The European Journal of General Practice

ISSN: 1381-4788 (Print) 1751-1402 (Online) Journal homepage: informahealthcare.com/journals/igen20

Family physicians' ability to detect a physical sign (hepatomegaly) from an unannounced standardized patient (incognito SP)

Francisco Borrell-Carrió, Benilde Fontoba Poveda, Elena Muñoz Seco, Jose Antonio Prados Castillejo, Miguel Pedregal González & Eva Peguero Rodríguez

To cite this article: Francisco Borrell-Carrió, Benilde Fontoba Poveda, Elena Muñoz Seco, Jose Antonio Prados Castillejo, Miguel Pedregal González & Eva Peguero Rodríguez (2011) Family physicians' ability to detect a physical sign (hepatomegaly) from an unannounced standardized patient (incognito SP), The European Journal of General Practice, 17:2, 95-102, DOI: 10.3109/13814788.2010.549223

To link to this article: https://doi.org/10.3109/13814788.2010.549223

Published online: 12 Jan 2011.

European Journal of General Practice, 2011; 17: 95–102

informa
healthcare

ORIGINAL ARTICLE

Family physicians' ability to detect a physical sign (hepatomegaly) from an unannounced standardized patient (incognito SP)

FRANCISCO BORRELL-CARRIÓ[1], BENILDE FONTOBA POVEDA[2], ELENA MUÑOZ SECO[3], JOSE ANTONIO PRADOS CASTILLEJO[4], MIGUEL PEDREGAL GONZÁLEZ[5] & EVA PEGUERO RODRÍGUEZ[6]

[1]*Clinical Sciences Department Campus Bellvitge, Faculty of Medicine, University of Barcelona, Catalan Institute of Health (ICS), Barcelona, Spain,* [2]*Vinyets Health Centre, Catalan Institute of Health (ICS), Barcelona, Spain,* [3]*Menorca Health Centre, Balearic Health Service (Ib-Salut), Psychology Department, University of the Balearic Islands, Menorca, Spain,* [4]*Lucano Health Centre, Andalusian Health Service (SAS), Andalusian Public Health School (EASP), Córdoba, Spain,* [5]*Public Health Research Department, University of Huelva, Huelva, Spain,* [6]*Castelldefels Health Centre, Catalan Institute of Health (ICS), Barcelona, Spain*

Abstract

Background: Little is known about the quality of the physical examination and its effectiveness in daily practice. *Objective:* To determine if family physicians (FPs) were able to detect an important physical sign (hepatomegaly) and to relate this result with other measures of quality. *Methods:* 57 of 104 invited FPs from the National Health Service of the Southern Barcelona Area agreed to schedule an unannounced Standardized Patient (SP) randomly into their daily practice. The SP presented with hepatomegaly and mild abdominal pain. After the visit clinical notes, medical orders, an audiotape of the visit and a checklist completed by the SP detailing items in the physical examination (PE) were analysed. The attainment of a number of quality standards was assessed. *Results:* The three major findings that resulted from this study were: (a) only 4 of the 57 FPs who examined the patient detected the hepatomegaly; (b) FPs performed better at history taking (84.24%) than at PE (26.35%); no correlation was found between the two; (c) diagnostic accuracy was associated with older age, years of experience, history taking skills and better performance at requesting diagnostic tests. Most FPs (88%) requested the appropriate tests. FPs who scored better on requesting diagnostic tests spent an average of four minutes more with the patient. None of the participants detected the SP.

Conclusions: Clinical hepatomegaly is difficult to detect, even by well trained FPs. Senior doctors scored better on physical examination.

Key words: *Clinical skills, physical examination, professional competence, process assessment (health care), standardized patient*

Introduction

Visits by unannounced standardized patients (SP) randomly inserted into the daily practice seem to be the best method to assess actual performance in medical practice (1–7). These methods are expensive and require the physician's willingness to be assessed (8). Since 1991, many studies have evaluated professional performance and its relation to clinical outcome. Rethans et al., (9) compared what a doctor actually did in daily practice (performance) and what he or she was capable of doing in a formal examination (competence), showing better competence than performance. In another study, the extent to which clinical notes in the FP's medical records reflected their actual performance during patient encounters was assessed with SPs (10), proving that the audit of clinical notes was a weaker method of assessing quality of care. Other studies highlighted important differences between physicians when evaluating the same SP (11), but consistency when the same doctor evaluated the

Correspondence: Borrell-Carrió F. CAP Cornellà. C/Bellaterra 39-41 2ª Planta. 08940 Cornellà de Llobregat. Spain. Tel: 011-34-934711000. E-mail: fborrell.cp.ics@gencat.cat.

(Received 15 June 2010; accepted 8 December 2010)

ISSN 1381-4788 print/ISSN 1751-1402 online
DOI: 10.3109/13814788.2010.549223

same problem on different days (12). Most of this research has focused on history taking and communication skills (13), even though physical examination (PE) is a crucial part of the encounter, especially when a physical finding can guide to a relevant diagnosis (such as liver tumour, liver cirrhosis or fatty liver).

This study aimed at exploring the extent to which physicians were able to detect significant hepatomegaly in real practice, and whether this physical exam skill was related to the quality of history taking, physical examination, diagnostic testing and diagnostic accuracy. Our initial hypothesis was that physicians capable of documenting 'hepatomegaly' in the medical records would be those with better clinical skills. We selected hepatomegaly because it is a straightforward sign to detect on examination compared with other physical examination findings.

Methods

Study design

An observational study of physicians´ performance using an unannounced female standardized patient (SP) was carried out. The SP was covertly introduced into each physician's practice with the physician's prior consent to assess actual practice (14).

Recruitment of health centres, physicians and encounters

The Southern Metropolitan Area of Barcelona has 12 accredited health centres for teaching Family Medicine Residency Program, and 41 non-accredited. We randomly selected 10 accredited centres, and two non-accredited, with 159 physicians involved of which 104 were interested in participating in our study; 61 of them signed the authorization. To be included in the study, doctors should remain in their posts for at least 12 months prior to the study.

An agreement to receive a SP presenting as a real patient on an unannounced day was signed by 61 doctors. They also gave consent to audiotape the encounter and to have their clinical notes reviewed. In case they uncovered the SP during the visit, the doctors and the SP agreed to notify to investigators.

The clinical encounters were excluded from the study if any of the following circumstances occurred: the SP visit was uncovered, if part of the visit was completed by another professional, if the SP did not follow the standard script, and poor audiotape quality.

The standardized patient

A single SP was used for all of the encounters. The SP was a 48-year-old female actor who had a significant hepatomegaly of 6 cm below the ribs in the midclavicular line, verified by ultrasonography (15), without jaundice or other stigmata of liver disease. She signed a confidentiality statement and was provided with a false identity, including a clinical record with previous clinical notes as if she was a regular patient in the health centre. The SP waited for her visit as any ordinary patient, and during the visit she recorded the meeting with a hidden microphone. When the visit ended, she filled out a questionnaire on the physical examination that had been carried out.

The training of the SP (two months) was conducted according to the methodology of the Catalan Institute of Health Studies (IES, collaborating institution of the ECFMG) (16). Throughout the six-month duration of the study, four meetings with the SP were held to assess the reliability of her performance (intra-rater reliability). The actress played a middle-aged woman with a 10-day pain in the right upper abdomen that increased after meals, with urine a little bit darker than usual and with occasional nausea. She was trained to convey the belief that she was feeling well, without important concerns. If the doctor reached a diagnosis only with her clinical symptoms without performing a clinical examination and he/she prescribed a medication, the patient would say 'don't you need to explore my abdomen (tummy) today?' However, if despite this comment the doctor carried on prescribing, the patient would not insist. If this verbal prompt had been given, it must be reported by SP after the interview.

Data collection and analysis

Immediately after each visit we collected the information shown in Table I. All the medical notes, audiotapes and test requests had to be intelligible to be included. Audiotapes were immediately analysed to detect technical errors (poor quality), and also to check SP's accuracy portraying her role. Quality standards were based on clinical records, (not audiotapes), and they were developed by a group of 6 experienced doctors (2 of them gastroenterologists). History taking, physical examination, final tests requested, and diagnosis impression were evaluated using a weighted score (zero to ten, see Table II) based on ECFMG standards and other sources (17–19). Doctors could have detected the hepatomegaly but they could choose not to write the diagnostic impression until ultrasound confirmation, so we reviewed all reports to the radiologist searching for 'hepatomegaly' or another similar word.

Definition of variables

Doctors older than 45 years old were considered senior doctors.

Table I. Information collected at the end of each encounter.

1. Report of physical examination (SP)
2. Audio recording of the meeting (SP)
3. Physician's personal data and clinical experience
4. Years in his/her job (seniority)
5. If prior training in a family medicine residency programme
6. If the physician was accredited as teaching tutor
7. If a student or resident were present at the encounter
8. Duration of interview
9. Clinical notes (medical record)
10. Tests requested
11. Treatment
12. Total number of visits (and average length) made by doctor the same day he/she visited our SP

Note: Data for items 1, 2 and 7 were collected by the SP.

Table II. Quality standard and percentage of compliance.

	Score	*n*	%
History taking			
1. Since when ... duration	1	54	94.7
2. How is the pain? Characteristics	1	46	80.7
3. Radiating pain	1	15	26.3
4. Improves or worsens with...?	1	37	64.9
5. Where is the pain located?	1	57	100
6. High temperature?	1	29	50.9
7. Any previous disease or surgical procedure	0.5	42	73.7
8. Vomiting or diarrhoea	0.5	48	84.2
9. Allergies?	0.5	39	68.4
10. Medication she was taking	0.5	16	28.1
11. Impact on daily activities	0.5	18	31.6
12. Smoking or drinking	0.5	27	47.4
13. Changes in her urine colour	1	31	54.4
Physical examination			
14. Superficial abdominal palpation of 4 quadrants	2	57	100
15. Deep abdominal palpation in inspiration+expiration (Murphy sign).	2	24	42.1
16. Rebound sign	1	12	21.1
17. Abdominal percussion (at least once)	1	10	17.5
18. Bilateral lung auscultation (at least one location)	1	11	19.3
19. Cardiac auscultation	1	8	14
20. Bilateral lumbar fist percussion	1	8	14
21. Abdominal auscultation (at least once with stethoscope)	0.5	12	21.1
22. Jugular engorgement	0.5	4	7
Final procedures (tests requested)			
23. ALT and AST and GGT and bilirubin	3	27	47.4
24. ALT+/−bilirubin or any other combination with AST or GGT	2	10	17.5
25. Abdominal ultrasound	3	30	52.6
26. B and C Hepatitis serology	1	50	87.7
27. Full blood count	1	38	66.7
28. Creatinine	1	38	66.7
29. Glycaemia	1	38	66.7
Diagnostic orientation (only one option)			
30. Pain in right hypochondrium with hepatomegaly	10	4	7
31. Hepatomegaly	7	0	0
32. Pain in right hypochondrium	5	46	80.7
33. No diagnostic orientation or misdiagnosis	0	7	12.3

We evaluated the quality of the encounters based on the following four points: (a) basic clinical skills in history taking and physical examination; (b) diagnostic impression; (c) final tests requested at the end of the interview; and (d) global quality of the encounter (Table III).

We defined 'sufficient history taking' as an encounter with equal or more than 5 points over 10, but only if the 'how,' 'when' and 'where' of abdominal discomfort were recorded correctly (items 1, 2 and 5 in Table II).

We defined 'sufficient physical examination' as an encounter with equal or more than 5 points over 10, but only if the abdomen was examined (item 14, Table II). We defined 'sufficient tests requested' as an encounter with at least a blood test or an ultrasonography requested.

Owing to the above definitions, the overall quality of the encounter was defined as 'correct' if the encounter had sufficient history taking, physical examination (PE) and final test requests. We defined an 'encounter in need of improvement' as the encounter with adequate final test requests, but with an insufficient history taking or PE. Finally, we considered 'unexpected success' as any encounter with adequate final test requests but with inappropriate history taking and PE, and 'unacceptable' the encounter with inadequate final test requests (Table IV).

Finally, we analysed the relationship between the overall quality of the encounter and the diagnostic impression.

Statistical analyses

Participant characteristics and clinical skills were measured and summarized using frequencies (percentages) for binary variables, with the mean and the standard deviation (SD) for normally distributed continuous variables. We compared these characteristics using the χ^2 or Fisher test for binary variables and the independent two-sample *t* test for normally distributed continuous variables, Welch test for normally distributed variables with unequal variances, or non-parametric tests (Mann Whitney's U test). We assessed differences among groups by using a one-way ANOVA analysis of variance for continuous variables, with Bonferroni's correction for multiple comparisons, or its non-parametric equivalent

Table III. Quality features: basic skills and diagnoses.

A. Basic skills
   1. Sufficient history taking: equal or more than 5 points, but only if the 'how', 'when' and 'where' of abdominal discomfort were recorded.
   2. Sufficient physical examination: equal or more than 5 points, but only if the abdomen was examined.
   3. Sufficient diagnostic tests: at least either blood tests or ultrasonography requested.

B. Diagnostic impression
   (a) no diagnosis
   (b) abdominal pain and/or upper abdominal pain
   (c) hepatomegaly

C. Diagnostic tests
   (a) 'appropriate': when a blood test and an ultrasound were requested
   (b) 'in need of improvement': when only either the blood test or the ultrasound were requested
   (c) 'inappropriate': when neither were requested.

(Jonckheere-Terpstra). To compare two quantitative variables we used univariate linear regression, verifying the conditions of application and the residual normality. We calculated kappa coefficient (intra-rater reliability) of SP, when assessing the physical examination performed by trainer doctors.

Ethical requirements

The study was approved by the Research Ethics Committee of the Jordi Gol i Gurina Foundation (Catalan Health Institute).

Results

The study invited 104 physicians to participate. The encounter with the SP was completed by 61 physicians, but 4 were rejected because of poor audiotape quality. The remaining 57 encounters constituted the study encounters. No differences were found in either age or sex between the participants and the overall population of physicians in the Southern Metropolitan Area of Barcelona. 45 of the physicians taking part in the study (78.9%) were accredited as Family Medicine tutors, (they were responsible for tutoring family medicine residents). The baseline characteristics of the participants are shown in Table V.

Physicians completed an average of 34.5 ± 9.8 visits per day and an average of 7 minutes per patient (measured on the day of the SP encounter). There was no significant association between the number of visits and the physician's age ($P = 0.53$). The average time spent with our SP was 9.28 min ± 5.5, exceeding almost by 3 min the average time for other encounters of the same day (7 ± 3.4 min). The encounters with older doctors were longer ($P < 0.001$). Physicians never reported having uncovered the SP. Intra-rater reliability of our SP assessing clinical examinations was K = 0.61 to 0.63.

Basic clinical skills

Physicians scored better at history taking (84.24% obtained 'sufficient history taking') than at physical examination (26.35% obtained 'sufficient PE') ($P < 0.01$). The complete dataset is summarized in Table VI. Two participants obtained the highest score in history taking and one in PE.

All participants performed at least superficial palpation of the abdomen not needing verbal SP prompt to perform it, and 12 carried out an auscultation. 6 participants (10%) performed a more complete physical examination, which included cardiopulmonary examination, jugular engorgement and costovertebral angle tenderness. 12 (21%) physicians tested for abdominal rebound tenderness and 24 (42%) looked for Murphy's sign. Senior doctors had better scores on history taking ($P = 0.02$) and physical examination ($P = 0.03$). Longer visits were also associated with higher scores ($P = 0.001$).

Fewer total visits during the day correlated with better history taking scores ($P = 0.03$). Doctors with residency training had better scores on physical examination ($P = 0.026$) and on history taking, but this association fell just outside statistical significance ($P = 0.053$). Physicians working at health centres accredited for teaching obtained better scores in history taking ($P = 0.03$). Figure 1 summarizes these findings.

Table IV. Overall quality of the encounter.

	History taking	Physical examination	Final tests requested	*n* (%)
Correct encounter	> or = 5 points	> or = 5 points	Blood tests or ultrasound	17 (30)
Encounter in need of improvement	History taking or PE less than 5 points		Blood tests or ultrasound	28 (49)
Encounter showing unexpected success	< 5 points	< 5 points	Blood tests or ultrasound	5 (8.7)
Unacceptable	< or > 5 points		Neither blood tests or ultrasound	7 (12.2)

Table V. Characteristics of the physicians who participated in the study.

	Mean	SD
Physician's age	40.60	5.95
Experience in primary care (years)	13.60	5.70
Time in the workplace (years)	9.58	5.53
Interview duration (minutes)[a]	9.28	5.56
Number of visits[b]	34.58	9.80

[a]The average time spent with our SP.
[b]Number of visits on the day of the encounter.

Diagnostic impression

Abdominal pain was the diagnosis given by 80.7% of the family physicians, 7% (4 participants) documented abdominal pain with hepatomegaly, and 12% did not report any diagnosis. Senior physicians obtained better scores for diagnostic impression ($P = 0.04$).

The 4 participants who reported hepatomegaly did not differ significantly from the rest in their scores on history taking and physical examination. The only physicians who informed the radiologist of the presence of hepatomegaly were these 4 doctors, all of them accredited as Family Medicine tutors.

The association between the duration of the encounter and the accuracy of the diagnosis was not statistically significant, nor between the scores of history taking and physical examination.

Requested final tests

66% of participants requested a full blood count and blood chemistry tests, which included a liver panel. An abdominal ultrasonography was requested by 52%. Both blood tests and ultrasonography were requested by 18 (31.5%) physicians. 7 participants (12%) did not request further tests.

The encounters for those who ordered both types of test lasted 12.6 ± 3.2 min, compared to 8.14 ± 2.67 minutes for those who did not order either type of test ($P < 0.05$), a difference of almost four minutes.

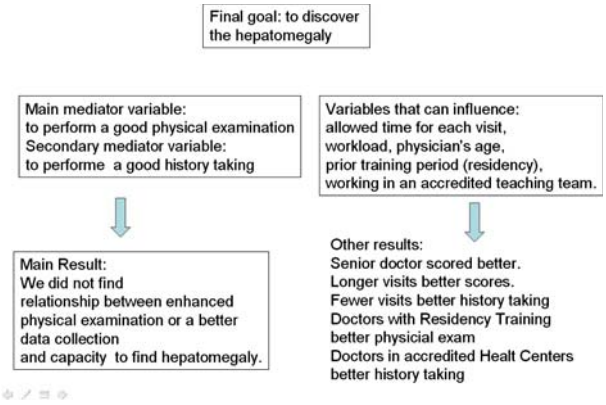

Figure 1. Main results.

Overall quality

The overall quality of the encounter is shown in Table IV. In the case of two of the seven encounters considered 'unacceptable,' the physicians had shown good clinical skills but had not completed the interview with the request of diagnostic tests appropriate to the clinical case.

Discussion

This study was carried out mainly with accredited doctors (Family Medicine Residency). The three major findings that resulted from this study were: (a) only 4 of the 57 FPs who examined the patient detected the hepatomegaly; (b) FPs performed better at history taking (84.24%) than at PE (26.35%) and no correlation was found between the two; (c) diagnostic accuracy was associated with older age, years of experience, history taking skills and better performance at requesting diagnostic tests. Most FPs (88%) requested the appropriate tests. FPs who scored better on requesting diagnostic tests spent an average of four minutes more with the patient. None of the participants detected the SP. The following variables were associated with a successful encounter: longer duration of the interview, lower number of visits per day, older physician's age, longer time in the same workplace, and working in an accredited health centre. More experienced and older physicians were more likely to be tutors of the Family Medicine Residency Program and to work in accredited health centres.

If we consider the request of appropriate diagnostic tests as the best indication of the search for a more accurate diagnosis, most of encounters met this criterion. A major concern was that 5 of the 57 participants (almost 10%) scored poorly both on history taking and physical examination, but requested abdominal ultrasonography and/or blood tests. We have labelled these cases as 'unexpected success.' It is possible that these physicians routinely order tests without a sound basis for doing so, perhaps they were practicing defensive medicine (20), or they might have detected the hepatomegaly without recording their finding. By contrast, two encounters illustrated the opposite phenomenon, where good clinical skills were not followed by any further diagnostic testing.

Detecting hepatomegaly

Only 4 doctors detected hepatomegaly. These four professionals did not score above average on clinical

skills; therefore, one of our hypotheses was rejected: detection of hepatomegaly was neither a marker for quality of overall clinical skills nor associated with quality of subsequent care. Detection of hepatomegaly might depend upon other unmeasured factors.

First, hepatomegaly may be difficult to detect in clinical practice. There is evidence from a number of studies that the detection of hepatomegaly depends on several variables. Sajjad et al. found that liver volume calculated on the basis of CT scans correlated to liver size based on the hepatologist's clinical exam (21). Zoli et al. noted that correlation between physical examination and ultrasonographic evaluation of liver size was low in healthy controls, but excellent for those who had hepatomegaly (22). However, other authors agreed that PE is not a reliable method to detect hepatomegaly (23–25). In our case, highly qualified doctors in an environment of time constraint were unable to detect a clinically obvious hepatomegaly. Were the clinicians mainly interested in whether the point of maximum tenderness was epigastric or over the gall bladder and thus failing to notice the liver edge? Alternatively, were they avoiding hurting the patient with a clinical examination, which was unlikely to influence the further diagnostic tests that they had already decided to request?

Second, doctors may consider physical examination data less important than verbal information and tests results. Eric Cassell has labelled as 'soft data' the facts arising from the history, as opposed to the 'hard data' that arise from tests (26). Currently, physical examination data may be regarded as even 'softer' than verbal data, when there are available 'objective tests' such as CT scans and ultrasound.

Third, the PE was conducted in a psychological context where our SP was not transmitting a special concern about her liver, but rather the expectation that everything was fine. Typically, in this context a physician might think 'I do not trust my finding because I want the patient to be healthy', or 'I won't find anything relevant in the PE'. Even if the doctor might perceive the hepatomegaly, he/she might think that 'the patient may have hepatomegaly, but I did not expect that based on the symptoms, so it is better to perform an ultrasound without documenting this finding on the clinical record, to avoid confusing other colleagues or even put my reputation at risk with false information'. Campbell and Croskerry call this phenomenon 'framing effect,' a decision being influenced by the way in which the scenario is presented or 'framed,' and 'ascertainment effect,' when thoughts are preconceived by expectations (27,28). It can be caused either by the physician not wanting to stress the patient, or by the concern about his/her own reliability and reputation.

Table VI. Assessment of clinical skills in 57 physicians.[a]

	Mean	SD	Minimum	Maximum
History taking (maximum score: 10)	6.39	1.59	2	9.5
Physical examination (maximum score: 10)	3.84	1.40	2	8
Final tests score (maximum score: 10)	6.23	3.21	0	10
Diagnosis score (maximum score: 10)	4.74	2.20	0	10

[a]A panel of experts weighted the diagnostic and therapeutic behaviour of physicians.

Limitations

This study has some limitations. First, it cannot be considered representative of the Spanish Primary Care system because a large percentage of doctors were accredited for postgraduate teaching. Second, other studies with more participants may discover a stronger link between clinical skills and the ability to find specific physical signs. Third, our approach allows us to measure errors of omission, but not errors of commission. Therefore, it was not possible to determine the extent to which hepatomegaly was diagnosed incorrectly in a patient with abdominal pain but without hepatomegaly, or the extent to which unnecessary tests were requested. The results may not be applicable to settings where physicians can spend more time with patients, even though the relationship between encounter time and the quality of the physical examination has not been established. None of the participants reported that they had discovered the SP, but we do not know if any of them suspected her presence. This fact could have been found out by calling each participant a few days after the visit of the SP. Finally, the patient was a 'new' patient and the results might not apply in the context of a more established relationship. It has not been established if a physical examination differs between a patient who the doctor knows well and a new unexpected patient.

Implications

Future studies should determine the extent to which highly trained doctors are able to detect signs such as murmurs, thyroid nodules or pulmonary abnormalities in their daily practice, and what factors influence their performance. The frequency of undetected physical findings should be addressed to ascertain its effect on quality of care and to guide efforts to improve clinical performance. Since many countries are considering revalidation programmes, Incognito SP would be useful to assess basic clinical skills, especially if SP reliability is considered (29).

As clinical encounters become more pressured and complex, the detection and documentation of obvious physical findings such as hepatomegaly is likely to suffer. With the availability of sophisticated diagnostic testing, physicians may discount the value of physical findings that might alter clinical decisions. Since many patients in primary care present with symptoms but few clinically important physical findings, the teaching of physical examination skills in Family Medicine Residency and Continuing Medical Education Programs should promote skills to overcome the 'ascertainment effect,' and other attitudinal barriers to accurate physical diagnosis. This may be particularly important for the less experienced young physicians.

Acknowledgments

The authors should like to thank the members of the nominal expert groups conducted by Professor Rafael Azagra (MD), Jordi Cebrià Andreu (MD) and Josep Maria Bosch Fontcuberta (MD). The authors are grateful to Dr José María Martínez Carretero from the Catalan Institute of Health Studies (IES) and Dr Ronald M. Epstein (Rochester University), who provided important methodological support, and to the Health Centre directors, Dr X. Bayona and Dr C. Pujol, as well as Professor Randol Barker (MD), Dr Elena Barquero and Dr Leonore Novotny (MD), who helped with the editing, Fermín Quesada and Pablo Bonal participated in writing the script of the Standardized Patient. The authors also thank the 61 doctors who volunteered to participate in the study, with no other compensation than to improve our knowledge on clinical practice. The authors are grateful to the IDIAP Jordi Gol for funding the translation of the study into English.

This research was supported by FISS grant number.02/10054, 18 December 2002 (Prados JA Principal Investigator), and a 2003 REAP grant, Carlos III Health Institute, 'Assessment of Health Technology' (number REAP-22/2002).

Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

References

1. Hays RB, Davies HA, Beard JD, Caldon LJ, Farmer EA, Finucane PM, et al. Selecting performance assessment methods for experienced physicians. Med Educ. 2002;36:910–7.
2. Barragán N, Violan C, Martín Cantera C, Ferrer-Vidal Cortella D, González Algas J. Designing a method for the evaluation of clinical competence in primary care. Aten Primaria 2000;26:590–4.
3. Gorter S, Rethans JJ, van der Heijde D, Scherpbier A, Houben H, van der Vleuten C, et al. Reproducibility of clinical performance assessment in practice using incognito standardized patients. Med Educ. 2002;36:827–32.
4. Tamblyn RM, Klass DJ, Schnabl GK, Kopelow ML. The accuracy of standardized patient presentation. Med Educ. 1991;25:100–9.
5. Badger LW, deGruy F, Hartman J, et al. Stability of standardized patients' performance in a study of clinical decision making. Fam Med. 1995;27:126–31.
6. McLeod PJ, Tamblyn RM Gayton D, Roland G, Snell L, Berkson L, et al. Use of standardized patients to assess between-physician variations in resource utilization JAMA 1997;278:1164–8.
7. Luck, Jeff. Using standardised patients to measure physicians' practice: Validation study using audio recordings. Br Med J. 2002;325:679–83.
8. Tamblyn, Robyn M. Use of standardized patients in the assessment of medical practice Canadian Medical Association; Association Médicale Canadienne 1998;158:205–7.
9. Rethans JJ, Sturmans F, Drop R, van der Vleuten C, Hobus P. Does competence of general practitioners predict their performance? Comparison between examination setting and actual practices Br Med J. 1991;303:1377–80.
10. Rethans JJ, Martin E, Metsemakers J. To what extent do clinical notes by general practitioners reflect actual medical performance? A study using simulated patients. Br J Gener Pract. 1994;44:153–6.
11. McLeod PJ, Tamblyn RM, Gayton D, Grad R, Snell L, Berkson L, et al. Use of standardized patients to assess between-physician variations in resource utilization JAMA 1997;278:1164–8.
12. Rethans JJ, Saebu L. Do general practitioners act consistently in real practice when they meet the same patient twice? Examination of intradoctor variation using standardised (simulated) patients. Br Med J. 1997;314:1170–3.
13. van den Brink-Muïnen A, van Dulmen AM, Bensing JM , Maaroos HI, Plawecka L, Krol ZJ, et al. The Eurocommunication Study, Utrecht, Nivel, 2003. Available at http://www.nivel.nl/pdf/EurocommunicationII.pdf (accessed 2 December 2010).
14. Miller G. The assessment of clinical skills/ competence/ performance. Acad Med. 1992;65:63S–7S.
15. Michie C, Alu S, Wild K, Hampsheir R, Chabonnaud P, Harvey D. Should we estimate liver span in the right midclavicular line or the midline? J Paediatr Child Health 1995;31:241–4.
16. ECFMG Educational Comission for Foreign Medical Graduates. Philadelphia, ECMFG. Available at "http://www.ecfmg.org/cert/index.html" http://www.ecfmg.org/cert/index.html (accessed 2 October 2010).
17. Trowbridge RL, Rutkowski NK, Shojania KG. Does this patient have acute cholecystitis? JAMA 2003;289:80–6.
18. Berger MY, van der Velden JJ, Lijmer JG, de Kort H, Prins A, Bohnen AM. Abdominal symptoms: Do they predict gallstones? A systematic review. Scand J Gastroenterol. 2000; 35:70–6.
19. McGee S. Evidence-based physical diagnosis. Philadelphia, Pa: WB Saunders; 2001.
20. Borrell F, Paez C, Suñol R, Orrego C, Gil N, Marti M. Clinical errors and adverse events: a perception from primary care physicians. Aten Primaria 2006;38:25–32.
21. Sajjad S, Garcia M, Malik A, Van Thiel DH. An assessment of the accuracy of hepatic and splenic size based upon a clinician's physical examination, a radiologist's impression and the actual liver and spleen volumes calculated by CT scanning. Dig Dis Sci. 2008;53:1946–50.

22. Zoli M, Magalotti D, Grimaldi M, Gueli C, Marchesini G, Pisi E. Physical examination of the liver: Is it still worth it? Am J Gastroenterol. 1995;90:1428–32.
23. Mangione S. Physical diagnosis secrets. Philadelphia, Pa: Hanley and Belffus Inc.; 2000. pp. 347–53.
24. Ariel IM, Briceno M. The disparity of the size of the liver as determined by physical examination and by hepatic gamma scanning in 504 patients. Med Pediatr Oncol. 1976; 2:69–73.
25. Tucker WN, Saab S, Rickman LS, Mathews WC. The scratch test is unreliable for detecting the liver edge. J Clin Gastroenterol. 1997;25:410–4.
26. Cassell EJ. The nature of suffering and the goals of medicine. New York: Oxford University Press; 1991. p. 96.
27. Campbell SG, Croskerry P, Bond WF. Profiles in patient safety: A 'perfect storm' in the emergency department. Acad Emerg Med. 2007;14:743–9.
28. Leblanc VR, Brooks LR, Norman GR. Believing is seeing: The influence of a diagnostic hypothesis on the interpretation of clinical features. Academic Med. 2002;77:Suppl. 67–9.
29. Rethans JJ, Gorter S, Bokken L, Morrison L. Unannounced standardised patients in real practice: A systematic literature review. Med Educ. 2007;41:537–49.