

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

Jeff Luck, John W Peabody



The full version of this article appears on bmj.com

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% ($\kappa=0.81$). 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.¹⁻⁴ They offer the advantage of measuring quality while completely controlling for variation in case mix.^{5,6} Although standardised patients have long been used to evaluate medical students and residents, their use in actual clinical settings is relatively

new.⁷ 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. We did a validation study to determine whether standardised patients perform as well in the clinic as they do in medical education settings.³

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 (box).

Training of standardised patients

We trained 45 professional actors, approximately six per case scenario, as standardised patients. The

Veterans Administration, Greater Los Angeles Healthcare System, 11 301 Wilshire Blvd, Los Angeles, CA 90073, USA

Jeff Luck
assistant professor

Institute for Global Health, 74 New Montgomery St, San Francisco, CA 94105, USA

John W Peabody
deputy director

Correspondence to: John W Peabody
peabody@psg.ucsf.edu

BMJ 2002;325:679-82

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

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

Characteristic	Agreement	κ (95% CI)
Condition:		
Chronic obstructive pulmonary disease	92.3	0.83 (0.71 to 0.95)
Depression	88.1	0.75 (0.66 to 0.84)
Diabetes	95.2	0.89 (0.77 to 1.00)
Vascular disease	91.5	0.81 (0.69 to 0.93)
Site:		
A	90.8	0.81 (0.71 to 0.91)
B	91.4	0.83 (0.72 to 0.94)
C	92.8	0.81 (0.71 to 0.91)
D	89.8	0.78 (0.66 to 0.90)
Physicians' level of training:		
Second year resident	91.6	0.81 (0.72 to 0.90)
Third year resident	91.2	0.82 (0.73 to 0.91)
Staff physicians and teaching physicians	91.0	0.81 (0.69 to 0.93)
Domain of visit*:		
History taking	91.2	0.81 (0.74 to 0.88)
Diagnosis	89.3	0.69 (0.52 to 0.86)
Treatment and management	92.7	0.85 (0.73 to 0.97)
All visits	91.3	0.81 (0.75 to 0.87)

*Comparable items in the physical examination domain were too few to yield a significant result.

training protocol involved several steps and is described in detail elsewhere.⁵ The actors were trained to complete a checklist of 35-45 items that might be performed or discussed by the physician (see bmj.com). 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.

Audio recording of visits

Of the 45 trained actors we successfully recorded 40, using a digital "pen" recorder concealed on the patient. Each actor was recorded once. 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 trained medical records abstractor scored each transcribed recording using the same quality criteria as in the standardised patients' checklist.

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. 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. Per-

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

Standardised patients	Audio recording		Total
	Items done	Items not done	
Items done	743	71	814
Items not done	38	406	444
Total	781	477	1258
	Sensitivity 95.1%	Specificity 85.1%	

centage 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.

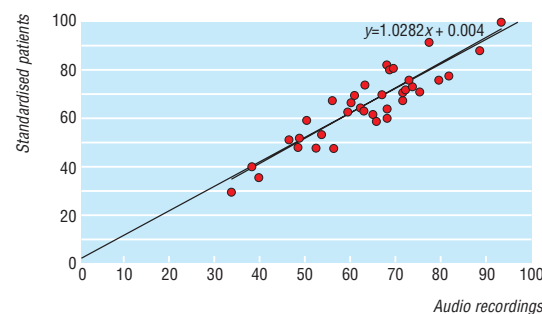
Results

The overall rate of agreement between corresponding items on the standardised patients' checklists and the recording transcripts was 91% ($\kappa=0.81$) (table 1). Rates of agreement varied little by medical condition, site, level of physicians' training, or domain. The figure 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). 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. The area under a receiver operator characteristic curve, constructed by plotting the sensitivity and specificity values for each recorded visit, was 90% (see bmj.com).

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.^{8,9} 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.



Percentage of items on checklist done by physician, as rated by standardised patients and as indicated by audio recordings of visits

Many studies have turned to standardised patients when highly accurate measures of quality are needed.¹⁰ 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.^{11–13}

Standardised patients are already considered the criterion standard for evaluating competence in specialties and have become part of national certification examinations in the United States. And while the accuracy of standardised patients is assumed to be high, it has not been prospectively evaluated.^{14 15}

We found that standardised patients were well calibrated to actual recordings of clinical encounters. No apparent systematic bias was seen by medical condition, site, level of physicians' training, or domain of the encounter. Intermethod reliability was uniformly high. Standardised patients showed excellent sensitivity, specificity, and operating characteristics.

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.¹⁶ 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 17 18}

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.¹⁹ The expert judges seem to use a compensatory model, where very good performance on some cases compensates for performing poorly on other cases.²⁰ 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.^{21 22}

Conclusions

We believe standardised patients are particularly useful to validate innovative methods of quality measure-

What is already known on this topic

Standardised patients are valid and reliable reporters of physicians' practice in the medical education setting

However, validating standardised patients' measurements of quality of care in actual primary practice is more difficult and has not been done in a prospective study

What this study adds

Reports of physicians' quality of care by unannounced standardised patients compare well with independent assessment of the consultations

ment, 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.²³ Ultimately, accurate and affordable measurements of clinical practice underlie any effort to provide better quality for patients.²⁴

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: See bmj.com

Funding: This research was funded by Grant IIR 98118-1 from the Veterans Affairs Health Services Research and Development Service. From July 1998 to June 2001 JWP was the recipient of a senior research associate career development award from the Department of Veterans Affairs.

Competing interests: None declared.

- 1 Beullens J, Rethans JJ, Goedhuys J, Buntinx F. The use of standardized patients in research in general practice. *Fam Pract* 1997;14:58-62.
- 2 Rethans JJ, Martin E, Metsemakers J. To what extent do clinical notes by general practitioners reflect actual medical practice? A study using simulated patients. *Br J Gen Pract* 1994;44:153-6.
- 3 Kopelow ML, Schnabl GK, Hassard TH, Tamblyn RM, Klass DJ, Beazley G, et al. Assessing practicing physicians in two settings using standardized patients. *Acad Med* 1992;67(10 suppl):S19-21.
- 4 Woodward CA, McConvey GA, Neufeld V, Norman GR, Walsh A. Measurement of physician performance by standardized patients. *Med Care* 1985;23:1019-27.
- 5 Glassman PA, Luck J, O'Gara EM, Peabody JW. Using standardized patients to measure quality: evidence from the literature and a prospective study. *Jt Comm J Qual Improv* 2000;26:644-53.
- 6 Carney PA, Dietrich AJ, Freeman DH Jr, Mott LA. The periodic health examination provided to asymptomatic older women: an assessment using standardized patients. *Ann Intern Med* 1993;119:129-35.
- 7 Badger LW, deGruy F, Hartman J, Plant MA, Leeper J, Ficken R, Templeton B, et al. Stability of standardized patients' performance in a study of clinical decision making. *Fam Med* 1995;27:126-31.
- 8 Luck J, Peabody JW, Dresselhaus TR, Lee M, Glassman P. How well does chart abstraction measure quality? A prospective comparison of standardized patients with the medical record. *Am J Med* 2000;108:642-9.
- 9 Dresselhaus T, Luck J, Peabody JW. The ethical problem of false positives: a prospective evaluation of physician reporting in the medical record. *J Med Ethics* 2002;5:291-4.
- 10 Grant C, Nicholas R, Moore L, Salisbury C. An observational study comparing quality of care in walk-in centres with general practice and NHS Direct using standardised patients. *BMJ* 2002;324:1556.
- 11 Russell NK, Boekeloo BO, Rafi IZ, Rabin DL. Unannounced simulated patients' observations of physician STD/HIV prevention practices. *Am J Prev Med* 1992;8:235-40.
- 12 Russell NK, Boekeloo BO, Rafi IZ, Rabin DL. Using unannounced simulated patients to evaluate sexual risk assessment and risk reduction skills of practicing physicians. *Acad Med* 1991;66:87-95.
- 13 O'Hagan JJ, Botting CH, Davies LJ. The use of a simulated patient to assess clinical practice in the management of a high risk asthmatic. *N Z Med J* 1989;102:252-4.

- 14 Norman GR, Davis DA, Lamb S, Hanna E, Caulford P, Kaigas T. Competency assessment of primary care physicians as part of a peer review program. *JAMA* 1993;270:1046-51.
- 15 Norcini JJ, Blank LL, Arnold GK, Kimball HR. The mini-CEX (clinical examination exercise): a preliminary investigation. *Ann Intern Med* 1995;123:795-9.
- 16 Finlay IG, Scott NC, Kinnnersley P. The assessment of communication skills in palliative medicine: a comparison of the scores of examiners and simulated patients. *Med Educ* 1995;29:424-9.
- 17 Dauphinee WD. Assessing clinical performance: where do we stand and what might we expect. *JAMA* 1995;274:741-3.
- 18 Gordon JJ, Saunders NA, Hennrikus D, Sanson-Fisher RW. Interns' performances with simulated patients at the beginning and the end of the intern year. *J Gen Intern Med* 1992;7:57-62.
- 19 Ross L, Clauser B, Margolis MJ, Orr NA, Klass D. An expert judgment approach to setting standards for a standardized-patient examination. *Acad Med* 1996;71(10 suppl):S4-6.
- 20 Margolis MJ, De Champlain AF, Klass DJ. Setting examination-level standards for a performance-based assessment of physicians' clinical skills. *Acad Med* 1998;73(10 suppl):S114-6.
- 21 Colliver JA, Barnhart AJ, Marcy ML, Verhulst SJ. Using a receiver operating characteristic (ROC) analysis to set passing standards for a standardized-patient examination of clinical competence. *Acad Med* 1994;69(10 suppl):S37-9.
- 22 Van der Vleuten CP, Norman GR, De Graff E. Pitfalls in the pursuit of objectivity: issues of reliability. *Med Educ* 1991;25:110-8.
- 23 Peabody JW, Luck J, Glassman P, Dresselhaus TR, Lee M. Comparison of vignettes, standardized patients, and chart abstraction: a prospective validation study of 3 methods for measuring quality. *JAMA* 2000;283:1715-22.
- 24 Fihn SD. The quest to quantify quality. *JAMA* 2000;283:1740-2.

(Accepted 1 August 2002)

Spontaneous talking time at start of consultation in outpatient clinic: cohort study

Wolf Langewitz, Martin Denz, Anne Keller, Alexander Kiss, Sigmund Rüttimann, Brigitta Wössmer

Division of Psychosomatic Medicine, Department of Internal Medicine, University Hospital, CH-4031 Basle, Switzerland

Wolf Langewitz
executive director
Alexander Kiss
medical director
Brigitta Wössmer
head psychologist

University Hospital, Zurich, Switzerland
Martin Denz
chief executive consultant

Forel-Klinik, Ellikon an der Thur, Switzerland
Anne Keller
consultant psychiatrist

Department of Internal Medicine, Kantonsspital, Schaffhausen, Switzerland
Sigmund Rüttimann
head

Correspondence to: W Langewitz
wlangewitz@uhbs.ch

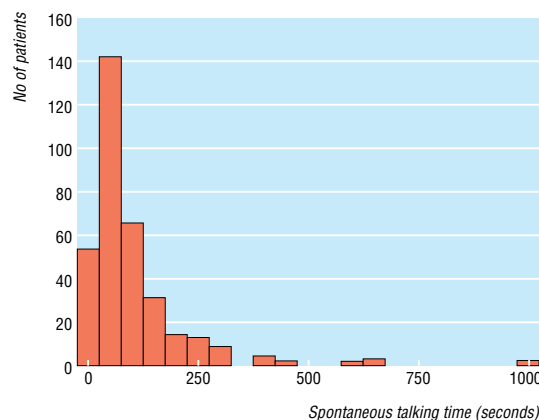
BMJ 2002;325:682-3

The average patient visiting a doctor in the United States gets 22 seconds for his initial statement, then the doctor takes the lead.¹ This style of communication is probably based on the assumption that patients will mess up the time schedule if allowed to talk as long as they wish to. But for how long do patients actually talk, at least initially? We found only one study, from a neurological practice, investigating this question.² The author reported one minute and 40 seconds. We examined how long it would take outpatients at a tertiary referral centre to indicate that they have completed their story—for example, with a statement such as: “That’s all, doctor!” if uninterrupted by their doctors.

Participants, methods, and results

We investigated a sequential cohort of patients from the outpatient clinic of the department of internal medicine at the university hospital in Basle. The study protocol was approved by the university's ethics committee. Inclusion criteria were sufficient knowledge of the German language, first contact with the outpatient clinic, and mental competence. We informed doctors about the purpose of the study and told patients that we were interested in their opinion concerning the service provided. We asked doctors to activate a stop watch surreptitiously at the start of the communication and press it again when patients indicated that they wanted the doctor to take the lead (for example, by saying: “What do you think, doctor?”). Patients did not know that a timer was being used. Doctors were trained for one hour in basic elements of active listening, such as waiting, use of facilitators like “hmm-hmm,” nodding, or echoing. They were told not to ask questions during the initial phase of the consultation. To comply with their consultation schedule they were advised to interrupt if a patient talked for more than five minutes.

Within three months 406 out of a total of 1137 patients fulfilled the inclusion criteria; 33 were later judged as not correctly classified. Of the remaining 373, 20 patients did not give informed consent; for nine



Spontaneous talking time of 331 patients at start of consultation in outpatient clinic

patients doctors did not register talking time; and data on talking time were lost for nine patients. We analysed spontaneous talking time in 335 patients who had been seen by 14 doctors. Of the 330 patients who provided sociodemographic data, 176 (53%) were female, mean age was 42.9 years (SD 18.2 (95% confidence interval 17 to 84) years). The sociodemographic characteristics were typical of patients seen at this hospital.³ The 11 male and three female doctors had worked a mean of 58 (26) months in the clinical field, with a mean of 38 (19) months spent in internal medicine.

Mean spontaneous talking time was 92 seconds (SD 105 seconds; median 59 seconds; figure), and 78% (258) of patients had finished their initial statement in two minutes. Seven patients talked for longer than five minutes. In all cases doctors felt that the patients were giving important information and should not be interrupted. No other sociodemographic variable (education, income, civil status, type of employment, and sex) had a significant influence on spontaneous talking time except for age ($r_s=0.41$; $P < 0.001$; 17-29 years: 77 (105) seconds; 30-49 years: 92 (93) seconds; 50-87 years: 108 (114) seconds).