Do general practitioners act consistently in real practice when they meet the same patient twice? Examination of intradoctor variation using standardised (simulated) patients

Jan-Joost Rethans, Lars Saebu

Centre for Research on Quality Assurance in General Practice, Department of General Practice, University of Limburg, PO Box 616, 6200 MD Netherlands

Jan-Joost Rethans, *assistant professor*

Department of General Practice, University of Trondheim, Norway

Lars Saebu, *general practitioner*

Correspondence to: Dr Rethans.

BMJ 1997;314:1170–3

Abstract

Objective: To assess the variation within individual general practitioners facing the same problem twice in actual practice under unbiased conditions.
Design: General practitioners were consulted during normal surgery hours by a standardised patient portraying a patient with angina pectoris. Six weeks later the same general practitioners were consulted again by a similar standardised patient portraying a similar case. The patients reported on the consultations.
Setting: Trondheim, Norway.
Subjects: Of 87 general practitioners invited by letter, 28 (32%) agreed to participate without hesitation; nine others (10%) wanted more information before consenting. From these 24 were selected and visited.
Main outcome measures: Number of actions undertaken from a guideline in both rounds of consultations. Duration of consultations.
Results: The mean (range, interquartile range) guideline score, total score, and duration of consultation were not significantly different between the first and second patient encounters for the group as a whole. For individual doctors the mean (SD) difference was −0.09 (3.36) for the guideline score, 0.30 (8.1) for the total score, and −0.87 (9.01) for consultation time.
Conclusions: The study shows that assessment of performance in real practice for a group of general practitioners is consistent from the first round of consultations to the second round. However, significant variation occurs in performance of individual physicians.

Introduction

Variation between doctors is a reflection of the individual's art of medicine but may also be a threat to the scientific basis of practice.[1] Variation in performance may be studied between countries,[2] regions,[3] hospitals,[4] practices, and doctors.[5 6] To try to minimise the variation between doctors national bodies have produced guidelines for good medical practice, both for medical specialties and general practice.[7]

Variation of performance is an important consideration in assessment of competence of general practitioners. The performance of doctors varies across different medical problems.[8] For example, a doctor's performance in dealing with a patient with a urinary tract infection does not predict his or her performance with a patient with diarrhoea. This phenomenon has been labelled content specificity[9] and is one of the main reasons why doctors are examined on different areas of medicine and with different problems.[10]

When assessing doctors' management of a single problem we need to know whether the doctor consistently performs to the assessed standard. Intradoctor (or intraobserver) variation may lead to different results when a doctor is faced with an identical problem twice. Few studies have addressed this problem, and their results are ambiguous. When medical students and specialists were presented with a clinical problem twice by standardised patients the correlation was only 0.60 between the two presentations.[11] With medical students test-retest reliability on the same station of an objective structured clinical examination was 0.66-0.88.[8] In a study with two independent clinical assessments by a single clinician (three months apart) of the same set of 100 fundus photographs, 88 of 100 patients received identical assessment.[12] Repetition of identical tasks by medical students within the same exam did not improve their scores.[13] However, these studies were run in examination laboratory settings and may be biased since the subjects knew they were being tested and were likely to recognise the second presentation. In addition, performance under examination circumstances may differ from performance in practice.[14] To overcome these problems we did a study to find out whether and to what extent intradoctor variation—that is the variation within doctors facing a similar problem twice—in real life general practice exists under unbiased conditions.

Subjects and methods

We used standardised patients for this study because this method has proved to be reliable, valid, feasible, and acceptable in general practice.[15 16] A standardised

role of an elderly patient with angina pectoris was constructed. The role focused on the medical history with no abnormal physical signs and normal laboratory and electrocardiographic findings. Two healthy women, aged 69 and 70, were selected as standardised patients and paid to participate. They signed written consent to keep all medical and personal information about the general practitioners in the project strictly for research purposes.

The patients were trained to present a standardised complaint and to score history taking, physical and laboratory examination, instructions given to the patient, treatment, and follow up against a guideline on managing angina pectoris. This guideline was based on relevant general practice literature (such as the guidelines of the Dutch College of General Practitioners) and discussed with two experienced general practitioners and an experienced cardiologist.[17] The guideline contained only items considered necessary to manage angina pectoris as presented by the standardised patients.

To ensure the reliability and consistency of scoring by the standardised patients we used standard procedures.[18 19] In brief, reports of standardised patients during training (before and between the first and second round) were compared with reports of a panel of doctors about the same consultation. These reliability and consistency κ scores were 0.85 (maximum $\kappa = 1.0$). Several scores were used to assess the performance of the general practitioners. Firstly, a guideline score—that is, the number of items of the guideline performed by the general practitioner in a consultation. Secondly, a total score—that is, all items (guideline plus non-guideline items) performed by a general practitioner in a consultation. Patients also recorded the duration of visits in minutes using a wristwatch with stopwatch facilities.

One year before the actual visits all 87 general practitioners in Trondheim, Norway, were informed by letter about the objectives of the study and invited to give written acceptance of standardised patients into their practices. The dates, number, and content of the visits were not mentioned. For budgetary reasons it was decided beforehand that 24 general practitioners would participate.

Patients took their original health insurance identifying papers and enlisted in the practices of the selected general practitioners by using techniques reported earlier.[16 20] The general practitioners were visited by the standardised patients in two rounds in March and May 1994. Patient A visited 12 of them in the first round and the other 12 in the second round, while patient B visited the doctors in the reverse order. All participating general practitioners were presented with similar standardised presentations twice.

The Wilcoxon signed rank test (paired design) was used to look for differences in the doctors' performances in the first and second round. To assess intradoctor variation, the scores of individual doctors on the two rounds were analysed by the Bland and Altman method.[21] The Wilcoxon signed rank test (paired design) was used to assess whether the two standardised patients showed any consistent difference in the way they scored for consultations for the guideline score (the most important score).

Table 1 Number of general practitioners performing items listed in the guideline for angina pectoris during consultations with two standardised patients

	First consultation (n=23)	Second consultation (n=23)
History:		
Onset of pain	23	23
Location of pain, radiation	22	23
Nature of pain	22	20
Duration of pain	20	19
Physical activity initiates pain	21	23
Rest relieves pain	22	21
Other factor aggravates pain	15	11
Previous diseases	20	17
Smoking habits	16	15
Physical examination:		
Blood pressure	21	23
Pulse	7	9
Heart auscultation	21	22
Lung auscultation	18	18
Laboratory examination:		
Haemoglobin	20	15
Treatment/patient information:		
Tell diagnosis	23	22
Nitroglycerin prescription	19	20
Symptoms on worsening	3	6
Lifestyle advice	7	11
Correct use of prescription	18	20
Side effect of prescription	14	17
Follow up:		
Control visit	18	17

Results

Of the 87 doctors asked to participate, 53 (61%) replied. Twenty-eight (32%) answered yes without any further information; nine others asked for more information before agreeing.We selected 24 doctors from those that agreed. After a visit in the second round one general practitioner reported having detected the patient. This left 23 general practitioners and 46 visits for analysis.

Table 1 shows the performance of general practitioners for each item of the guideline in each consultation. Table 2 gives the guideline and total scores and consultation times in the two rounds. We found no significant difference between the first and second round for any of the items or scores assessed. However, to assess intradoctor variation the scores of individual physicians during the first round have to be compared with their individual scores during the second round. This is indicated by the standard deviations in table 2. For example, the standard deviation of the guideline score is 3.36, suggesting that the average within doctor difference for number of guideline items scored is around 3; the average inconsistency in total score is around 8 and the average difference in length of consultation around 9 minutes. These data indicate substantial intradoctor variation between the two rounds. Means (interquartile range) of the guideline scores for the two standardised patients were 16.22 (14 to 19) and 16.04 (14 to 18). These were not significantly different by Wilcoxon signed rank test (paired design), suggesting the two patients showed no consistent differences.

Table 2 Mean number of actions scored by standardised patients for consultations with 23 general practitioners and mean differences between two consultations

	Mean	Range	Interquartile Range	Mean (SD) difference	95% CI
Guideline score:				–0.09 (3.36)	–6.81 to 6.81
First consultation	16.0	11-20	14-18		
Second consultation	16.2	10-21	14-19		
Total score:				0.30 (8.10)	–15.9 to 16.5
First consultation	28.2	13-39	24-32		
Second consultation	27.9	14-47	17-34		
Consultation time (min):				–0.87 (9.01)	–18.89 to 17.15
First consultation	16.8	10-40	14-20		
Second consultation	17.6	10-40	14-20		

Discussion

We believe that this is the first study of intradoctor variation in real practice using standardised patients presenting similar problems. This design is the only way to ensure subjects do not know they are being observed, thus removing an important source of bias. In examination or test settings subjects would easily spot the second presentation.

Clearly, this study has some limitations. There were only 23 general practitioners and only one standardised problem was presented twice, resulting in 46 consultations. However, the few studies set in examination conditions that have used more comparisons have produced ambiguous results. Getting funding for a larger study incorporating more patients and comparisons would be difficult until a pilot study such as this one has been done. Only 32% of the doctors approached agreed to participate without further hesitation, which may mean that the participants reflect a more competent sample of general practitioners.

We believe, however, that our results are valid as the doctors were unaware that they were being assessed. The results show that the assessment of performance was consistent from the first round of consultations to the second round. This means that anyone wanting to give feedback to a group of practitioners on their management of a particular problem would probably need to do only one assessment. However, for assessment of performance of a single physician the results are quite different. We found appreciable intradoctor variation in the management of the two patients. Analysis showed that the personality of the two standardised patients had no effect on the results. A further study using more problems and more presentations of the same problem would give a better indication of whether intradoctor variation is a problem. This may in turn lead to reassessment of the way cases of sampled for examination and licensing of doctors and for quality assessment.

Does the variation matter

A further question is to what extent the intradoctor variation found in this study is a problem? Different scorings (for example a weighted score) may have resulted in different results. The panel which constructed the guideline thought all guideline items were essential and therefore distinguished only between these and non-guideline items. Earlier studies with standardised patients that used more differentiated scores (obligatory, intermediate, and superfluous items) found no differences between these scores.[14] Our data should act as a stimulus for careful thinking about differentiated scores of guidelines. Some may argue that only evidence based items are important to record in this type of study, but in general practice this might result in only one or two items per case. All other items are then reflections of the individual performance of a doctor.

To try to find an explanation for the differences in the results of individual general practitioners between the two consultations we carried out some secondary analyses—for example, to determine if there were different outcomes for visits before or after lunch. These analyses all gave negative results. Our data showed two consultations of 40 minutes, which is unusually long. Although we do not know exactly what happened in these consultations, it seems likely that the doctor received a telephone call during these visits. Since these 40 minutes conversations could have a relatively large effect in the inconsistency in the duration of consultation we performed the same calculations for the duration of visits without these consultations and by substituting the 40 minutes by 30 minutes (30 minutes being the second longest consultation). Although the standard deviations were reduced to 5.78 (without these visits) and to 6.99 (for 30 minutes), the conclusions remained the same. We discussed our results with several groups of general practitioners and received reactions such as "this is just real practice and so it should be" or "on Monday after a sleepless night doctors perform differently from Tuesdays after a good rest."

In conclusion this study shows that intradoctor variation occurs in day to day practice. The implications of this variation remain undetermined, and documentation of what is really going on in doctors' surgeries remains a great challenge.

We thank Arnold Kester (department of biostatistics, University of Limburg) and the General Practitioners Writers Association (in particular Professor Robin Hull) for their help with this paper.

Funding: Norwegian Fund for Quality Assurance (Kvalitetssikringsfondet), grant number 93007.

Conflict of interest: None.

1 Anderson TF, Mooney G, eds. *The challenge of medical practice variation.* London:Macmillan, l990.

2 McPherson K, Strong PM, Epstein A, Jones L. Regional variations in the use of common surgical procedures: within and between England and Wales, Canada and the United States of America. *Soc Sci Med* 1981;18:273-88.

3 Wennberg J. McPherson K, Caper P. Will payment based on diagnostic-related groups control hospital costs? *N Engl J Med* 1984;311:295-300.

Key messages

- Variation in the performance of doctors is a potential problem in ensuring patients receive agreed best standards of care
- This study assesses the intradoctor variation in treating two standardised patients presenting with similar conditions in real practice
- For a group of general practitioners performance in the two consultations was consistent
- The performance of individual doctors differed when facing the same problem twice

4 McPherson K. Variation in hospital rates: why and how to study them. In: Ham C, ed. *Health care variations: assessing the evidence*. London: Kings Fund Institute, 1988:120-34.
5 Marinus A. *Inter-doktervariatie in de huisartspraktijk*. Meppel: Krips Repro, 1993. (Dissertation with summary in English.)
6 Rethans JJ, Sturmans F, Drop R, Vleuten vd C. Assessment of the performance of general practitioners by the use of standardised (simulated) patients. *Br J Gen Pract* 1991;41:97-9.
7 Grimshaw J, Russell I. Achieving health gain through clinical guidelines. l. Developing scientific valid guidelines. *Quality in Health Care* 1993;2:243-8.
8 Roberst J, Norman G. Reliability and learning from the objective structured clinical examination. *Med Educ* 1990;24:219-23.
9 Elstein A, Shulman L, Sprafka S. *Medical problem solving*. Cambridge, MA: Harvard University Press, 1978.
10 Newble D, Jolly B, Wakeford R, eds. *The certification and recertification of doctors. Issues in the assessment of clinical competence*. Cambridge: Cambridge University Press, 1994.
11 Norman GR, Tugwell P, Feightner JW, Muzzin LJ, Jacoby LL. Knowledge and clinical problem-solving. *Med Educ* 1985;19:344-56.
12 Aoki N Horibe H, Ohno Y, Hayakawa N, Kondo R. Epidemiological evaluation of funduscopic findings in cerebral diseases. III. Observer variability and reproducibility for funduscopic findings. *Jpn Circ J* 1977:41:11.
13 Hodder RV, Rivington RN, Calcutt LE, Hart IR. The effectiveness of immediate feedback during the objective structured clinical examination. *Med Educ* 1989;23:184-8.
14 Rethans JJ, Sturmans F, Drop R, Vleuten vd C, Hobus P. Does competence of general practitioners predict their performance? Comparison between examination setting and actual practice. *BMJ* 1991;303:1377-80.
15 Vleuten vd CPM, Swanson DB. Assessment of clinical skills with standardised patients: state of the art. *Teaching and Learning in Medicine* 1990;2:58-76.
16 Rethans JJ, Drop R, Sturmans F, Vleuten vd C. A method for introducing standardised (simulated) patients into general practice consultations. *Br J Gen Pract* 1991;41:94-6.
17 Rutten FM, Bohnen AM, Hufman P, Bruinsma M, Leerink HJG, Strootman FA, *et al*. NHG standard Angina Pectoris. *Huisurts Wet* 1994;37:398-406.
18 McClure CI, Gall EP, Meredith KE, Gooden MA, Boyer JT. Assessing clinical judgement with standardised patient. *J Fam Pract* 1985;20:457-64.
19 Rethans JJ, Boven van CPA. Simulated patients in general practice: a different look at the consultation. *BMJ* 1987;294:809-12.
20 Saebu L, Rethans JJ, Johannessen T, Westin S. Standardiserte pasienter I allmennpraksis. En ny metode for kvalitetssikring I Norge. *Tidsskr Nor Laegefor* 1995;115:3117-9.
21 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;i:307-10.

(Accepted 17 February 1997)

Survey of general practitioners' opinions on treatment of opiate users

Ann Davies, Peter Huxley

The government's *Guidelines of Good Clinical Practice in the Treatment of Drug Misuse*[1] gave detailed guidance on the treatment of drug users and signalled a shift in official policy away from drug treatment clinics and towards treatment in primary care for opiate users.

Even though general practitioners are now seeing more drug users,[2] there is little research on general practitioners' attitudes and practice in respect of opiate treatment in primary care. The last substantial report was by Glanz in 1985.[3] We therefore conducted a survey of general practitioners' attitudes to opiate treatment.

Subjects, methods, and results

We sent a postal questionnaire to all 341 general practitioners in three districts in Greater Manchester. One district was inner city and the other two were metropolitan boroughs. The questionnaire focused on general practitioners' opinions on and knowledge of opiate use and their prescribing behaviour. In all, 270 general practitioners returned their questionnaires, giving a response rate of 79%. The response rate was 75% (58/77) from district A, 83% (95/115) from district B, and 79% (117/149) from district C. Of those general practitioners who responded, only 6% of general practitioners had not seen any opiate users in the past year; 16% had seen one user, 24% had seen two to three users, 31% had seen 4-10 users, 14% had seen 11-20 users, 5% had seen 22-50 users, 3% had seen 60-100 users, and 1% had seen 104-250 users.

Table 1 shows the responses to statements about treating opiate misuse. Younger general practitioners and general practitioners in contact with support services had more positive attitudes to opiate users ($t = -3.34$, $P < 0.05$, df=239 and $\chi^2 = 8.56$, $P < 0.05$, df=1 respectively) Overall general practitioners were twice as likely to hold positive attitudes (64%, 172) as negative attitudes (30%, 82).

Comment

It is encouraging that twice as many general practitioners hold positive attitudes as negative attitudes when dealing with opiate users. These general practitioners are making use of support services offered to them by specialist agencies and they are generally pleased with these services. However, many general practitioners felt they would become more involved in treatment if more specialist services existed. We cannot assess whether the relation between positive attitudes and greater contact with support services is causal from our data, but it is encouraging that the relation is positive. A positive attitude may be related to a better (short term) treatment outcome; the second part of the study will explore this possibility.

Only two thirds of general practitioners were familiar with the government's guidelines on drug misuse, and only two fifths had actually read them. This finding is similar to that of Bell *et al* five years ago.[5] It is disappointing that although general practitioners are now seeing more opiate users than previously, few have read the guidelines. Most general practitioners said that they needed more training in dealing with opiate users and thought that they lacked the necessary knowledge and skills to deal effectively with users. The guidelines can be effective only if general practitioners are provided with sufficient training to be confident about their ability to treat opiate users.

School of Psychiatry and Behavioural Sciences, University of Manchester, Manchester M13 9PL

Ann Davies, *researcher*

Peter Huxley, *professor of psychiatric social work*

Correspondence to: Ann Davies.

BMJ 1997;314:1173–4

Table 1 Opinions and treatment behaviour of responding general practitioners

Attitude statements	No (%) responding				General practitioners who agree are more likely to be*
	Agree	Disagree	Neutral	95% CI	
I prescribe for opiate users (n=264)	211 (80)	53 (20)		75% to 85%	Younger (P<0.05)†Those not in contact with support services more likely never to prescribe (P<0.01)
Local community drug teams would encourage treatment (n=250)	194 (78)	28 (11)	28 (11)	73% to 83%	
Community drug teams provide good service (n=263)	177 (67)	39 (15)	47 (18)	61% to 73%	Those not in contact with support services (P<0.05)
General practitioners require more training (n=265)	167 (63)	34 (13)	64 (24)	57% to 69%	Those not in contact with support services (P<0.05)
More training would encourage treatment (n=249)	161 (65)	41 (16)	47 (19)	59% to 71%	
General practitioners feel they lack necessary knowledge to prescribe (n=264)	160 (61)	47 (18)	57 (21)	55% to 67%	
Shared care provides best services (n=264)	159 (60)	57 (22)	48 (18)	54% to 66%	Those in contact with support services (P<0.05)
Clinics in surgery run by specialist workers would encourage treatment (n=245)	134 (55)	64 (26)	47 (19)	49% to 61%	
Treatment is a specialist care service (n=261)	130 (50)	76 (29)	55 (21)	44% to 56%	Prescribing for fewer than 9 opiate users (P<0.05)and those not in contact with support services (P<0.05)
Primary aim is to help opiate users become drug free (n=262)	124 (47)	71 (27)	67 (26)	41% to 53%	Those in contact with support services (P<0.05)
Treatment of opiate users is beyond competence of general practitioners (n=266)	121 (45)	93 (35)	52 (20)	39% to 51%	Prescribing for fewer than 9 opiate users (P<0.01)
General practitioners should refer all on to specialist service (n=263)	114 (43)	98 (37)	51 (20)	37% to 49%	Those not in contact with support services (P<0.01)
Have read government guidelines (n=256)	110 (43)	146 (57)		37% to 49%	
Opiate users should be removed from practice list (n=264)	29 (11)	197 (75)	38 (14)	7% to 15%	Those not in contact with support services (P<0.05)

*All P values obtained by χ^2unless stated otherwise .† P value determined by analysis of variance.

We thank Len Bowers, Tom Carnwath, Mike Donmall, Mary Hopper, Mike Smith, Pat O'Dea and Hadi Mohamad, all the general practitioners who participated in this study, and the community drug teams for their help.

Funding: Regional Health Authority.

Conflict of interest: None.

1 Medical Working Group on Drug Dependence. *Guidelines of good clinical practice in the treatment of drug misuse.* London: Department of Health and Social Security, 1984.

2 Donmall MC, Millar T. Problem drug use in the North West 1990-1992. *Druglink* 1993;8 (July/August):8-10.

3 Glanz A. Findings of a national survey of the role of general practitioners in the treatment of opiate misuse: dealing with the opiate misuser. *BMJ* 1986;293:486-8.

4 Kirkwood BR. *Essentials of medical statistics.* Oxford: Blackwell Scientific, 1988.

5 Bell G, Cohen J, Cremona A. How willing are general practitioners to manage narcotic misuse? *Health Trends* 1990;2:56-7.

(Accepted 21 June 1996)

One hundred years ago

Poisonous ices and sweets

Whatever may be the composition of the stuff sold in the streets to children as ice cream and under the name of "hokey pokey" experience has shown that it is in many instances of a nature or in a condition that renders it extremely dangerous and not unfrequently fatal to the consumers. In the Islington Coroner's Court an inquest was held last week by Dr. Danford Thomas on a child who died two days after eating some of this noxious stuff, and the medical evidence attributed her death to the effects it produced. The jury in returning a verdict of death by misadventure added a rider suggesting that vendors of this commodity should be placed under strict inspection. This is certainly a case in which the services of public analysts should be utilised, though there may be some uncertainty as to whether the definition of the term "food" given in the Sale of Food and Drugs Act is sufficiently comprehensive, as it now stands, for that purpose. Prosecutions for the sale of sweetmeats containing paraffin wax have been successful in several instances, and a grocer was last week fined two guineas for selling them, but the material of which the sweetmeats were represented to consist being chocolate it came distinctly within the scope of the legal definition and more fully than the ice creams sold under the name of "hokey pokey" might do. This, however, is a point to which attention should be given whenever the Food and Drugs Act Amendment Bill comes before Parliament, and meanwhile it is desirable that the articles in question should be submitted to examination as to their capability of causing injury. (*BMJ* 1897;ii:236.)