Using standardized patients to measure professional performance of physicians

MARIE-DOMINIQUE BEAULIEU1,2, MICHELLE RIVARD3, EVELINE HUDON1,4, DANIELLE SAUCIER5, MARTINE REMONDIN2 AND ROBERT FAVREAU6

1Department of Family Medicine, Universite de Montreal, 2Unité de Recherche Evaluative, Centre de Recherche du Centre Hospitalier de l'Universite de Montreal, 3Department of Social and Preventive Medicine and Groupe de Recherche Interdisciplinaire en Sante' (GRIS), Universite de Montreal, Quebec, 4Equipe de Recherche en Medecine Familiale, Cite de la Sante, Laval, Quebec, 5Department of Family Medicine, Universite Laval, Sainte-Foy, Quebec, 6Aventis Pharma, Laval, Quebec, Canada

Abstract

Objective. To determine the nature of inaccuracies likely to occur when standardized patients (SPs) are used to measure physician behaviour and to evaluate the potential impact of these inaccuracies on estimates of physician performance.

Design. Secondary analysis from a randomized controlled trial.

Setting. Family physicians' offices.

Study participants. Eighteen individuals, each portraying one of two patient scenarios, made a total of 179 visits to 92 family physicians who were participating in a separate randomized controlled trial to evaluate the impact of an educational workshop on implementation of preventive guidelines.

Main outcome measures. Accuracy of SPs' portrayal of the assigned scenarios and accuracy of their coding of physician performance, determined on the basis of audiotapes of the visits and correlated with indicators of physicians' preventive practices.

Results. Accuracy of portrayal of the patient scenario was 84.8% for the male SPs and 93.5% for the female SPs. Inaccuracies in portrayal had no impact on physician performance scores. Accuracy of coding of physician performance was 90.5% for the female SPs (kappa = 0.66) and 90.1% for the male SPs (kappa = 0.68). Coding inaccuracies occurred most frequently for assessment of alcohol consumption and advice against smoking.

Conclusion. SPs can provide valid information about physicians' professional performance. However, standardization of their activities must not be taken for granted. It may be more difficult to obtain standardized coding for counselling activities, an aspect of physician visits for which SPs are particularly appropriate.

Keywords: outcome measures, physician performance, quality of care, standardized patients

Measuring physicians' performance in the context of clinical care is an important methodological challenge for health care researchers. The standardized patient (SP), initially proposed by Barrows [1] as an instrument for evaluating medical education, has become a popular method in health services research, despite its high cost. The use of SPs has two main advantages. Firstly, the SP represents a standardized stimulus, never be completely avoided with surveys, clinical vignettes, and direct observation [5,6]. Thus controlling for case-mix variability and some non-medical factors, such as the emotional tone of an interview, that are responsible for variations in medical practices [2,3]. Secondly, SPs can themselves serve as a measurement tool, providing access to information that cannot be obtained from traditional sources, such as the quality of case histories, performance of the physical examination, and counselling activities [4]. Moreover, this method controls for the Hawthorne effect and other biases related to the situation of being observed, such as memory and social desirability biases, which can never be completely avoided with surveys, clinical vignettes, and direct observation [5,6].

However, these advantages are realized only if the physicians are unaware that they are being observed and if the SPs remain standardized, both as 'storytellers' and as
‘performance coders’ [7]. Unfortunately, few researchers using
SPs provide evidence that these conditions have been met,
and few researchers report the potential impact of deviations
in the SPs’ performances on outcome measurements. Indeed,
knowledge about SPs as a source of data in health care
research is limited. Still, as with any instrument relying on
human factors, SPs are subject to variability. Tamblyn et al.
[8–11] are among the few researchers who have studied this
phenomenon, showing that inaccuracies do occur in portrayal
of the standard scenario and in recording the encounter,
and that these inaccuracies affect estimates of physician
performance.

We conducted a comparative trial to evaluate the impact
of a continuing medical education (CME) workshop on the
preventive practices of family physicians, using unannounced
SPs as a standardized stimulus for physicians and as a means
of measuring physicians’ behaviour [12]. In this article, we
report and discuss four aspects of SP-based research: (1) our
experience with the SP research method; (2) the rate
of detection of SPs (unblinding) and the conditions associated
with detection; (3) the accuracy of portrayal of various aspects
of the SP scenario and its impact on the estimates of physician
performance; and (4) the accuracy of coding and the impact
of coding errors on the estimates of physician performance.

Methods

Context

Ninety-two physicians who were accepting new patients into
their practices participated in the main study, which was
designed to evaluate the impact of an educational workshop
on the implementation of recommendations made by the
Canadian Task Force on Preventive Health Care (CTFPHC)
[13]. Participating physicians were informed that within a
period of 6 months after they agreed to enter the study, they
would receive visits from two unannounced SPs, who would
record the visits by means of a hidden microphone. The
physicians were randomly assigned to attend the CME work-
shop either before or after the SP visits. A total of 179 SP
visits were carried out between June 1998 and September
1999.

The study was approved by the Ethics Committee at the
Research Centre of the University of Montreal Health Centre
(Centre Hospitalier de l’Université de Montréal).

Recruitment of standardized patients

Eighteen individuals (nine women, nine men) were recruited
from volunteer associations and senior citizens’ clubs. They
were interviewed to confirm their suitability to participate as
SPs (understanding of the study’s objectives; capacity to
organize one’s thought; absence of issues with the medical
profession). They had to be comparable in terms of physical
characteristics that might influence a physician’s behaviour
during a medical checkup (e.g. ethnic origin, body mass index
between 22 and 25) and they had to be free from any specific
abnormality or disease that could trigger a particular response
from the physician (e.g. arrhythmia, high blood pressure,
scars from major surgery). However, they did not undergo a
physical examination.

Scenario development and standardized patient
training

On the basis of work by Hutchison et al. [14], and with their
permission, we developed two scenarios: a recently retired
man between 55 and 65 years of age who had not seen a
physician for 5 years, and a woman between 55 and 65 years
of age who was worried because one of her friends had just
had a heart attack and who was looking for a new family
physician because hers had recently retired. The SPs were
given an excuse to avoid undergoing a rectal or pelvic
examination.

The SPs were given a detailed description of the scenario
[7], consisting of a standard opening sentence, information
that could be shared spontaneously with the physician, in-
formation that was to be provided only if the physician asked
particular questions, a specific emotional tone for the visit,
and a closing sentence. The SPs kept their own names,
birth dates, and employment records, but marital status
was standardized. Each scenario included 11 key points of
standardized information to be addressed in the visit (Table
1).

During training, the SPs first engaged in role-playing with
physicians on the research team (including M.-D.B. and E.H.).
Then, they practised with physicians who were not involved
in the study. The latter interviews were recorded on video.
The external physicians met with the research team and
suggested ways to make the scenarios and the SP portrayals
more realistic. The video recordings of the female scenario
were reviewed with the female SPs, while those of the male
scenario were reviewed with the male SPs. In addition, the
research team compiled a video illustrating the best examples
for each of the 11 key points in the scenario. This part of
the training was spread over three meetings and lasted a total
of 10 hours.

Next, with the help of video recordings, the SPs were
trained to record the performance of the physicians on a
coding grid until their coding became uniform. During this
process, ambiguities in the coding criteria were clarified. This
session lasted 4 hours.

Finally, each SP made an impromptu visit to another
physician who had agreed to participate in the training.
During this visit the SP practised all aspects of the scenario:
arriving at the clinic, turning on the hidden microphone (in
a handbag for the women or in a leather pouch for the men),
going through the interview and physical examination, and
coding the physician’s performance. The audio recordings
and notes about SP performance made by the physicians
were analysed by the research assistant (M.R.), who rated
the SP’s accuracy of portrayal of the scenario and independently
coded the physician’s performance. A group feedback session
lasting about 3 hours was held after these practice visits.

SPs were allowed into the field when they reached 90% accu-
rate in portrayal and coding. During the project, the
Table 1  Accuracy of standardized patient (SP) scenario portrayal\textsuperscript{1}

<table>
<thead>
<tr>
<th>Key information point</th>
<th>Male scenario</th>
<th>Female scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n = 44 visits\textsuperscript{2}</td>
<td>n = 38 visits\textsuperscript{2}</td>
</tr>
<tr>
<td><strong>Mean accuracy (%)</strong></td>
<td><strong>Range</strong></td>
<td><strong>Mean accuracy (%)</strong></td>
</tr>
<tr>
<td>All SPs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Opening statement</td>
<td>75.0</td>
<td>58.8–100</td>
</tr>
<tr>
<td>Family history</td>
<td>95.2</td>
<td>60.0–100</td>
</tr>
<tr>
<td>Lifestyle</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smoking</td>
<td>100</td>
<td>–</td>
</tr>
<tr>
<td>Alcohol consumption</td>
<td>100</td>
<td>–</td>
</tr>
<tr>
<td>Exercise</td>
<td>90.2</td>
<td>80.0–100</td>
</tr>
<tr>
<td>Eating habits</td>
<td>89.2</td>
<td>71.4–100</td>
</tr>
<tr>
<td>Male SPs only</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lack of knowledge of prostate-specific antigen</td>
<td>81.3</td>
<td>54.5–100</td>
</tr>
<tr>
<td>Gastric problems</td>
<td>80.0</td>
<td>50.0–100</td>
</tr>
<tr>
<td>Excuse to undergo rectal examination next time</td>
<td>78.1</td>
<td>16.7–100</td>
</tr>
<tr>
<td>Urination habits</td>
<td>100</td>
<td>–</td>
</tr>
<tr>
<td>Female SPs only</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hormone replacement therapy</td>
<td>100</td>
<td>–</td>
</tr>
<tr>
<td>Request for information on hormone replacement therapy</td>
<td>88.0</td>
<td>60.0–100</td>
</tr>
<tr>
<td>Date of last mammography examination</td>
<td>90.0</td>
<td>50.0–100</td>
</tr>
<tr>
<td>Excuse to undergo pelvic examination next time</td>
<td>95.8</td>
<td>83.3–100</td>
</tr>
<tr>
<td>All SPs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Closing statement</td>
<td>43.9</td>
<td>0–100</td>
</tr>
<tr>
<td>Overall accuracy</td>
<td>84.8</td>
<td>76.2–87.7</td>
</tr>
</tbody>
</table>

\textsuperscript{1}Calculated as the percentage of the 11 key points for which the SP demonstrated correct behaviour.

\textsuperscript{2}All SP visits were audiotaped except 12. The number of tapes reviewed to determine accuracy of portrayal was chosen to be representative of the number of SPs and each SP’s number of completed visits. This proved to be 44 of the 89 visits by male SPs and 38 of the 90 visits by female SPs.

Fieldwork of standardized patients

At the time each participating physician was recruited, the research team determined the specifics of making an appointment with the physician and passed this information to the SPs, who made their own appointments. When making an appointment, each SP used a local address after ‘scouting around’ in the neighbourhood of the physician’s office.

The SP visits occurred on average 4 months before or after the CME workshop. On average, there was a 1-month interval between the visits of the two SPs to each physician. However, for 24 physicians, the two SP visits occurred on the same day, by SPs who presented themselves as a couple. This strategy was used for physicians who accepted few new patients and for those working in clinics with several participating physicians, in an attempt to prevent detection of the SPs. In these instances, the two SPs went into the physician’s office separately.

Each of the SPs had a medical insurance card issued for the purposes of the project issued by the Quebec Health Insurance Board (la Régie de l’Assurance Maladie du Québec), so that physicians would bill for the visits in the usual manner. Physicians were given an ‘alert card’ to be sent to the research team if they thought they had detected an SP.

Every 3 months, the SPs met with the research team for 2 hours. During these quarterly gatherings, the scenarios and coding were reviewed, and the participants shared their impressions and gave examples of how they had handled certain situations. The meetings also helped to maintain participants’ enthusiasm for the project. Of the 18 individuals initially recruited and trained, 10 (five men, five women) completed the whole study. Eight SPs left the study; four left during/after training, while four left during the course.
of visits to study physicians. They cited personal and health reasons for withdrawing from the study.

**Indicators of physician performance**

Physician performance was evaluated on the basis of 21 items from the CTFPHC recommendations for the scenarios portrayed by the SPs [13]. The SPs coded information for the 13 items related to counselling, physical examination, and vaccination, using a coding grid similar to one used by Hutchison et al. [14]. The grid was organized as a set of questions on each of the items (e.g. ‘During your visit to the office, did the doctor advise you about wearing your seat belt?’), which were to be coded as ‘done’ or ‘not done’. They completed the grid immediately after leaving the physician’s office, attaching to it any prescriptions they received for screening tests; these prescriptions accounted for the other eight items from the CTFPHC recommendations. Scores were calculated for the preventive manoeuvres recommended by the CTFPHC (termed the A/B score) and for the manoeuvres recommended for exclusion (the D/E score).

**Analyses**

Assessment of the detection rate was based on all 179 visits. We explored the association of SP detection with the interval between the two visits, the presence of other participating physicians within the same clinic, the location of the clinic, and individual SPs. All the visits were taped except 14 (the SPs forgot to start the recorder for 12 visits, and the audio was inaudible for two visits).

Throughout the study the research assistant listened to all the tapes in order to detect departure from the scenario and to give appropriate feedback to the SPs. However, in order to stay within the budget limits, it was decided from the start that portrayal accuracy and inter-rater agreement between SPs and the research assistant would be formally evaluated on half of the visits for each SP, at regular intervals from the beginning to the end of the study. Detected visits were excluded from this analysis. Thus, 82 (46%) of the visits (representing each SP according to his or her total number of visits) were formally evaluated by the research assistant, to determine the accuracy of the scenario’s portrayal and reproducibility of the coding. Analyses were performed separately for the male and female scenarios because both cases were different as far as the clinical cues given to the physicians and the expected performance. However, we did not have any hypothesis related to gender effect on the results and did not explore this possibility.

Accuracy of portrayal was calculated as the percentage, for each visit, of the 11 key points for which the SP demonstrated correct behaviour. The association between accuracy of portrayal and the A/B and D/E scores was non-linear and was therefore evaluated using the Spearman correlation coefficient [15].

Accuracy of coding of physician performance was evaluated on the basis of the 10 items that could be validated by listening to the audiotape recordings (one item, ‘clinical breast examination’, coded by SPs was related to the physical examination and was therefore excluded from this calculation). The research assistant used a coding grid identical to the one used by the SPs. The overall percentage of agreement and the inter-rater agreement, as estimated by the kappa statistic [16], were calculated for each item. For points where the research assistant disagreed with the SP’s coding, final coding was determined by the research team as a whole. This coding was the basis for the analysis. The association between the accuracy of coding and the A/B and D/E scores was evaluated using the Spearman correlation coefficient [15]. An ANOVA model [15] was fitted to explore the association between disagreement on coding and the performance scores attributed to the physician on the basis of SP coding. SPSS software (version 9) [17] was used for all the analyses. Visits made by couples were merged with the other visits.

**Results**

**Detection of standardized patients**

Physicians returned a total of 16 alert cards, of which 14 were for SPs detected by eight physicians. The detection rate was therefore 8%. Detected visits generally took place sooner after a previous SP visit than undetected ones (0.6 versus 2 months, respectively), but this difference was not statistically significant. Of the 14 detected visits, eight involved couples who visited the physician on the same day. Therefore, 17% of the 48 visits involving couples were detected, whereas only six (5%) of the 131 visits made by individuals were detected. No SP was detected more frequently than any other. Detected visits occurred more frequently in practice settings outside large urban centres (five of the eight physicians). One physician reported contradictions in the history. Another had a feeling of being evaluated, and a third found a similarity between the SPs case history and the scenario in the CME workshop.

**Accuracy of portrayal of scenarios and impact on performance scores**

Overall accuracy of portrayal of the scenario was high: 93.5% for women and 84.8% for men. Accuracy at the beginning of the interview was slightly lower among the male SPs than among the female SPs (75.0% versus 94.4%, respectively). One male SP was much less successful in avoiding a rectal examination than the others (83.3% deviation). Another SP often digressed from the scenario with respect to his knowledge of the prostate-specific antigen test (45.5% deviation). He had a greater tendency than the others to chat socially with the physician. This issue was discussed with him and he was able to adjust his behaviour. Most of the errors made by female SPs related to the mammography and the end-of-interview items. Two of the female SPs who completed the study made errors in reporting the date of the most recent mammography examination, whereas the other three female SPs always reported the correct date. The accuracy of portrayal at the end of the interview was more variable among the female SPs, ranging from zero (one woman always neglected
to ask for electrocardiography if the physician did not order it) to 100% (average 74.4%). Concluding the interview was also a problem for the men (63.9% accuracy). The SPs told us that it could be difficult to get the final remark in when the physician was concluding the visit rapidly.

Accuracy of portrayal was not significantly correlated with physician performance scores for male SPs (correlation coefficients were 0.17 for the A/B score and 0.18 for the D/E score). For the female SPs, accuracy of portrayal was somewhat associated with the A/B score (correlation coefficient = 0.30; \( P = 0.09 \)) but not associated with the D/E score (correlation coefficient = 0.06).

Reproducibility of coding and impact on physician performance scores

Table 2 summarizes the reproducibility analysis. Overall agreement between the SPs and the research assistant, as well as the kappa value, were high for visits by both female SPs (agreement = 90.5%, kappa = 0.66) and male SPs (agreement = 90.1%, kappa = 0.68). Inter-rater agreement was not as good (<80%) for items concerning assessment of the amount of alcohol consumed, and counselling to stop smoking for men and counselling about physical activity for women. In almost every case, the differences in coding between the SP and the research assistant favoured the physician (i.e. the SP was less severe than the research assistant in his or her judgement of the performance of counselling).

There was no association between the inter-rater agreement and the physician performance scores, except for the A/B score for visits by male SPs, which reached marginal statistical significance (correlation coefficient = \(-0.29; \ P = 0.05\)) again, the SPs’ judgements on items related to counselling were more favourable to the physicians than those of the research assistant.

Discussion

Our results confirm that SPs can be used to evaluate the performance of complex medical activities in the course of health services research. However, despite the many precautions taken, some SPs were detected. In addition, standardization was not always perfect, which affected some of the measures of physician performance.

In many respects these results are consistent with, or better than, those previously reported. The detection rate in this study (8%) was lower than the rates reported by other researchers using the unannounced patient technique (typically 10–20%) [14,19]. We attribute the low detection rate to the quality of standardization for the scenarios and to the specific preparation that the SPs received for making appointments (e.g. details about the referral process and a visit to the neighbourhood before the appointment). In fact, the main factors associated with detection were related to physician characteristics, rather than those of the SPs, and to the length of time between the two SP visits to each physician. Like other authors who have studied detection-related factors [20], we noted that physicians who accepted few new patients, as well as those who received closely spaced SP visits, were more likely to detect the SPs. Researchers have little control over such factors. The strategy we used to reduce detection by physicians who were accepting few new patients, i.e. sending SPs as a couple, did not work well so we do not recommend it. There remains the question of the interval between visits. Clearly, it is desirable to separate SP visits by more than 1 month if more than one visit to each physician is planned. Although this was possible for most of the physicians in our study, it can be difficult, depending on the restrictions inherent to the protocol (e.g. if the visits have to be made before the physician attends a previously scheduled workshop). Conversely, prolonging the study may increase the physician dropout rate, may be a source of fatigue for the SPs, and may require further checks of SP standardization.

In terms of accuracy of portrayal of the scenarios and accuracy of coding of physician performance, we are among the few medical researchers who have rigorously evaluated the SP technique with a large number of physicians and sites. Several studies of this type have been carried out in the context of medical education, but in these studies the physicians were generally informed that they were dealing with SPs [21]. Carney et al. [22] reported high levels of inter-rater agreement and high kappa values in a study similar to ours, which examined a training programme for family physicians to improve cancer control skills. However, those researchers provided no details about the nature of the deviations or their impact on study outcomes.

Some of the kappa values in this study were low despite acceptable values for percentage agreement. This discrepancy can be explained by the asymmetry of the margins for these variables, but it raises the issue of whether kappa values are valid as a measure of inter-rater agreement under these circumstances [23,24].

To the best of our knowledge, only the studies conducted by Tamblyn et al. [8] are comparable to ours [9–11]. Like them, we have emphasized that deviations from the scenarios and deviations in the coding of physician performance affect the performance scores. The impact in our study was less significant than that observed by Tamblyn et al. [8], probably because our basic scenario was less complicated (a visit for a medical checkup, rather than a consultation with one of two elderly SPs suffering from osteoarthritis, one of whom also had melena). In both our study and those of Tamblyn et al. [8–11], the deviations generally favoured the physicians.

The deviations from the standard male scenario, the most frequent of which involved the prostate-specific antigen (PSA) test and the rectal examination, did not influence the D/E score, which consisted of only four items, including prescription of the PSA test. However, deviations from the standard female scenario were associated with more favourable A/B scores. We suggest that the decision to prescribe the PSA test probably depends more on the physician than on his or her perception of what the patient expects. More generally, we cannot assume what impact deviations will have...
Table 2  Accuracy of physician performance coding as reported by standardized patient (SP) and research assistant, and disparities between these reports

<table>
<thead>
<tr>
<th>Item evaluated</th>
<th>Visits by male SPs ($n = 44$)</th>
<th></th>
<th>Visits by female SPs ($n = 36$)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Agreement between SP Kappa (95% CI)</td>
<td>Disagreement in</td>
<td>Agreement between SP Kappa (95% CI)</td>
<td>Disagreement in</td>
</tr>
<tr>
<td></td>
<td>% (range)</td>
<td>favor of MD, % (range)</td>
<td>% (range)</td>
<td>favor of MD, % (range)</td>
</tr>
<tr>
<td>Evaluation of quantity of alcohol consumed</td>
<td>74.0 (37.5–100)</td>
<td>26.0 (0–62.5)</td>
<td>85.0 (62.5–100)</td>
<td>12.0 (0–33.0)</td>
</tr>
<tr>
<td>Use of CAGE questionnaire&lt;sup&gt;1&lt;/sup&gt;</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Blood pressure measurement</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Advice on eating habits</td>
<td>97.0 (75.0–100)</td>
<td>3.0 (0–25.0)</td>
<td>87.0 (71.0–100)</td>
<td>8.0 (0–29.0)</td>
</tr>
<tr>
<td>(fat and fibre)</td>
<td>0.89 (0.69–1.10)</td>
<td>0.74 (0.46–1.01)</td>
<td>0.74 (0.46–1.01)</td>
<td>0.74 (0.46–1.01)</td>
</tr>
<tr>
<td>Advice on calcium intake</td>
<td>NA</td>
<td>0</td>
<td>92.0 (86.0–100)</td>
<td>0.82 (0.58–1.06)</td>
</tr>
<tr>
<td>Advice on physical activity</td>
<td>88.0 (75.0–100)</td>
<td>10.0 (0–22.0)</td>
<td>81.0 (67.0–100)</td>
<td>11.0 (0–33.0)</td>
</tr>
<tr>
<td>Recommendation to stop smoking</td>
<td>79.0 (62.5–100)</td>
<td>12.0 (0–18.0)</td>
<td>90.0 (83.0–100)</td>
<td>3.0 (0–14.0)</td>
</tr>
<tr>
<td>Offer of smoking cessation therapy</td>
<td>88.0 (75.0–100)</td>
<td>7.0 (0–25.0)</td>
<td>96.0 (86.0–100)</td>
<td>4.0 (0–14.0)</td>
</tr>
<tr>
<td>Offer of support to stop smoking</td>
<td>93.0 (80.0–100)</td>
<td>7.0 (0–20.0)</td>
<td>83.0 (71.0–100)</td>
<td>13.0 (0–17.0)</td>
</tr>
<tr>
<td>D2T5 booster vaccine</td>
<td>97.0 (87.5–100)</td>
<td>0</td>
<td>96.0 (80.0–100)</td>
<td>4.0 (0–20.0)</td>
</tr>
<tr>
<td>Global agreement</td>
<td>90.1 0.68</td>
<td>0.65 (0.02–1.28)</td>
<td>90.5 0.66</td>
<td>0.66</td>
</tr>
</tbody>
</table>

95% CI, 95% confidence interval; NA, not applicable.

<sup>1</sup>The CAGE questionnaire [18], a tool to assist assessment of alcohol use, consisting of four questions: (C) Have you ever felt you ought to cut down on drinking?; (A) Have people annoyed you by criticizing your drinking?; (G) Have you ever felt bad or guilty about drinking?; and (E) Have you ever had a drink first thing in the morning to steady your nerves or get rid of a hangover (eye-opener)? Reproduced with permission [18].
Table 3  Recommendations for research with standardized patients (SPs)

<table>
<thead>
<tr>
<th>Aspect of study</th>
<th>Objectives/rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>SP selection</td>
<td>Evaluate individual's capacity for organized thinking</td>
</tr>
<tr>
<td>Selection interviews</td>
<td>Ensure homogeneity and cooperation</td>
</tr>
<tr>
<td>Physical examination</td>
<td>Detect physical abnormalities that could influence the behaviour of physicians</td>
</tr>
<tr>
<td>Sufficient numbers</td>
<td>Avoid exhaustion and estimate dropout rate</td>
</tr>
<tr>
<td>SP training</td>
<td>Avoid chit chat</td>
</tr>
<tr>
<td>Case design: detailed scenarios,</td>
<td>Reduce risk of SPs introducing other elements</td>
</tr>
<tr>
<td>including opening and closing</td>
<td>Ensure that clinical clues given to the physicians are the same</td>
</tr>
<tr>
<td>sentences and psychosocial details</td>
<td>Reduce stress associated with the first ‘field’ experience</td>
</tr>
<tr>
<td>Portrayal: role-playing with</td>
<td>Give feedback on the veracity of the role-play</td>
</tr>
<tr>
<td>experienced clinicians</td>
<td>Aim for high concordance (&gt;80%) for all items evaluated</td>
</tr>
<tr>
<td>Checklist development: coding practice</td>
<td></td>
</tr>
</tbody>
</table>

Making an appointment
- Involve physician’s secretary in difficult cases
- Ensure that SP is prepared to supply an address and a telephone number
- Ensure a delay of at least 1 month between visits

SP follow-up
- Note health problems occurring in the course of the project
- Provide support to SP (with regard to problems encountered or emotions experienced)
- Check coding periodically
- Have research assistant listen to audiotapes periodically
- Allow exclusion of deviant SPs
- Quality of audio recording and positioning of equipment during consultation
- Required microphone potency
- Positioning of equipment (on a table close to the physician)

Validation/evaluation
- Maintenance of standardization
- Identification of incidents that could jeopardize the validity of any visits (which would have to be excluded from the analysis)
- Allow for validation of SP portrayal and coding
on performance scores; they must be observed. This situation emphasizes that each SP scenario is unique. It could be interesting to explore further to what extent gender differences affect the accuracy of coding, a phenomenon into which our data provide no insight.

Many deviations in coding were related to counselling, an element of medical performance that is particularly significant with respect to prevention. Furthermore, in theory, using SPs is an especially appropriate way to obtain information about this kind of activity, for which files and clinical vignettes are less valid [4]. The SPs were less severe in their assessment of counselling performance than was the research team, and they were sometimes satisfied by very superficial inquiries or remarks (e.g. ‘Do you occasionally drink?’ or ‘You should stop smoking’). We cannot determine from these data whether such inquiries or statements would have satisfied real patients. However, Tamblyn et al. [10] reported a high correlation between SPs and real patients seeing the same physician in Aventis Pharma to attend conferences where preliminary works were presented. This study was made possible with respect to prevention. Furthermore, in theory, using SPs want to thank our standardized patients and all the physicians of counselling than do actual patients. Nonetheless, the use of SPs is certainly the best way to evaluate this aspect of medical performance.

There may be ways to reduce even further this human variability in SP portrayals and in coding. For questions related to counselling, researchers should probably spend more time working with the SPs to define what constitutes an acceptable performance on the part of the physician who is being studied. Furthermore, we believe it is essential for SPs to undergo a physical examination as part of the initial assessment so that any abnormality that might elicit a reaction from the physicians can be detected. Some individuals digressed from the scenarios more frequently than others. In general, these people enjoyed chatting with the physicians, and it was during these conversations that they deviated from the scenario. One such patient could not modify his behaviour despite our feedback, and he often asked his own personal questions rather than those stipulated by the scenario. After four visits, we did not assign him to visit any more physicians, citing scheduling difficulties, although we continued to invite him to the medical practice. We were able to detect such problems only by listening to the tape recordings of the interviews, which supports the practice of recording the interviews or at least part of them. In Table 3, we have summarized what we learned and our advice to other researchers planning to use the SP method.

In conclusion, the SP is a powerful instrument for evaluating medical performance when traditional measures are not appropriate or when it is important to control variability in the clinical picture. However, this method is itself subject to variation and must therefore be applied with the greatest possible rigour. During the study, researchers should always provide feedback to SPs for a subsample of their visits, so as to maintain standardization of the scenarios and their coding of physician performance.

Acknowledgements

We are grateful for the vision and guidance of Dr Claude Beaudoin, now deceased, in the early stages of this study. We also thank Drs Robyn Tamblyn, Brian Hutchison, and Christel Woodward, who kindly agreed to share with us their experiences on standardized patient methodology. Finally, we want to thank our standardized patients and all the physicians who participated in the study. This study was made possible by a grant from the Medical Research Council of Canada and by Aventis Pharma, a participant in the MRC-PMAC programme.

Financial disclosure

M.-D.B., D.S. and M.R. received financial support from Aventis Pharma to attend conferences where preliminary results of the study were presented.

References


Accepted for publication 10 February 2003