Written Simulation of Patient–Doctor Encounters. 2. Assessment of the Performance of General Practitioners

H M JACOBS, M M KUYVENHOVEN, F W M M TOUW-OTTEN AND J C VAN ES

Jacobs H M (Department of General Practice, University of Utrecht, Mariahoek 6, 3511 LD Utrecht, The Netherlands), Kuyvenhoven M M, Touw-Otten F W M M and van Es J C. Written simulation of patient-doctor encounters. 2. Assessment of the performance of general practitioners. *Family Practice* 1984, 1:

Two rating procedures were used to judge the performance of 19 general practitioners confronted with a written simulation of patient-doctor encounters.[1] The simulation comprised five patients with vague complaints. A group of three 'expert' general practitioners judged the attention that the doctors paid to somatic aspects or to causes of the complaints. A second group of 18 experts judged the extent to which the therapeutic procedures of the general practitioners might induce a risk of unnecessary harm to the patient. Both rating procedures were shown to be reliable for three of the five simulated patients. A weak correlation between these issues was established for these three patients. The problems of judging the behaviour of the practitioners with the other two patients are discussed. The performance of the general practitioners was relatively constant. Variability between individuals substantially exceeded variability within an individual with respect to the attention given to somatic aspects, but these variabilities were roughly equal with respect to the risk of causing unnecessary harm.

A written simulation of patient-doctor encounters for registration and comparison of the performance of general practitioners has been described in a previous article.[1] It has long been held to be an important rule in medicine that a physician must avoid overlooking a relevant somatic aspect of a complaint: 'it seems fairly clear that the rule in medicine may be stated as: when in doubt continue to suspect illness.'[2] Over the past 10 years the profession as well as the public has begun to realize that improved diagnostic and therapeutic methods not only enhance health but may also have disadvantages and entail risks. Thus, while adequate examination and treatment of the somatic aspects of a complaint are important, the risk of inducing unnecessary harm to the patient should also be reduced to a minimum. Two rating procedures have been developed to evaluate these two aspects of the behaviour of general practitioners taking part in the written simulations.

Another problem which currently confronts general practitioners is 'to define what we mean by good care.'[3] In the absence of absolute rules, behaviour in the simulation was rated by experienced general practitioners drawn from the staff members of the University of Utrecht and active members of the Dutch College of General Practitioners,[4] who would be expected to be aware of and be capable of judging the two facets of the performance of general practitioners.

Because the criteria applied by the expert assessors remain implicit, it is important to establish the value which can be attached to the ratings. The two rating procedures were therefore submitted to reliability analysis to determine whether differences between the scores could be attributed to differences in the performance of the 19 general practitioners or to factors such as differences between experts, differences between patients or measurement error. The question of whether differences (if any) between experts were based on differences of opinion or on measure-

Department of General Practice, University of Utrecht, Mariahoek 6, 3511 LD Utrecht, The Netherlands

ment error was considered, and the extent to which these affected the scores was calculated. At the same time, consideration was given to how consistent was the behaviour of the general practitioners with the five patients.

METHOD

Nineteen general practitioners took part in the written simulation of patient-doctor encounters.[1] Two aspects of the performance were rated:

1. *Attention to somatic aspects or causes of the complaint.* This was rated independently by three university staff members who considered both the diagnostic and therapeutic procedures of the general practitioners.[1] The rating was made on a five-point scale ranging from 'entirely insufficient' (1 point) to 'very good' (5 points).
2. *Risk of unnecessary harm to the patient.* This was rated independently by 18 general practitioners. Seven of the assessors were university staff members and 11 were active members of the Dutch College of General Practitioners. This group of assessors was larger because the second factor is less well-defined in general practice, comprising as it does somatic, intrapsychic, relationship and social factors. The rating was limited to the therapeutic procedures because this part of the performance is most relevant in rating the risk of unnecessary harm and because of its high content validity in this simulation. The experts gave their opinion on a five-point scale ranging from: 'In view of the patient's complaint and the predictable positive effects of this strategy, I expect no risk of unnecessary harm to the patient' (1 point) to 'In view of . . . I expect a very grave risk of unnecessary harm to the patient' (5 points).

The reliability of the rating procedures was analysed and expressed as the intra-class correlation coefficient, $\hat{r}$.[6] This coefficient describes the extent to which the differences in the average ratings of the experts can be attributed to factual differences in the performance of the 19 general practitioners being tested.

The ratings of all the experts were analysed in terms of the relative frequency of the different response categories. This was done to check the tendency of the experts to give certain ratings regardless of the performance of the different practitioners being tested. For each expert the number of times he rated the different response categories was counted and all the experts were ordered according to this variable. For those experts who had very high (or low) scores it was reasonable to assume that they had tended to give certain ratings regardless of the performances of the general practitioners.[7]

The seven simulated patients included five patients with vague complaints (patients A, B, C, D, E), an 'instruction' patient with sinusitis and a 'test' patient with acute appendicitis. Details about the simulated patients were presented in the previous paper.[1]

RESULTS

The Reliability of the Procedures

1. *Attention to somatic aspects or causes of the complaint.* The rating procedure for this had a high reliability coefficient ($\hat{r}_3 = 0.83$). This means that 83% of the differences in the rating of the attention paid to somatic aspects of the complaint for the five patients was based on factual differences in the behaviour of the general practitioners. With patient B, for example, 11 of the general practitioners referred the patient for X-ray to make sure that there were no distortions between the fourth and fifth vertebrae. There were differences in rating between individual experts but these could be ascribed to errors of measurement. Intrajudge variation was seen when the judges sometimes gave different ratings to similar performances. There was no significant overall interjudge variation.

An analysis of the rating procedure for individual patients revealed reliable ratings for patients A, B, C and E ($0.80 \leqslant \hat{r}_3 \leqslant 0.89$). For patient D, however, reliability was very low, not because of rating differences between the experts but because of a relatively higher influence of measurement errors owing to the small number of experts.

2. *Risk of unnecessary harm.* On the basis of the response set analysis[7] the ratings for risk of harm given by three of the experts were eliminated as they were felt to be incapable of weighing the advantages and disadvantages of the therapeutic procedure in an unequivocal way. The reliability of the rating procedure taken in its entirety was relatively high ($\hat{r}_{15} = 0.89$) for the remaining 15 experts. But there were differences of opinion about the amount of risk of unnecessary harm when the reliability with individual patients was considered. The experts probably disagreed

about the estimation of the severity of the complaints. The discrepancy between experts was particularly important for patient E; the intra-class correlation coefficient for this patient was only 0.59. The reliability ratings for the other patients were satisfactory. The average ratings of the risk of harm between the experts for patients A, B, C and D (91, 82, 80 and 77% respectively) were based on differences in the therapeutic procedures of the general practitioners. Reliability would have been very low if only a single expert had rated the performance ($0.09 \leqslant \hat{r}_1 \leqslant 0.39$).

Further descriptions will be limited to the performance of the general practitioners with patients A, B and C.

Differences in the Quality of Care

The performances of the 19 general practitioners differed substantially (Figure 1). A description of the performance of general practitioners 4 and 12 in relation to patient A may serve to illustrate which behaviour was qualified as good and which as less good.

General practitioner 4 tried to establish whether there were gynaecological, urological or intestinal problems. He asked the patient whether she was afraid of something serious, and what she thought was wrong with her (complaint perception). He also considered the possibility of a correlation between her complaint and her psychosocial functioning. In the second encounter he continued this approach. Concluding the encounters, he explained to the patient what was wrong with her, prescribed a spasmolytic and gave precise instructions.

General practitioner 12 focused his history-taking and examination on the digestive tract and urological problems. He explained to the patient that she was suffering from chronic irritation of the caecum and prescribed a spasmolytic drug. In the second encounter he placed different accents, asking some questions about the course of the complaint and paying much attention to her intrapsychic functioning. In the conclusion he stated that he considered her to be rather listless and gloomy, and prescribed a neuroleptic drug.

1. *Attention paid to somatic aspects.* The performance of general practitioner 4 was rated as sufficient in terms of the attention paid to somatic aspects, whereas that of practitioner 12 was rated as insufficient. Unlike general practitioner 4, practitioner 12 disregarded gynaecological problems as a possible cause of the complaint and he also placed a different emphasis in the two encounters. In the first he focused entirely on possible somatic causes of the complaint. In the second he switched to asking questions about the course of the complaint and then focused on psychological aspects of the patient.

2. *Risk of unnecessary harm.* This risk was rated as high in the procedure of general practitioner 12; the prescription of a neuroleptic drug for 10 days for a patient with vague complaints of listlessness was not thought to be indicated. The experts rated the risk of unnecessary harm to the patient from practitioner 4 as lower than that of practitioner 12. The fact that practitioner 4 considered referring the patient to an internist may have played a role in the expert's rating of some

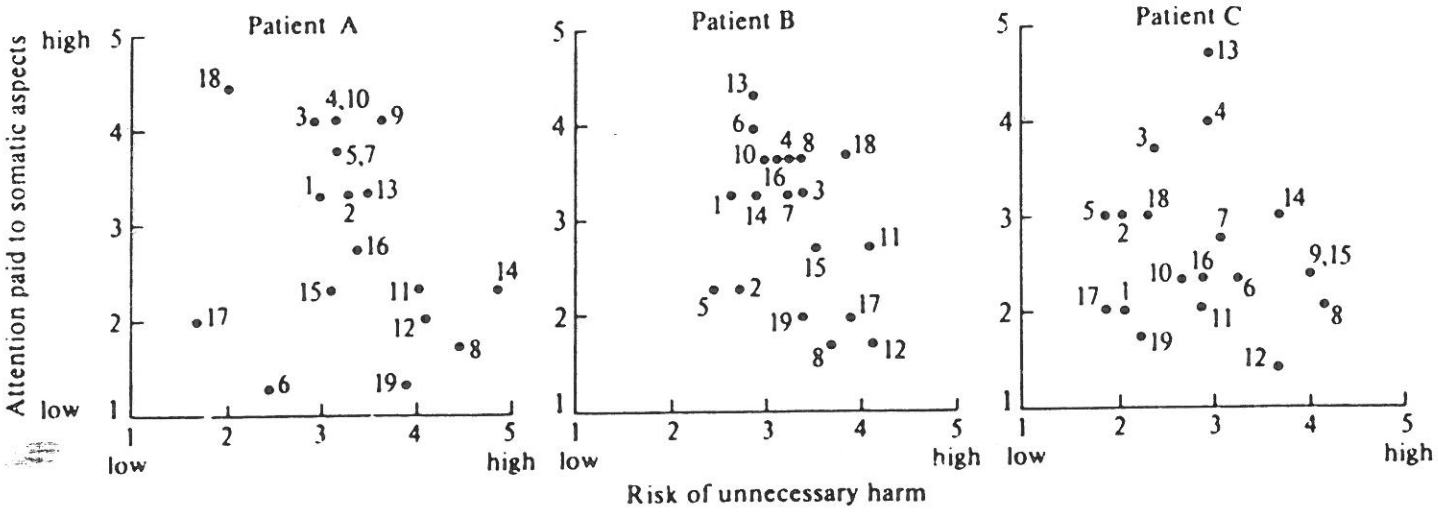

FIGURE 1 *Mean of experts' ratings of the 19 general practitioners (indicated by their reference numbers) for the attention paid to somatic aspects and the risk of causing unnecessary harm with patients A, B and C*
Kendall's rank correlation coefficients: patient A, $\tau = -0.26$; patient B, $\tau = -0.25$; patient C, $\tau = -0.04$

risk of unnecessary harm since they rated this risk higher in the procedures of doctors who referred patients to a specialist (see Figure 2).

The correlation between the two facets of performance is important for rating the overall quality of performance. Figure 1 shows that the therapeutic procedures of the general practitioners who paid sufficient attention to somatic aspects or causes of the complaint induced less risk of unnecessary harm than the procedures of the practitioners who did not pay sufficient attention to these aspects. The mean of the experts' ratings of risk of unnecessary harm for patients A, B and C are shown in Table 1, in relation to the mean ratings of the attention paid to somatic aspects of the complaint.

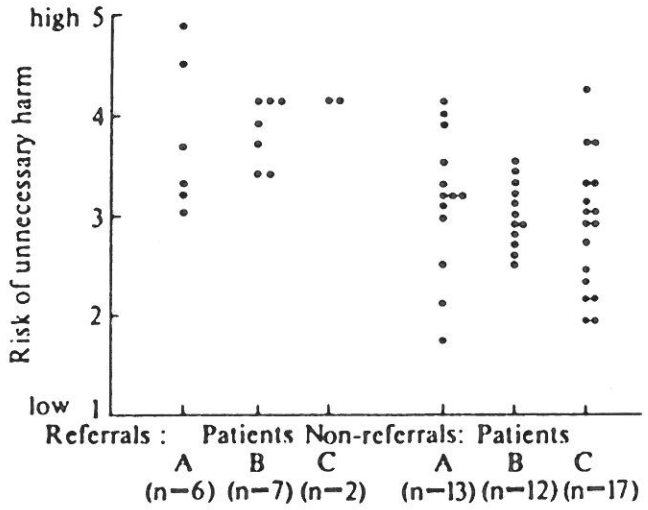

FIGURE 2 *The relation between the experts' ratings of the risk of unnecessary harm and referral or non-referral of patients A, B and C by the 19 general practitioners*
Mean overall rating for referral = 3.83, r_{sd} = 0.52
Mean overall rating for non-referral = 2.98, r_{sd} = 0.60

TABLE 1 *Mean ratings of the risk of unnecessary harm for patients A, B and C, subdivided according to the degree to which the general practitioners paid attention to somatic aspects of the illness*

Mean ratings of attention paid to somatic aspects[b]	Mean ratings of the risk of unnecessary harm[a]					
	Patient A		Patient B		Patient C	
	n	Mean	n	Mean	n	Mean
1.0-1.9	3	3.63	2	3.90	2	3.00
2.0-2.9	6	3.52	6	3.35	10	3.13
3.0-3.9	5	3.24	9	3.22	5	2.68
4.0-5.0	5	3.04	2	2.85	2	3.00

n = number of general practitioners.
[a] Low risk defined as a mean rating < 3.5, high risk > 3.5.
[b] Insufficient attention to somatic aspects defined as a mean rating < 3.0, sufficient attention > 3.0.

Stability of the Performance of the General Practitioners

1. *Attention paid to somatic aspects.* The general practitioners were relatively consistent in their attention to somatic aspects or causes of complaints; variance component analysis showed the differences between individual doctors (0.47) far exceeded the variability within each doctor (0.18).

2. *Risk of unnecessary harm.* This aspect of the performance of the practitioners showed a certain stability; the variability in the risk of inducing harm was roughly equal between doctors and within each doctor (0.18 and 0.17 respectively).

The finding of some intra-individual variability implies that the performance of the general practitioners was not entirely consistent. Intra-individual variability comprises two independent effects: a patient effect and a doctor effect. Only a small proportion of this variability was related to the patient problems and probably depended on differences between the simulated patients, such as differences in the complexity of the complaint or differences in the possible risks. The intra-individual variability in the performance of the general practitioners was largely specific to the doctor and was probably due to individuals' different estimates of the severity of the three problems of patients A, B and C.

DISCUSSION

The results show that it is possible, with the aid of expert general practitioners, to form reliable conclusions about qualitative aspects of the performance of general practitioners. There were problems, however, with the ratings of two of the patients (D and E). One problem concerned the number of experts required to rate any aspect of performance, as for example in the case of patient D. Pain in the elbow at the level of the lateral epicondyle is usually caused by hypertonicity of the arm muscles. The experience of general practitioners is that this is related not only to mechanical overload but also to a general increase in muscle tone resulting from increased emotional tension. The complaint evoked a fairly unequivocal diagnosis of 'tennis-elbow'; the diagnostic and therapeutic procedures of the 19 general practitioners showed a low level of differentiation. With this patient, however, three experts proved to be an insufficient number to

obtain reliable ratings of the attention paid to somatic aspects of the complaint.

The rating problems with patient E demonstrated that a low level of differentiation in practitioner performance is not the only factor which calls for a large number of experts. This patient showed a relatively complex set of symptoms. Although the mean of the ratings of 15 experts was used, the risk of unnecessary harm was unreliably judged since there was a discrepancy in the rating of the therapeutic procedures for this patient. An analysis of the experts' ratings revealed that they held different views on the necessary medical procedure in this case: whether the irregular vaginal blood loss was an indication for referral to an internist or a gynaecologist. Views on the risk of unnecessary harm are rather diffuse within the profession; a smaller number of experts would suffice to rate an aspect of performance where the standards are more explicit.

The selection of expert general practitioners is important in the rating procedures described. Three of the experts were found to be incapable of rating the performance of general practitioners in a reliable way. The selection could be improved by paying more attention to factors such as experience with rating procedures.

Both the attention paid to somatic aspects of the complaint and the induction of a risk of harm to the patient were relatively stable aspects of the performance of the 19 general practitioners. There was some intra-individual variability in the performance with different patients, but this was less than the inter-individual variability. This fact contrasts with the findings of the 'Patient Management Problems' simulation (PMPs),[8,9] in which intra-individual variability is usually higher. The relatively high stability of performance in this written simulation of patient-doctor encounters may be due to the fact that, unlike PMPs, this simulation comprises patients of only one category, those with vague complaints. In support of this is the finding that only a small proportion of the intra-individual variability in the performance of the general practitioners with the different patients was related to differences in patient problems. This variability was based largely on doctor-specific factors.

This written simulation of patient-doctor encounters provides a general impression of the therapeutic and diagnostic procedures which general practitioners use in response to patients with vague complaints.

ACKNOWLEDGEMENTS

We are indebted first of all to Ph. D. J. C. Maatsch, Director of the Office of Medical Education Research and Development, Michigan State University, for his important contribution to the development and construction of the two rating procedures. We thank Dr C. Spreeuwenberg, general practitioner and associate of the Department of General Practice, for his valuable and inspiring cooperation in the preparation of this publication. We also thank the university staff members and members of the Dutch College of General Practitioners who contributed to the construction and execution of the rating procedures, and in particular M. M. van Nunen, general practitioner in Hoensbroek-Noord. Finally we thank J. Dessens, sociologist at the Sociological Institute of the University of Utrecht, for his statistical advice on the variance component analysis.

REFERENCES

1 Kuyvenhoven M M, Jacobs H M, Touw-Otten F W M M, van Es J C. Written simulation of patient-doctor encounters. 1. Research instrument for registration of the performance of general practitioners. *Fam Practice* 1984; 1: (this issue).

2 Scheff T J. Decision rules, types of error and their consequences in medical diagnosis. *Behav Sci* 1963; 8: 97-107.

3 Hodgkin K. Evaluation in primary care. *Update* 1980; 20: 963-972.

4 Spreeuwenberg C. De toekomst van de huisartsgeneeskunde. *Medisch Contact* 1978; 33: 1205-1208.

5 Bridgstock M. Social theory and measures of the quality of medical care in general practice. *Soc Sci Med* 1979; 13A: 269-375.

6 Ebel R L. Estimation of the reliability of ratings. In: Ebel R L, ed. Essentials of educational measurement. Englewood Cliffs, NJ: Prentice Hall Inc., 1972: 116-131.

7 Galtung J. Theory and methods of social research. London: George Allen and Unwin, 1970.

8 McQuire C H, Page G. The assessment of clinical performance by written and oral simulations. Report to the Faculty 1972-73. Center for Educational Development, University of Illinois College of Medicine, Chicago, Illinois, 1973.

9 Elstein A S, Shulman L S, Sprafka S A. Medical problem solving: an analysis of clinical reasoning. Cambridge, Mass.: Harvard University Press, 1978.