

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

new.<sup>7</sup> 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.<sup>8</sup> Under these controlled conditions standardised patients have been validated to ensure that they perform consistently.<sup>9,10</sup> Well trained standardised patients effectively and convincingly imitate medical conditions and are remarkably consistent performers, showing high inter-rater agreement and excellent operating characteristics.<sup>11,12</sup>

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.<sup>3</sup> 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.

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.com 2002;325:679

### 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.<sup>5</sup> 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

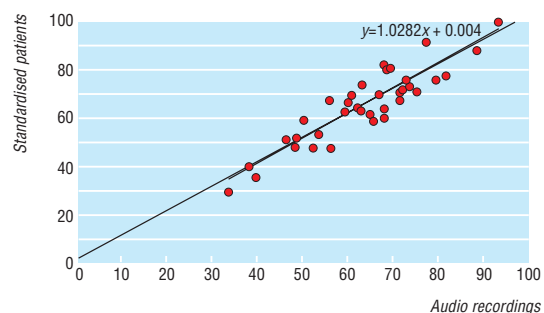
A total of 1258 quality measurement items were compared. The items were aggregated into four domains corresponding to stages of a visit: history taking, physical examination, diagnosis, and treatment and management. An additional 287 items in the physical examination domain were recorded on the standardised 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  $\kappa$  values to further quantify the degree of agreement. Percentage agreement and  $\kappa$  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

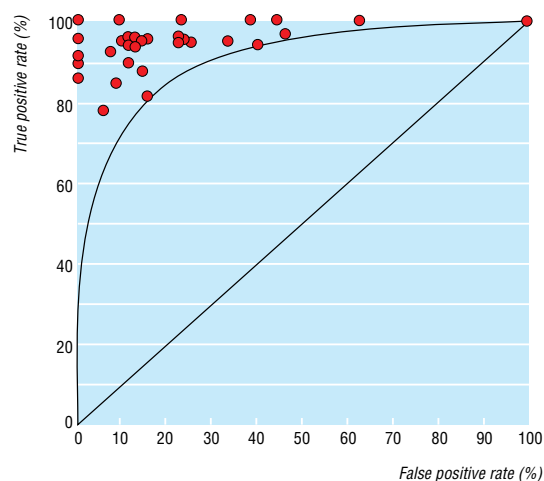
The overall rate of agreement between corresponding items on the standardised patients' checklists and the recording transcripts was 91% ( $\kappa=0.81$ ) (table 1). Agreement rates for the four conditions ranged from 88% for depression ( $\kappa=0.75$ ) to 95% for diabetes ( $\kappa=0.89$ ). Agreement rates for individual sites ranged from 90% ( $\kappa=0.78$ ) to 93% ( $\kappa=0.81$ ). Agreement rates and  $\kappa$  values also varied little by physicians' training

### Box 2: Checklist for evaluating quality of a consultation for chronic obstructive pulmonary disease with a mild exacerbation and history of hypertension

Duration of dyspnoea  
 Severity of dyspnoea  
 Any similar previous episodes  
 History of asthma, emphysema, or chronic obstructive pulmonary disease  
 Medications taken  
 Presence of fever  
 Presence of cough  
 Quality of cough  
 History of hypertension  
 History of high cholesterol concentrations  
 Exposure to allergens or other irritants at workplace  
 Smoking history  
 Alcohol use  
 Last flu or tetanus shots  
 Last Pneumovax  
 Marital status  
 Job or other social history  
 Blood pressure (both arms)  
 Palpation of jugular vein distension or point of maximal impulse  
 Chest auscultation  
 Lung auscultation  
 Examination of digits for cyanosis or clubbing  
 Examination of lower legs for oedema  
 Peak flow evaluation  
 Pulse oximetry (or arterial blood gas analysis)  
 Rectal/prostate examination  
 Diagnosis of chronic obstructive pulmonary disease, emphysema, or bronchitis  
 Discussion of severity of chronic obstructive pulmonary disease  
 Diagnosis of hypertension or discussion of need to continue taking medication for hypertension  
 Consulted with an attending physician  
 Verified proper use of inhaler(s)  
 Told to drink more fluids  
 Told to call or return if symptoms worsen  
 Told needed oxygen or intravenous fluids or to go to the emergency room (*not necessary*)  
 Wanted to admit patient to hospital (*not necessary*)  
 Discussed smoking cessation  
 Counselling on diet  
 Counselling on exercise  
 Recommended, advised, or referred to have colon cancer screening  
 Follow up appointment recommended



**Fig 1** Percentage of items on checklist done by physician, as rated by standardised patients and as indicated by audio recordings of visits



**Fig 2** Receiver operator characteristic curve for standardised patients with respect to audio recordings

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

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

Characteristic	Agreement	$\kappa$ (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.

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.<sup>13 14</sup> 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.<sup>15</sup> 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.<sup>16-18</sup>

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.<sup>19 20</sup>

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

**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%

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.<sup>21</sup> 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.<sup>1</sup> 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.<sup>5 22 23</sup>

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.<sup>24–26</sup> 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.<sup>27</sup> We did not use rating scales because of our concerns over the potentially more subjective nature and lower inter-rater reliability of global ratings.<sup>27</sup>

#### 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

### 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

standards.<sup>24</sup> The expert judges seem to use a compensatory model, where very good performance on some cases compensates for performing poorly on other cases.<sup>25</sup> 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.<sup>26 28</sup>

### 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.<sup>29</sup> Ultimately, accurate and affordable measurements of clinical practice underlie any effort to provide better quality for patients.<sup>30</sup>

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 Yeretsian, Christina Conti, and Molly Bates were responsible for implementation and assisted with data collection. Elizabeth O'Gara and Julianne Arnall were responsible for standardised patient training and

scheduling. Ed LaCalle and Dan Bertenthal assisted with data analysis.

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 Gomez JM, Prieto L, Pujol R, Arbizu T, Vilar L, Pi F, et al. Clinical skills assessment with standardized patients. *Med Educ* 1997;31:94-8.
- 9 Vu NV, Steward DE, Marcy M. An assessment of the consistency and accuracy of standardized patients' simulations. *J Med Educ* 1987;62:1000-2.
- 10 Colliver JA, Swartz MH, Robbs RS, Lofquist M, Cohen D, Verhulst SJ. The effect of using multiple standardized patients on the inter-case reliability of a large-scale standardized-patient examination administered over an extended testing period. *Acad Med* 1998;73(10 suppl):S81-3.
- 11 Thompson JC, Kosmorsky GS, Ellis BD. Field of dreamers and dreamed-up fields: functional and fake perimetry. *Ophthalmology* 1996;103:117-25.
- 12 Colliver JA, Swartz MH. Assessing clinical performance with standardized patients. *JAMA* 1997;278:790-1.
- 13 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.
- 14 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.
- 15 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.
- 16 Russell NK, Boekeeloo BO, Rafi IZ, Rabin DL. Unannounced simulated patients' observations of physician STD/HIV prevention practices. *Am J Prev Med* 1992;8:235-40.
- 17 Russell NK, Boekeeloo 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.
- 18 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.
- 19 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.
- 20 Norcini JJ, Blank LL, Arnold GK, Kimball HR. The mini-CEX (clinical examination exercise): a preliminary investigation. *Ann Intern Med* 1995;123:795-9.
- 21 Finlay IG, Scott NC, Kinnersley 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.
- 22 Dauphinee WD. Assessing clinical performance: where do we stand and what might we expect. *JAMA* 1995;274:741-3.
- 23 Gordon JJ, Saunders NA, Henrikus 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.
- 24 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.
- 25 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.
- 26 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.
- 27 Cohen DS, Colliver JA, Marcy MS, Fried ED, Swartz MH. Psychometric properties of a standardized-patient checklist and rating-scale form used to assess interpersonal and communication skills. *Acad Med* 1996;71(10 suppl):S87-9.
- 28 Van der Vleuten CP, Norman GR, De Graff E. Pitfalls in the pursuit of objectivity: issues of reliability. *Med Educ* 1991;25:110-8.
- 29 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.
- 30 Fihn SD. The quest to quantify quality. *JAMA* 2000;283:1740-2.

(Accepted 1 August 2002)