General Practice Observed

Peer review of consultations in primary care: the use of audiovisual recordings

J E VERBY, P HOLDEN, R H DAVIS

British Medical Journal, 1979, **1**, 1686-1688

Summary and conclusions

A rating scale was designed to measure performance in interviewing techniques in primary care. Peer review of audiovisual recordings of their consultations showed that a group of experienced general practitioners achieved significantly higher scores on the rating scale compared with a group of similarly experienced general practitioners who did not observe their recordings. The higher scores were obtained at the expense of longer consultations.

The traditional five-minute appointment system in general practice needs to be reconsidered if an improved interviewing technique results in a more favourable outcome.

Introduction

The ability to communicate with a patient effectively has been described as the artistic aspect of clinical care.[1] It is important in defining a patient's problems and hence in arriving at a diagnosis. It is equally important when telling a patient what is wrong and enlisting co-operation in the management of their illness. Numerous studies have shown the relative inefficiency of the consulting techniques of doctors trained in the traditional model.[2-4]

Maguire *et al*[5] have shown that audio and more effectively audiovisual equipment can be used together with simulated patients to improve the interviewing techniques of students. Audiovisual equipment enables a doctor to observe his behaviour in the consulting room with patients.[6]

We have tried to show similar techniques may be used in primary care, when the consultation assumes a central role and when a holistic approach within severe constraints of time is necessary. We present the results of a study among a group of experienced principals in general practice. In a separate communication[7] we report our findings with a group of registrars in general practice (trainee assistant general practitioners).

General Practice Unit, Welsh National School of Medicine, Cardiff CF3 7PN, UK

J E VERBY,* MD, director, Rural Physician Research Programme, University of Minnesota Medical School

P HOLDEN, BA, DIP PSYCH, senior clinical psychologist, West Dorset Health Care District

R H DAVIS, DM, FRCGP, professor of general practice, Welsh National School of Medicine

*This study was undertaken while Professor Verby was on sabbatical leave and attached to the general practice unit.

Hypothesis

The study was designed to test the hypothesis that peer review of videotape recordings of doctor-patient consultations would modify the interviewing techniques and skills of experienced primary doctors.

Method

Seventeen experienced primary doctors each had two 30-minute consultation sessions recorded on videotape at intervals of between three and five months. All the doctors were members of practices engaged in vocational training for general practice. They were all aware that the study was designed to measure the quality of interviewing skills and techniques, but did not know precisely what criteria of measurement were to be used.

The videotape recordings were made in the doctor's own consulting room during an ordinary consulting session.[8] The consent of the patients to record the consultation and to use it for research and educational purposes was obtained. Five of the doctors were members of the general practice unit of the Welsh National School of Medicine who had had regular experience of peer review.[9] They formed the experimental group, who in the interval between the two recordings met weekly and discussed each others' recordings. We joined them at these meetings but were careful not to disclose the criteria used for measurement.

The remaining 12 doctors were members of teaching practices in South and Mid-Glamorgan. They constituted the control group and did not see their recordings or to our knowledge receive instruction or undertake any form of peer review exercise.

Measurement

To measure the interviewing techniques and skills that were displayed by the doctors, a general practice interview rating scale was designed. It consisted of 17 items, each of which was rated on a four point scale (table I). The ratings were converted into scores 0-3, a maximum total of 51 points being available.

The items in the rating scale were selected from two sources. Maguire *et al*[10] devised an interview rating scale for use in their studies on history-taking skills learnt by medical students. Where items in their scale were relevant to family practice they were included in the general practice interview rating scale. The other source, a medical interview skills checklist in regular use as a training aid on the Minnesota Rural Physician Associate Program,[11] provided further items for the rating scale. Four independent raters were trained to use the rating scale until acceptable levels of inter- and intra-rater reliability were achieved. Each rater was provided with a manual that laid down detailed criteria for assigning a rating (A-D) to each item. The manual was always used in conjunction with the rating scale to ensure maximum reliability between and within the raters.

Reliability data were computed on a sample of 16 videotapes made

TABLE I—*General practice interview rating scale*

	A : B : C : D	
1 Beginning of interview poor	—:—:—:—	Beginning positive and smooth
2 Seating arrangement open	—:—:—:—	Seating arrangement closed
3 Body posture bad	—:—:—:—	Body posture good
4 Appropriate use of eye contact	—:—:—:—	No eye contact or excessive use of eye contact
5 Frequent use of jargon	—:—:—:—	Absence of jargon
6 Frequently interrupted patient	—:—:—:—	Did not interrupt patient
7 Did not use facilitation	—:—:—:—	Frequent use of facilitation
8 Encouraged patient to keep to relevant matter	—:—:—:—	Not able to keep patient to relevant matter
9 Good clarification	—:—:—:—	Lack of clarification
10 Did not cover psychosocial areas	—:—:—:—	Covered psychosocial areas
11 Avoided personal issues	—:—:—:—	Did not avoid personal issues
12 No empathic statements	—:—:—:—	Frequent use of empathic statements
13 Picked up leads	—:—:—:—	Failed to pick up leads
14 Time spent in silence inappropriate	—:—:—:—	Appropriate use of silence
15 Good question style	—:—:—:—	Inappropriate question style
16 Warm	—:—:—:—	Cold
17 Ending of interview smooth and definite	—:—:—:—	Ending of interview abrupt or imprecise

during the study, each of which was seen and rated by two raters independently of each other. The same videotapes were also rated about two months apart by the same rater. The Pearson Product Moment Correlation Coefficient for inter-rater reliability on total scores was +0·87 and that for intra-rater reliability was +0·80. Each item was rated with a high degree of reliability, although correlation coefficients were not computed on this data because only four points were available for comparison, making such coefficients of dubious value. With the exception of two items, however, all were given the exact same rating or one point different on at least 92% of occasions during these reliability checks.

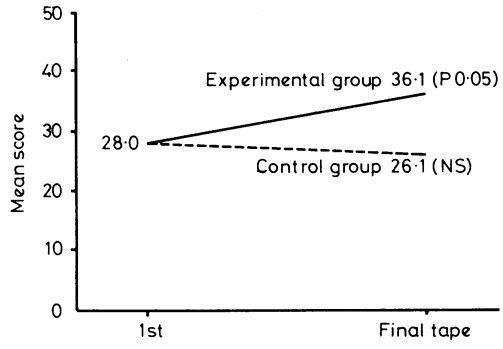
The four raters were given videotapes to rate in a balanced design such that no rater saw a videotape of the same doctor more than once except where she was undergoing an intra-rater reliability check. They were allowed to spend as long as they liked with each tape, going over the same ground several times until they were satisfied that their ratings accorded with the criteria laid down in the manual. Generally a rater took about one hour to rate a videotape lasting half an hour. Every consultation was rated separately, and the raters were kept blind about the group to which each doctor was allocated.

Results

Altogether 146 complete consultations were recorded and rated. The number of separate consultations recorded on each audiovisual tape varied between two and five. The mean score obtained on the first two consultations on any tape was similar to the mean score obtained on the first five consultations on each tape (SE 0·032). Each doctor completed at least two consultations on each tape. The remainder of the analyses were performed using the mean scores of the first two consultations on each tape.

The mean of the total scores obtained by both the experimental and the control groups on the first recording was the same (28·0). On the second recordings the experimental group achieved a much higher score (36·1) than the control group (26·1). The amount by which the experimental group increased its mean score was statistically significant ($P < 0·05$). The control group's mean score on its second recording did not differ significantly from its mean score on the first recording (figure).

An analysis of the scores obtained for each of the variables showed that on the first recordings there were no significant differences between the two groups on any of the 17 variables. On the second recordings, however, the experimental group of doctors achieved higher scores than the control group on all the variables. For six items the difference was statistically significant (table II). These were items 7 (use of facilitation), 9 (clarification), 13 (ability to pick up verbal and non-verbal leads), 15 (question style), 16 (warmth), and 17 (ending the interview). Differences between groups on total scores or on scores for individual items were all computed by *t* test.



Mean scores of experimental and control groups showing movements from first to final videotapes.

TABLE II—*Mean scores achieved for each variable by experimental and control groups on each recording. Difference between groups is expressed as plus when value of experimental group was greater*

Variable	First recording			Second recording			P
	Experimental	Control	Difference	Experimental	Control	Difference	
1	1·4	1·21	+0·19	2·1	1·54	+0·57	NS
2	1·6	1·21	+0·39	2·0	1·5	+0·5	NS
3	1·8	1·71	+0·09	2·5	2·0	+0·5	NS
4	1·2	1·5	−0·30	2·2	1·75	+0·45	NS
5	2·3	2·63	−0·33	2·7	2·58	+0·12	NS
6	2·1	1·46	+0·55	2·3	2·08	+0·22	NS
7	1·9	1·54	+0·36	2·0	1·17	+0·83	<0·02
8	2·0	2·33	−0·33	2·7	2·29	+0·41	NS
9	2·2	2·42	−0·22	2·6	1·79	+0·81	<0·01
10	1·6	1·71	−0·11	1·5	0·96	+0·54	NS
11	1·3	1·04	+0·26	0·9	0·42	+0·48	NS
12	0·6	0·46	+0·14	0·5	0·17	+0·33	NS
13	1·6	1·67	−0·07	2·5	1·29	+1·21	<0·001
14	1·8	2·00	−0·20	2·1	1·63	+0·47	NS
15	1·6	1·88	−0·28	2·4	1·75	+0·65	<0·02
16	1·7	2·00	−0·30	2·7	1·96	+0·74	<0·05
17	1·3	1·29	+0·01	2·4	1·21	+1·19	<0·02
Total	1·65	1·65	+0·00	2·12	1·53	+0·59	

TABLE III—*Number of interviews, mean time of consultations, and range on first and second recordings of control and experimental group of doctors*

	First recording			Second recording		
	No of interviews	Mean time per interview (min)	Range	No of interviews	Mean time per interview (min)	Range
Control group n = 12	44	6·8	1·7-15·6	47	5·0	1·1-13·7
Experimental group n = 5	12	7·5	2·3-12·4	15	10·5	6·6-17·0
Total n = 17	56	6·9	1·7-15·6	62	7·8	1·1-17·0

The length of experience of general practice of both groups varied from four to 31 years. There was no correlation between the scores obtained on the first tapes and the length of experience in general practice. An analysis of both tapes of the control group and of the first tapes of the experimental group—that is, before any training—showed a positive correlation between the score achieved on an individual consultation and the time taken for that consultation ($Y = +0·3$). This correlation was significant ($< 0·01$ $P > 0·001$ $N = 95$). The average time taken for a consultation on the first recordings was similar for both the control (6·8 min) and the experimental group (7·5 min). On the second recordings, the control group took only an average of five minutes for each consultation whereas the experimental group took on average 10·5 minutes (table III).

Discussion

We have developed a rating scale that can be reliably used by independent lay people to measure certain components of the

consultation in primary care. We did not use all the theoretical variables[12] because some were clearly not appropriate and in others we could not achieve a satisfactory rater reliability. All the variables we included could discriminate between the behaviour of doctors during the consultation.

Our results show that a group of experienced general practitioners can achieve significantly higher scores on the rating scale after they have been able to observe their consultation techniques and after they have discussed the recordings in a peer group review. All the doctors in the experimental group were members of one group practice who had had experience of peer review. They were thus accustomed to accepting the critical comments of their colleagues and to this extent could cope better with the threatening experience of having the traditionally confidential primary care consultation exposed. They were consequently able to behave constructively, learn new skills, and expand existing ones because they were working in a support system.

It does not necessarily follow that the patients of these doctors subsequently fared any better than the patients of the control group of doctors, because even though there is evidence that improving the interviewing skills of medical students increases the quantity of data obtained from a patient,[13] these data still need to be analysed, synthetised, and translated into a plan of management. It seems reasonable, however, to assume that patients will benefit if the communication between them and their doctor is improved. Analysis of consultations recorded on audiotape has shown that some consultations were less effective than others because doctors used inappropriate interviewing techniques.[14] There is certainly a wealth of evidence to suggest that lack of understanding lies behind many of the complaints made by the public about their doctors.[15]

In very many consultations general practitioners appear to develop a pattern of consulting behaviour that is singularly consistent.[14] In our study fewer interviews were recorded but there was no indication that the clinical material of the consultations differed substantially between the control and experimental groups or between individual doctors' first and second recordings. There was no correlation between the length of experience of primary care of the doctor and the score achieved on the rating scale. This suggests that the behaviour of the doctor is established within a few years of entering practice and in the absence of any opportunity to examine critically his performance persists unaltered.

Recording the primary care consultation with its traditional concepts of confidentiality could be regarded as introducing an element of unreality that might bias the results of any study designed to show improvements in the techniques of doctors. Other observers[14] using audiotape found little evidence that the process of the consultation was inhibited. Our experience with audiovisual recordings showed that patients rapidly forgot that the consultation was being recorded. Doctors on the other hand did appear to remain conscious of the camera and when questioned afterwards said so. The continuing awareness of the doctor might be expected to result in a longer consultation than usual. The mean time of each consultation of both groups before any feedback was 6·4 minutes (range 1·1-15·6). These figures are comparable with those recorded in studies of consultation times in general practice in Britain.[16-19]

The scores achieved by both groups of doctors before any feedback show a positive correlation between the length of the consultation and the score obtained. Theoretically a short consultation could achieve a lower score because the doctor does not have time to demonstrate several variables—for example, silence. The constraints on time may clearly have been different for different doctors but so far as we know no change occurred in the organisation of the practices or appointments systems between the first and second recordings. The higher scores achieved by the experimental group of doctors after feedback were accompanied by a significant increase in the length of their consultations. This observation needs to be repeated because if it is true and positively related to outcome then it implies that "the ubiquitous five-minute appointment system"[16] needs to be questioned seriously.

It would have been impressive if we could have shown a correlation between the score achieved on the rating scale and some measure of outcome of the consultation. Unfortunately a questionnaire designed to measure the patients' satisfaction was unreliably answered. Nevertheless, there is evidence that outcome is affected by consulting techniques—for example, consultations that go wrong tend to be shorter than those that are satisfactory[14] and the outcome in psychotherapy has been shown to be dependent on the therapist displaying empathy, warmth, and genuineness.[20]

We thank Drs L Addicot, E R Collins, D E Davies, D Evans, S Gang, R H G Lloyd, S McHugh, D Parry, A S Parsons, G Phillips, B Price, T Reilly, R M Richards, S A Smail, N C H Stott, H A Thomas, and B B Wallace for allowing us to record their consultations; Mrs S Ayres, Mrs M Jones, Mrs D Ipsen, and Mrs P White for rating the audiovisual tapes; and Dr R R West for analysing the data.

This study was supported by a grant from the Welsh Scheme for the Development of Health and Social Research.

Requests for reprints and further details of the general practice interview rating scale should be addressed to Dr R Harvard Davis, General Practice Unit, Welsh National School of Medicine, Llanedeyrn, Cardiff.

References

1 Fienstein, A R, *Clinical Judgement*. Baltimore, Williams and Wilkins, 1967.
2 Maguire, G P, and Rutter, D R, *Lancet*, 1976, **2**, 556.
3 Anderson, J, *et al*, *Postgraduate Medical Journal*, 1970, **46**, 606.
4 Tapia, F, *British Journal of Medical Education*, 1972, **6**, 133.
5 Maguire, G P, *et al*, *Psychological Medicine*, 1978, **8**, 695.
6 Verby, J E, *Journal of American Medical Association*, 1976, **236**, 2413.
7 Verby, J E, Davis, R H, and Holden, P. In press.
8 Verby, J E, Davis, R H, and Marshall, R, *Journal of Audiovisual Media in Medicine*, 1979, **2**, 56.
9 Stott, N C H, and Davis, R H, *Journal of the Royal College of General Practitioners*, 1975, **25**, 888.
10 Maguire, G P, Clarke, D, and Jolley, B, *Medical Education*, 1977, **11**, 175.
11 Verby, J E, *Minnesota Rural Physician Associate Program*, University of Minnesota, Department of Family Practice, 1976.
12 Flanders, M A, *Helping Teachers Change their Behaviour*. Michigan, University of Michigan Press, 1963.
13 Rutter, D R, and Maguire, G P, *Lancet*, 1976, **2**, 558.
14 Byrne, P S, and Long, B E L, *Doctors Talking to Patients*. London, HMSO, 1976.
15 Cartwright, A, *Patients and Their Doctors*. London, Routledge, Kegan, Paul, 1967.
16 Buchan, I C, and Richardson, I M, Scottish Health Service Studies No 27. Edinburgh, Scottish Home and Health Department, 1973.
17 Royal College of General Practitioners, *Present State and Future Needs of General Practice*, 3rd edn. London, RCGP, 1973.
18 Floyd, C B, and Livesey, A, *Journal of the Royal College of General Practitioners*, 1975, **25**, 425.
19 Westcott, R, *Journal of the Royal College of General Practitioners*, 1977, **27**, 552.
20 Truax, C B, and Mitchell, K M, In *Handbook of Psychotherapy and Behaviour Change*, ed A E Bergin and S L Garfield. New York, John Wiley, 1971.

WORDS CALLIPER, caliper. In orthopaedics a device applied to a limb comprising two rods joined together proximally and each turned inward at the distal end to enter an attachment, such as a socket in the heel of the shoe. Also an obsolete instrument for measuring the dimensions of the female pelvis (hence usually called a pelvimeter), similarly comprising two curved arms hinged at one end. The word stems from CALIBRE, originally the diameter of a bullet or cannon ball, and by extension, the internal diameter of a gun, and later of any cylindrical structure, such as an artery or bronchus. Whence calliper, an instrument for measuring the calibre of a gun or other hollow structure, or the external diameter of a convex body, and by analogy the orthopaedic device defined above. CALIBRATE was originally to measure the calibre of a thermometer tube, whence to graduate a gauge or scale of any kind with allowance for its irregularities.