practice vocational training: a validity study. *Med Educ*. 1994;28:226-233.
25. De Champlain AF, Margolis MJ, King A, Klass DJ. Standardized patients' accuracy in recording examinees' behaviors using checklists. *Acad Med*. 1997;72(suppl 1):S85-S87.
26. Rethans JJ, Van Boven CP. Simulated patients in general practice: a different look at the consultation. *BMJ*. 1987;294:809-812.
27. Colliver JA, Swartz MH. Assessing clinical performance with standardized patients. *JAMA*. 1997;278:790-791.
28. McLeod PJ, Tamblyn RM, Gayton D, et al. Use of standardized patients to assess between-physician variations in resource utilization. *JAMA*. 1997;278:1164-1168.
29. Carney PA, Ward DH. Using unannounced standardized patients to assess the HIV preventive practices of family nurse practitioners and family physicians. *Nurse Pract*. 1998;23:56-76.
30. Swartz MH, Colliver JA. Using standardized patients for assessing clinical performance: an overview. *Mt Sinai J Med*. 1996;63:241-249.
31. Beullens J, Rethans JJ, Goedhuys J, Buntinx F. The use of standardized patients in research in general practice. *Fam Pract*. 1997;14:58-62.
32. Safran DB, Kosinski M, Tarlov AR, et al. The primary care assessment survey: tests of data quality and measurement performance. *Med Care*. 1998;36:728-739.
33. Sriram TG, Chandrashekar CR, Isaac MK, et al. Development of case vignettes to assess the mental health training of primary care medical officers. *Acta Psychiatr Scand*. 1990;82:174-177.
34. O'Neill D, Gerrard J, Surmon D, Wilcock GK. Variability in scoring the Hachinski Ischaemic Score. *Age Ageing*. 1995;14:242-246.
35. Glassman PA, Kravitz RL, Petersen LP, Rolph JE. Differences in clinical decision making between internists and cardiologists. *Arch Intern Med*. 1997;157:506-512.
36. Glassman PA, Rolph JE, Petersen LP, Bradley MA, Kravitz RL. Physicians' personal malpractice experiences are not related to defensive clinical practices. *J Health Polit Policy Law*. 1996;21:219-241.
37. Yager J, Linn LS, Leake B, Gastaldo G, Palkowski C. Initial clinical judgment by internists, family physicians, and psychiatrists in response to patient vignettes, I: assessment of problems and diagnostic possibilities. *Gen Hosp Psychiatry*. 1986;8:145-151.
38. Tait RC, Chibnall JT. Physician judgments of chronic pain patients. *Soc Sci Med*. 1997;45:1199-1205.
39. Colenda CC, Rapp SR, Leist JC, Poses RM. Clinical variables influencing treatment decisions for agitated dementia patients: survey of physician judgments. *J Am Geriatr Soc*. 1996;44:1375-1379.
40. Jones TV, Gerrity MS, Earp J. Written case simulations: do they predict physicians' behavior? *J Clin Epidemiol*. 1990;43:805-815.
41. Ryynänen OP, Myllykangas M, Kinnunen J, Takala J. Doctors' willingness to refer elderly patients for elective surgery. *Fam Pract*. 1997;14:216-219.
42. Lohr KN, ed. *Medicare: A Strategy for Quality Assurance*. Washington, DC: National Academy Press; 1990.
43. Peabody JW, Rahman O, Fox K, Gertler P. Quality of care in public and private primary health care facilities: structural comparisons in Jamaica. *Bull Pan Am Health Organ*. 1994;28:122-141.
44. Norman GR, Neufeld VR, Walsh A, Woodward CA, McConvey GA. Measuring physicians' performances by using simulated patients. *J Med Educ*. 1985;60:925-934.
45. Woodward CA, McConvey GA, Neufeld V, Norman GR, Walsh A. Measurement of physician performance by standardized patients. *Med Care*. 1985;23:1019-1027.
46. Carney PA, Dietrich AJ, Freeman DH, Mott LA. The periodic health examination provided to asymptomatic older women: an assessment using standardized patients. *Ann Intern Med*. 1993;119:129-135.
47. Rethans JJ, Sturmans F, Drop R, van der Vleuten C, Hobus P. Does competence of general practitioners predict their performance? comparison between setting and actual practice. *BMJ*. 1991;303:1377-1380.
48. Everitt DE, Avorn J. Clinical decision-making in the evaluation and treatment of insomnia. *Am J Med*. 1990;89:357-362.
49. Carey TS, Garrett J, for the North Carolina Back Pain Project. Patterns of ordering diagnostic tests for patients with acute low back pain. *Ann Intern Med*. 1996;125:807-814.
50. Sandvik H. Criterion validity of responses to patient vignettes: an analysis based on management of female urinary incontinence. *Fam Med*. 1995;27:388-392.
51. Beyth RJ, Antani MR, Covinsky KE, et al. Why isn't warfarin prescribed to patients with nonrheumatic atrial fibrillation? *J Gen Intern Med*. 1996;11:721-728.
52. Kopelow ML, Schnabl GK, Hassard TH, et al. Assessment of performance in the office setting with standardized patients. *Acad Med*. 1992;67(suppl):S19-S21.
53. Aronow DB, Cooley JR, Soderland S. Automated identification of episodes of asthma exacerbation for quality measurement in a computer-based medical record. *Proc Annu Symp Comput Appl Med Care*. 1995:309-313.
54. Luck J, Peabody JW, Tozija F, Pecelj G, Ponce N, Nordyke R. A comparison of the quality of care between the U.S. and a developing country [abstract]. In: Program and abstracts of the 22nd Annual Meeting of the Society of General Internal Medicine; April 29–May 1, 1999; San Francisco, Calif.
55. Cooper GS, Fortinsky RH, Hapke R, Landefeld CS. Primary physician recommendations for colorectal cancer screening. *Arch Intern Med*. 1997;157:1946-1950.
56. Carroll RG. Evaluation of vignette-type examination items for testing medical physiology. *Am J Physiol*. 1993;264(suppl):S11-S15.
57. Peabody JW, Gertler PJ, Leibowitz A. The policy implications of better structure and process on birth outcomes in Jamaica. *Health Policy*. 1998;43:1-13.
58. Socolar RR, Raines B, Chen-Mok M, Runyan DK, Green C, Paterno S. Intervention to improve physician documentation and knowledge of child sexual abuse: a randomized, controlled trial. *Pediatrics*. 1998;101:817-824.
59. Kanten DN, Mulrow CD, Gerety MB, Lichtenstein MJ, Aguilar C, Cornell E. Falls: an examination of three reporting methods in nursing homes. *J Am Geriatr Soc*. 1993;41:662-666.
60. Rethans JJ, Martin E, Metsemakers J. To what extent do clinical notes by general practitioners reflect actual medical performance? a study using simulated patients. *Br J Gen Pract*. 1994;44:153-156.
61. Katz JN, Chang LC, Sangha O, Fossel AH, Bates DW. Can comorbidity be measured by questionnaire rather than medical record review? *Med Care*. 1996;34:73-84.