Using standardised patients to measure physicians’ practice: validation study using audio recordings

Jeff Luck, John W Peabody

Abstract

Objective To assess the validity of standardised patients to measure the quality of physicians’ practice. Design Validation study of standardised patients’ assessments. Physicians saw unannounced standardised patients presenting with common outpatient conditions. The standardised patients covertly tape recorded their visit and completed a checklist of quality criteria immediately afterwards. Their assessments were compared against independent assessments of the recordings by a trained medical records abstractor. Setting Four general internal medicine primary care clinics in California. Participants 144 randomly selected consenting physicians. Main outcome measures Rates of agreement between the patients’ assessments and independent assessment. Results 40 visits, one per standardised patient, were recorded. The overall rate of agreement between the standardised patients’ checklists and the independent assessment of the audio transcripts was 91% (c=0.81). Disaggregating the data by medical condition, site, level of physicians’ training, and domain (stage of the consultation) gave similar rates of agreement. Sensitivity of the standardised patients’ assessments was 95%, and specificity was 85%. The area under the receiver operator characteristic curve was 90%. Conclusions Standardised patients’ assessments seem to be a valid measure of the quality of physicians’ care for a variety of common medical conditions in actual outpatient settings. Properly trained standardised patients compare well with independent assessment of recordings of the consultations and may justify their use as a “gold standard” in comparing the quality of care across sites or evaluating data obtained from other sources, such as medical records and clinical vignettes.

Introduction

Standardised patients are increasingly used to assess the quality of medical practice. They offer the advantage of measuring quality while completely controlling for variation in case mix. Although standardised patients have long been used to evaluate medical students and residents, their use in actual clinical settings is relatively new. In medical education standardised patients are introduced into a carefully controlled setting typically they are directly observed, work in a designated room, and evaluate students from a single school or training programme. Under these controlled conditions standardised patients have been validated to ensure that they perform consistently. Well trained standardised patients effectively and convincingly imitate medical conditions and are remarkably consistent performers, showing high inter-rater agreement and excellent operating characteristics.

Validating the use of standardised patients to measure quality in the actual practice setting is, however, challenging and to our knowledge has not been done. Direct observation in the clinic is difficult for a variety of reasons, including cost, a potential Hawthorne effect (physicians performing better under observation), and ethical problems linked to informed consent (J Peabody, sixth European forum on quality improvement in health care, Bologna, March 2001). We did a validation study to determine whether standardised patients perform as well in the clinic as they do in medical education settings. We introduced unannounced standardised patients into clinics and compared their reports of a physician’s practice with a covert audio recording of the same visit.

Box 1: Clinical scenarios portrayed by standardised patients

- Chronic obstructive pulmonary disease with a mild exacerbation and history of hypertension
- Chronic obstructive pulmonary disease with an exacerbation associated with productive sputum, slight fever, and past history of hypertension
- Type 2 diabetes with limited preventive care in the past and untreated hypercholesterolaemia
- Poorly controlled type 2 diabetes and early renal damage
- Congestive heart failure secondary to long standing hypertension and non-compliance with treatment
- New onset amaurosis fugax in patient with multiple risk factors
- Depression in an older patient with no other major clinical illness
- Depression complicated by substance abuse
Methods

Setting
The study sites were four general internal medicine primary care clinics in California. All staff physicians, teaching physicians, and second or third year residents were eligible. Of the 163 eligible physicians, 144 consented to see standardised patients at some time during the 1999-2000 and 2000-1 academic years. We used the sampling function of Stata to randomly select consenting physicians to whom standardised patients would present with one of eight different clinical cases, two cases each for four common outpatient conditions: chronic obstructive pulmonary disease, diabetes, vascular disease, and depression (box 1).

Training of standardised patients
We trained 45 professional actors, approximately six per case scenario, as standardised patients. The training protocol involved several steps and is described in detail elsewhere. We prepared detailed scripts for each case scenario and assigned each actor to one of the eight cases. Actors, in groups of three, underwent five training sessions. They were trained how to act as a patient and to observe and recall the physician’s actions during the visit. The actors were trained to complete a checklist of 35-45 items that might be performed or discussed by the physician (box 2). The actors completed the checklist immediately after the visit by marking each item as done or not done. Checklist items were based on quality measurement criteria derived from national guidelines on specific conditions and were arrived at by expert panel review and a modified Delphi technique (a formal method to determine the extent of consensus).

Audio recording of visits
Of the 45 trained actors we recorded 42, using a digital “pen” recorder concealed on the actor. Three actors left the study before completing their recorded visits. Each actor was recorded once. Two recordings were unusable because of difficulties with the recorders. In 27 of the 40 successfully recorded visits the physicians reported that they had detected the standardised patients.

The number of visits was similar across study sites, conditions, and physicians’ level of training. To minimise potential variation in performance, we asked the actors to wear the recorder for visits that were not recorded. A single transcriptionist transcribed all recordings. A trained medical records abstractor then scored each transcript using the same quality criteria as in the standardised patients’ checklist. A second trained medical record abstractor reviewed each transcript against the recording.

Analysis
A total of 1258 quality measurement items were compared. The items were aggregated into four domains corresponding to stages of a visit: history taking, physical examination, diagnosis, and treatment and management. An additional 287 items in the physical examination domain were recorded on the standardised patients’ checklists but not compared, because they were only visually observed and could not be verified in the audio recordings. We calculated the percentage of items in agreement between the standardised patients’ checklists and the recording transcripts. We calculated κ values to further quantify the degree of agreement. Percentage agreement and κ values were disaggregated by condition, site, physicians’ level of training, and domain. A calibration curve was constructed to assess variation across actors. Sensitivity and specificity were calculated for each visit and for all visits combined, taking the audio recording as “truth” in the calculation. A receiver operator characteristic curve was then constructed by plotting sensitivity and specificity for each visit and choosing the most conservative spline that circumscribed all data points.

Results
The overall rate of agreement between corresponding items on the standardised patients’ checklists and the recording transcripts was 91% (κ=0.81) (table 1). Agreement rates for the four conditions ranged from 88% for depression (κ=0.75) to 95% for diabetes (κ=0.89). Agreement rates for individual sites ranged from 90% (κ=0.78) to 93% (κ=0.81). Agreement rates and κ values also varied little by physicians’ training experience.
level. Agreement rates were similar for history taking (91%; $\kappa=0.81$), diagnosis (89%; $\kappa=0.69$), and treatment and management (93%; $\kappa=0.85$).

Figure 1 shows the variation among standardised patients. This calibration curve plots the percentage of checklist items done by the physician as noted by the standardised patient against the corresponding percentage indicated by the audio recording of that visit. Points cluster closely along the plotted regression line, which has an intercept of 0.4% and a slope of 1.03. (Perfect calibration would yield a line with intercept of 0% and a slope of 1.00.)

Sensitivity of standardised patients' assessments, compared against the audio recording transcripts, was 95%, and specificity was 85% (table 2). Table 2 also shows that about two thirds of the items where the two methods disagreed were reported as done by the standardised patient but determined to be not done according to the transcript.

Figure 2 shows the operating characteristics of standardised patients. Each data point represents the sensitivity and specificity values for one recorded visit. The area under the resulting receiver operator characteristic curve is 90%.

**Discussion**

Although patients and physicians alike desire improved quality, accurate measurement of quality remains problematic. Comparisons of quality across physicians and sites are hampered by imperfect adjustments for variation in case mix. Also, the underlying data on quality are of uncertain validity, because of logistical and ethical difficulties in directly observing physicians while they care for patients. Measurement of quality has therefore relied largely on medical records, which at best are incomplete and at worst falsified. Standardised patients, despite being costly to train and implement, overcome the first problem by providing presentations that are perfectly adjusted for case mix. They may also be able to overcome the second problem, if their validity in the outpatient setting can be shown.

Many studies have turned to standardised patients when highly accurate measures of quality are needed. Standardised patients are particularly well suited for cross system comparisons, such as comparing general practice with walk-in care or for assessing quality for potentially sensitive conditions such as sexually transmitted infections and HIV.

Table 2: Sensitivity and specificity of standardised patients' assessments, with respect to audio recordings of consultations

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chronic obstructive pulmonary disease</td>
<td>92.1%</td>
<td>0.83 (0.71 to 0.95)</td>
</tr>
<tr>
<td>Depression</td>
<td>88.1%</td>
<td>0.75 (0.68 to 0.84)</td>
</tr>
<tr>
<td>Diabetes</td>
<td>95.2%</td>
<td>0.89 (0.77 to 1.00)</td>
</tr>
<tr>
<td>Vascular disease</td>
<td>91.5%</td>
<td>0.91 (0.69 to 0.93)</td>
</tr>
</tbody>
</table>

Table 1: Agreement (%) between standardised patients' assessments and audio recordings of consultations

<table>
<thead>
<tr>
<th>Domain of visit*</th>
<th>Agreement</th>
<th>$\kappa$ (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>History taking</td>
<td>91.3%</td>
<td>0.81 (0.71 to 0.91)</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>89.3%</td>
<td>0.69 (0.52 to 0.86)</td>
</tr>
<tr>
<td>Treatment and management</td>
<td>92.7%</td>
<td>0.85 (0.73 to 0.97)</td>
</tr>
<tr>
<td>All visits</td>
<td>91.3%</td>
<td>0.81 (0.75 to 0.87)</td>
</tr>
</tbody>
</table>

*Comparable items in the physical examination domain were too few to yield a significant result.
high. Standardised patients showed excellent sensitivity, specificity, and operating characteristics.

We observed higher sensitivity than specificity—that is, the false positive rate of standardised patients’ assessments exceeded the false negative rate. Given the inherent trade off between sensitivity and specificity, we attribute this finding to our explicit instructions to, “when in doubt, give the provider the benefit of the doubt.” Alternatively, although the technical quality of the recordings was generally high, some false positives could be attributed to unclear speech (if doctor and patient spoke at the same time).

The design of our study helped mitigate technical issues that might have degraded the audio recording data. Although the physicians’ informed consent meant that some standardised patients were detected, we received no reports by physicians or standardised patients that the concealed recorder itself was detected. The actors were coached in precise placement of the recorder, particularly as they undressed during the visit. The accuracy of the transcript was ensured by the use of an experienced transcriptionist as well as a trained abstractor who independently reviewed each transcript against the recording.

Limitations of the study

We assessed only verbal communication. In future studies doctors may consent to unannounced visits that are video recorded. We did not measure within-actor variation. In the medical education setting such variation is managed by using standardised physicians to calibrate the standardised patients.27 Such results show that performances by a standardised patient are consistent from visit to visit. We believe from anecdotal evidence that this was the case in our study as well but have not measured it objectively.

Another issue that merits further study is how accurately standardised patients can measure quality through a single encounter—or even a short series of visits. Some studies suggest that a “first visit bias” may skew assessment of quality, since chronic diseases typically necessitate several visits and ongoing follow up.1 We deliberately used clinical scenarios that required immediate interventions, and we are separately analysing those items (particularly preventive care) that could be postponed to a future visit. Future research might assess how well standardised patients’ measurements of quality for a few selected cases can comprehensively assess an individual physician’s overall competence.5 22 25

We used explicit checklists of quality criteria to measure physicians’ performance. Other studies involving standardised patients have used different analytic approaches, such as global rating scales.24 25 While checklists and rating scales have different emphases—for example, technical versus interpersonal skills—some researchers argue that both these types are valid and reliable.25 We did not use rating scales because of our concerns over the potentially more subjective nature and lower inter-rater reliability of global ratings.27

Setting standards

Using standardised patients to measure quality raises the question of how to set standards for what is considered adequate clinical competence. Panels of expert judges have been shown to be reliable for setting standards.24 The expert judges seem to use a compensatory model, where very good performance on some cases compensates for performing poorly on other cases.27 Analysis of the receiver operator characteristics of standardised patients has also been used to set standards in performance assessments of students at examination level. Receiver operator characteristic analysis shows that standardised patients can differentiate between disparate levels of competence—for example, accurately discriminating between second and fourth year medical students.25 26

Conclusions

Standardised patients’ assessments seem to be a valid measure of the quality of physicians’ care for a variety of common medical conditions in actual outpatient settings. Concealed audio recorders were effective for validating standardised patients’ assessments. Properly trained standardised patients should be considered for comparative measurements of quality of care across sites when validity is essential. As the criterion standard, standardised patients can be used to evaluate the validity of data obtained from other sources, such as medical records and physicians’ (self) reports. We believe standardised patients are particularly useful to validate innovative methods of quality measurement, such as computerised clinical vignettes. Vignettes, like standardised patients, inherently control for case mix variation; and, once validated against actual clinical practice, vignettes can be more widely used because they are cheaper and do not require subterfuge.27 Ultimately, accurate and affordable measurements of clinical practice underlie any effort to provide better quality for patients.28

JL is assistant professor at the UCLA School of Public Health. JWP holds positions with the Veterans Affairs San Francisco Medical Center (staff physician), UCSF Department of Epidemiology and Biostatistics (associate professor), UCLA School of Public Health (associate professor), and RAND (senior social scientist). We thank the actors and the nurses, physicians, and staff at the study sites for their participation and Greer Rothman for preparation of the manuscript.

Contributors: JL and JWP conceived and designed the study, analysed and interpreted the data, drafted and revised the article, and reviewed the final version for publication. JWP will act as guarantor. Peter Glassman, Maureen Spell, Joyce Hansen, and Sharad Jain contributed to planning, coordination, and implementation of the study. Ojig Yeletsian, Christina Conti, and Molly Bates were responsible for implementation and assisted with data collection. Elizabeth O’Gara and Juliane Arnall were responsible for standardised patient training and