Quarterly Journal of Medicine, New Series 76, No. 281, pp. 935–949, September 1990

Clinical Judgment Analysis

JOHN R. KIRWAN, D. MARK CHAPUT DE SAINTONGE* and C. R. B. JOYCE†

*From the Rheumatology Unit, Department of Medicine, University of Bristol, *Department of Clinical Pharmacology and Therapeutics, London Hospital Medical College, and † Ciba-Geigy Limited, Basel*

Accepted 29 January 1990

SUMMARY

Judgment is central to the practice of medicine and occurs between making clinical observations and taking clinical decisions. Clinical judgment analysis has developed as a method of making statistically firm models of doctors' judgments. Computed models reveal the differential importance attached to items of clinical, social, or other data which are determinants of clinical decisions. These models can both reveal the causes of conflicts of judgment and may help resolve them in a way that unaided discussion cannot. Revealing experts' models to students speeds learning of diagnostic skills. Clinical judgment analysis offers a method of probing the judgments not just of students and doctors but also of patients who have shown systematic differences in their perceptions of risk and benefit. The power and relevance of clinical trials can be improved by the consistent application of judgment policies generated from both the trialists and those who will use their results.

INTRODUCTION

Like all living organisms, doctors use information from their environment to help them decide on appropriate actions. In the medical world this usually means choosing treatments for a diagnosed illness. The process of making a diagnosis is complex and not fully understood, but basically involves the collection of items of clinical data, usually symptoms, signs, and laboratory results, and then using this information to make a judgment about the probability of the presence of specified diseases. The selection and measurement of clinical data is subject to many errors and biases which can prejudice the quality of subsequent decisions. The problems of observer error, the need to minimize it and the means of doing so are generally well-known. However, even if perfect observation were possible, good judgment is still required if accurate diagnoses are to be reached. It is surprising that there are so few systematic studies of clinical judgment, even though it is central to the practice of medicine. Systematic variations in medical practice are common and are being revealed more often, as medical audit is more widely practiced. It seems unlikely that these variations can be

Address correspondence to D. Mark Chaput de Saintonge, Department of Clinical Pharmacology and Therapeutics, The London Hospital Medical College, Turner Street, London E1 2AD, United Kingdom.

simply attributed to observational errors: some will be related to differences between doctors in the clinical judgment policies which they operate.

We first report the methods which have been developed to separate judgment from observation and to measure variations in judgment between doctors and then go on to summarize how models of medical judgment have been developed for specific medical problems, show how these models can be used to improve physicians' judgments and review the reactions of the physicians who took part in the experiments. We conclude by considering future implications for medical education and medical practice.

The study of clinical judgment can be approached in many ways, but this paper centres on recent applications of Social Judgment Theory [1] to the problem of combining observations as a basis for action (i.e. making judgments). Social judgment analysis takes account of the fact that we work in a probabilistic environment in which the evidence we gather bears an imperfect relationship to its cause. In medicine this is exemplified by the variety of diseases which might produce a given symptom, and the variety of symptoms which may arise from a particular disease. It seems likely that the process of medical judgment involves interpreting these superficial characteristics on a probabilistic basis which reflects these underlying uncertainties. Brunswik [2] outlined and Hammond *et al.* [3] developed the 'lens' model as a graphical and mathematically rigorous model of judgment (Fig. 1).

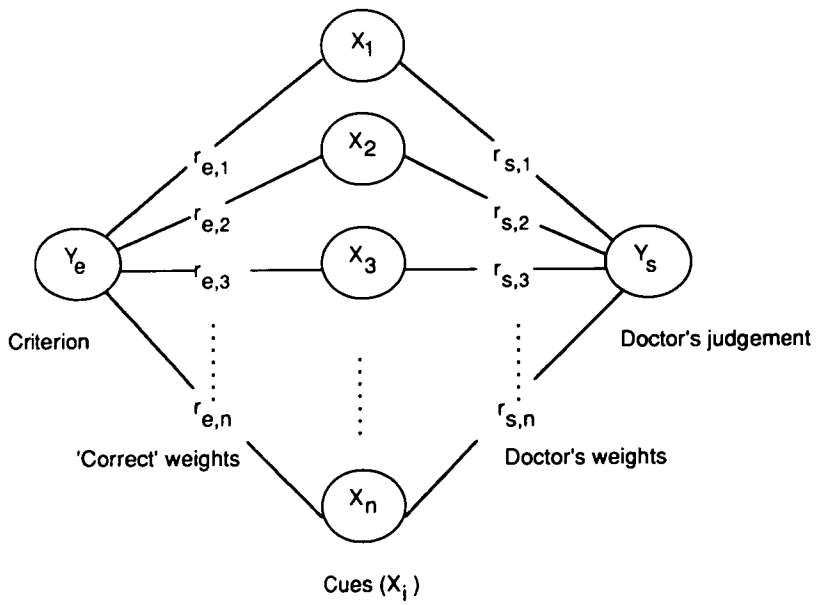

FIG. 1. Brunswik's lens model applied to disease assessment. The criterion to be judged might be the presence or the activity of a disease. This is manifest as any number of cues or indicants which traditionally take the form of symptoms, signs, or laboratory variables. The relationship between them and the criterion are usually indicated by the correlation coefficients ($r_{e,n}$). It is these cues that the doctor takes into account when making a judgment about the disease. His pattern of cue utilization is apparent from the correlations they have with his judgments ($r_{s,n}$). This lens model paradigm allows systematic differences between doctors judgments to be displayed in terms of the differences in the weights of importance attached to the various cues, and differences in the combination rule used to arrive at a final judgment. When there is some way to approach the criterion independently the doctor's weights can be compared with the 'correct' weights.

Brunswik argued that judgments are based on pieces of information bearing uncertain relationships to the underlying nature of the condition being assessed. These could best be described using multiple regression analysis to relate a series of judgments made by a particular observer (as the dependent variable) to the various data on which the judgments were made (independent variables). Such analyses should theoretically allow for non-linear and discontinuous relationships, but in practice linear regression has proved adequate for modelling the majority of judgment situations [4]. This approach has been increasingly and successfully applied to practical clinical problems and has allowed the construction of explicit models which often fit closely the judgments physicians actually make.

SEPARATING OBSERVATION FROM JUDGMENT

To some extent the processes of observation and judgment are always interwoven. For example, observing the murmur of aortic stenosis involves a judgment about how much of systole is occupied by the sound. To make proper comparisons between and within individual doctors the same clinical material would have to be presented to a number of observers on a single occasion, or a single observer on different occasions. With live patients this is difficult or impossible. Even ways of presenting records of the same material repeatedly, such as using videotape recordings, would not necessarily separate the functions of the 'observer' from those of the 'judge'. The material needs to be not so much pre-recorded as pre-scored and presented as laboratory test data, clinical observation, extracts from the patient's history etc. in a way that is identical for all judges. Such presentations may be made by a computer programmed to generate clinically plausible values or by using values taken from patient records and presented in a simple form. They are usually described as 'case vignettes', 'scenarios' or 'paper patients' and contain a scene setting element which is invariant across all cases and a variable element representing the clinical data under study [5, 6].

'Paper patients' (or their equivalent) must, however, simulate true clinical encounters sufficiently well for the judgments they induce to match those that the physician makes when seeing real patients with the same signs and symptoms. In order to establish whether this is the case three studies have been performed. The first compared judgments about disease severity made on patients in rheumatology clinics with those made by the same physicians on the same set of observations, presented as paper patients some weeks later [6]. Figure 2 shows a typical result, with a correlation coefficient between the two sets of judgments of $r = 0.9$. In the second study judgments about the severity of signs of otitis media were compared with those made when the real patient has been examined, and again a reasonable (though less good) agreement was obtained [5]. Test ordering behaviour has been studied in 98 family practitioners. The correlation between decisions predicted by the model and choices observed in practice was 0.74 [7]. Paper patients thus isolate judgments from the process of gathering information, allow comparisons between clinicians to be made on a standard basis, allow the testing of repeated judgments and reflect the judgments clinicians make when seeing real patients.

DIFFERENCES AND VARIATIONS IN JUDGMENT

Using paper patients it is a simple matter to demonstrate that doctors differ in their clinical judgments of the same material. When a random sample of 48 UK rheumatologists assessed the degree of change in disease severity in cases of rheumatoid arthritis the doctors' judgments showed major disagreements [8] (Fig. 3). Even when the clinicians were asked to

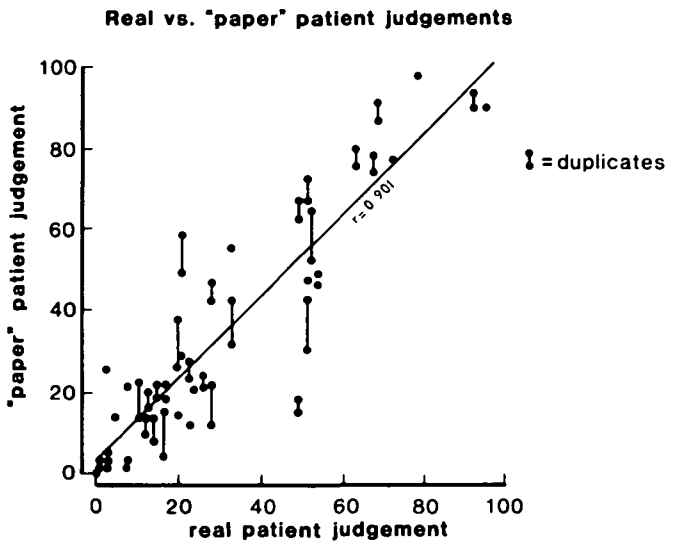

FIG. 2. Correlation between clinical scores (0–100) for real and equivalent 'paper' patients when judging 'current disease activity' in rheumatoid arthritis. (Reproduced by permission from ref. 6.)

indicate only 'clinically important' changes considerable disagreement remained. Such apparently major differences in judgment have also been found in rheumatologists in Australia [9] and Canada [10], amongst general physicians [11] psychiatrists [12] radiologists [13] and between general practitioners in the UK [5, 14].

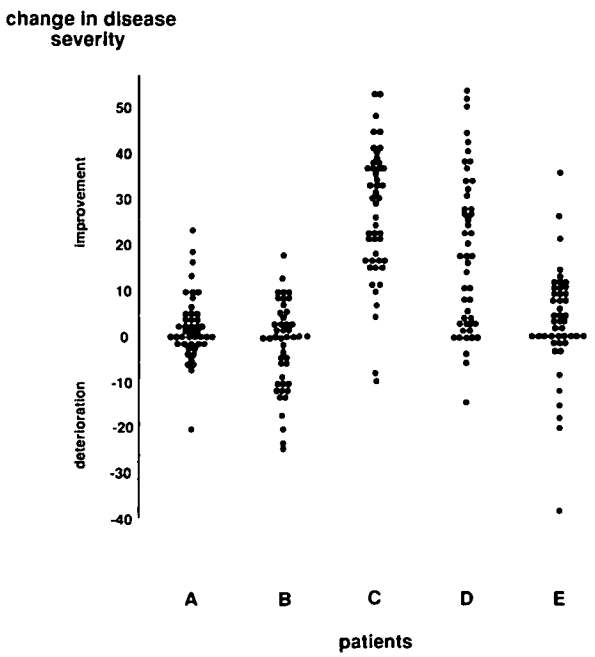

FIG. 3. Distribution of VAS scores from five patients. Variation among 48 physicians in their assessment of disease severity in five patients. Score +55 indicated maximum possible improvement and −55 maximum deterioration. (Reproduced by permission from ref. 8.)

These differences might result from inconsistency in the way some (or all) judges carry out their task. The reliability or reproducibility of physicians' judgments has been measured in studies that included duplicate sets of patient data [8, 10, 14]. In some tasks, the correlation between repeated judgments made by some physicians has not been significantly different from zero, but in the majority of cases physicians have proved reasonably ($r_s > 0.7$) or highly ($r_s > 0.85$) consistent. Individual clinicians often show a reasonably consistent judgment policy over many months (and perhaps years). In one study [15], seven physicians made judgments on identical sets of paper patients on two occasions one year apart. The stability of their judgments over one year ($r_s = 0.70$) compared favourably with the reliability of duplicate judgments on each occasion ($r_s = 0.76$). Similar stability was seen amongst a group of general practitioners. A trainee, however, progressively shifted her policy over a six month period so that it eventually approximated to that of her trainer [14]. This observation reminds us that some time-dependent shifts in judgment policy may be both expected and desirable. However, the fact remains that even highly consistent expert judges show marked differences in their judgments about identical data sets [8–10]. The reasons must lie in the way they make their judgments – the process by which each selects and combines the clinical data, their 'judgment policies'.

MAKING DIFFERENT JUDGMENT POLICIES EXPLICIT

A first step in obtaining explicit judgment policies might be to request each physician to describe his or her own approach to combining the data. Such descriptions could be in the form of weights allocated to a set of variables to indicate the contribution each makes to the overall judgment, or in the form of detailed and carefully considered descriptions of the use of each variable. Physicians are easily able to adopt either of these methods of describing their judgment policies as they perceive them to be, and large differences between their policies emerge with either method [14, 16].

Unfortunately, when the weights from these perceived policies are applied to observations on patients, or to the ratings of paper patients, they result in judgments which show relatively little correspondence with those made by the physicians when they actually see their patients or judge the paper patients in question [16]. In fact, these 'specified' weights may prove no better than giving equal weight to all the data, and explain only about 40 per cent of the variance in judgments. Thus, the descriptions provided by expert physicians of their judgment policies offer little real insight into the cause of those differences.

Regression modelling, based on the lens model paradigm, provides a practical approach to describing how doctors use predefined data to make judgments. The regression equations are used as models of the judgment process, though in fact they are only one way (among many) of combining the data to arrive at the same kind of output as the judge [17]. The models are descriptive in a mathematical rather than a psychological sense, though it is possible that they may also represent correct psychological models. In either case, they may provide the judges with novel insights.

The most appropriate way of expressing the contribution each clinical variable, or cue, makes to the model is open to debate [18]. The arguments seem to favour the relative contribution to R^2 (the square of the multiple correlation coefficient of the regression equation) [18, 19]. Here the change in R^2 which occurs when a cue is omitted from the equation is compared to that when each of the other cues are omitted in turn. Standardized regression coefficients may be equally good where intercorrelations between the predictor variables are close to zero. The risk of 'overfitting' and capitalizing on chance relationships

within the data which is inherent in all multiple regression approaches can be reduced by imposing a 'penalty' against equations containing many variables.

Regression models calculated in this way are frequently able to explain a high proportion of judgment variance: for example, the pooled value taken from one study of 89 rheumatologists was 73 per cent [16]. When models calculated from judgments made on an initial set of data were applied to a new set of patients they explained 88 per cent of the variance; however, rheumatologists' specified judgment policies, even described after all their judgments had been made, could explain only 34 per cent. It seems, therefore, that the policy equations provide an adequate and consistent model of clinical judgment with greater validity within the area of judgment tested than physicians' own perceived policies, and can therefore be used to compare the judgment policies of different clinicians with reasonable confidence. Further details of the use of linear models to analyse physicians' decisions have been recently reviewed [20].

COMPARING POLICY MODELS

Two rheumatologists who worked together in the same department and shared the care of the same group of patients took part in a study comparing clinical judgment of disease severity [21]. Their models showed little agreement in the relative importance of the clinical variables (Fig. 4). Furthermore, the specified policies described by the clinicians bore little resemblance to those calculated from their actual judgments. Indeed, the doctors held quite similar beliefs about their clinical behaviour and without deeper information it would be difficult to discover why they differed systematically in their assessments of certain patients. This information was provided by regression models developed from their judgments on 50 paper patients. These revealed that their actual policies were quite different – the patient's global assessment upon which rheumatologist B placed so much weight was totally ignored by rheumatologist A.

OF WHAT USE IS POLICY MODELLING?

Given that differences in expert clinical judgment can be appreciated and measured, to what further use can this information be put?

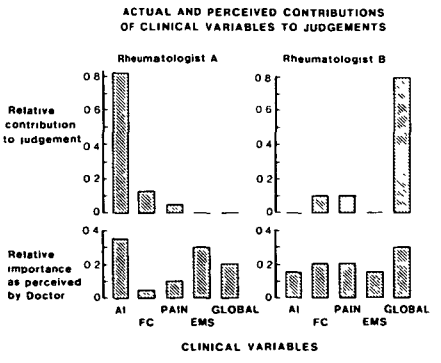

FIG. 4. Actual (modelled) and perceived contributions of clinical variables to two rheumatologists' judgments of disease severity. The clinical variables were: articular index (AI); functional capacity (FC); pain; early morning stiffness (EMS); patients global assessment (GLOB). (Reproduced by permission from ref. 21.)

The first benefit is in the individual's greater appreciation of the consistency of his own judgment processes. In addition, an explicit model enables him to examine the importance he attaches in practice to clinical data, to modify the model consciously and to apply the revised form with consistency. But further than that, analysis of clinical judgment provides a tool for helping to coordinate the judgment of several physicians. Diagnostic models based on actual judgments made by general practitioners will, for example, allow predictions to be made as to which cases of otitis media will cause disagreement in future (Fig. 5). This example illustrates the false agreement which might result if discussion rather than policy modelling were used. It will allow these doctors to consider how their management of identifiable patients might differ as a consequence of their different policies. It may be that the consequences will be clinically trivial; on the other hand they may feel it important to reach a consensus policy on diagnosis to avoid undesirable variations in treatment.

Two further investigations illustrate these advantages. In the first, two clinicians wished to coordinate their policies for the inclusion of patients in an international study of rheumatoid arthritis. They had agreed upon the entry criteria but sought further advice about even better coordination. Each was asked to judge the suitability for entry of 90 paper patients; the correlation between their judgments, a measure of their agreement, was $r=0.63$. The two clinicians then spent one hour discussing in detail the decisions each had made. Following this feedback on the outcome of their judgments, each separately assessed 30 further patients. The correlation between the second set of judgments was $r=0.64$, showing little change. Their clinical decisions were analysed separately and each was supplied with graphical representations of the clinical importance attached to variables in both of their judgment models. They met and discussed these models for a further hour before once more assessing 60 paper patients. The agreement between them after this 'process' feedback improved ($r=0.76$) [22]. This confirms earlier suggestions [23] that feedback provided by judgment analysis can improve agreement when unaided discussion or 'outcome' feedback fails.

The second example relates to routine clinical practice. Three physicians attempted to coordinate their judgment of disease severity when reviewing outpatients with rheumatoid

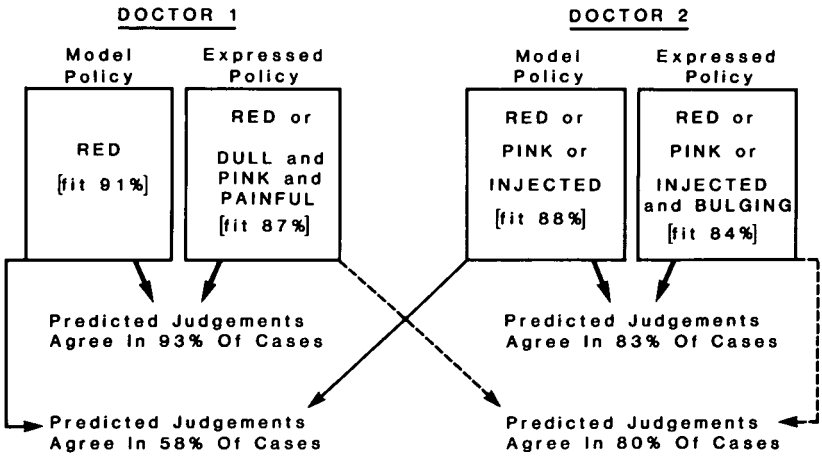

FIG. 5. Modelled and expressed (perceived) policies for two doctors using appearances of the tympanic membrane to diagnose otitis media. Application of their expressed policies would lead to a predicted agreement over 80 per cent of the cases. In reality they would agree over only 58 per cent as a comparison of their modelled policies shows. (Reproduced by permission from ref. 14.)

arthritis. At the end of a three month period, during which judgments and clinical data were collected, judgment policy models were calculated for each doctor and displayed in the clinic rooms. These models were applied to a fresh set of case data and used to predict the judgment that each doctor would make if he were perfectly consistent in the application of his policy. The correlation between the predicted judgments of one pair of doctors was particularly low indicating considerable differences in their judgment policies (Fig. 6). When assessing patients over the next three months they were able to see all three policy models, as well as predicted assessments for each consecutive incoming patient based on computer calculations performed and displayed 'real time' in the clinic. Each doctor was thus able to see what decisions his colleagues would have made had they been seeing the patient. At the end of this period their policy models were calculated once more. By applying both sets of policy models to data collected from further patients it was possible to examine changes in agreement over the three month period of feedback. Figure 6 shows that the correlation between two of the participants improved dramatically from $r=0.54$ to $r=0.99$ [24]. It appeared that real-time display of other physicians' assessments of each patient encouraged convergence of the judgment policies.

Another area where judgment analysis may help is in the design and conduct of clinical trials. Patients included in a trial must fairly represent the population of patients seen by those who read the report. In a study of published trials in bacterial otitis media, trial diagnostic criteria for admission were compared with those derived from the judgment analysis models of 27 general practitioners. Half of these doctors disagreed with the diagnosis in the majority of patients who would have been considered suitable for entry into the trials. It is difficult to know to what use these doctors could have put the results. At the moment diagnostic criteria for entry into trials are chosen solely by the trialists themselves. It seems likely that the relevance and value of clinical trials could be improved by modelling the diagnostic criteria if the doctors who will be using the results of the trial and making sure thay match the trial admission criteria. [25, 26]. Such methods would allow the construction of operational models of disease not just as they are seen by doctors but by patients, relatives and health care planners as well. Differences between such models may give insights into how cooperative health care may be planned better.

Trials lacking the power to detect clinically important effects have been frequently criticised. Although it has been tacitly assumed that everyone knew what these important effects were, recent studies (for example in the medical management of hypertension) have suggested that what doctors may consider a success, patients' relatives may rate as a failure [27].

But do doctors even agree upon the definition of treatment success? To find out, the judgment policies of 56 rheumatologists were modelled using 50 paper patients. Each 'patient' provided two sets of measurements on ten clinical variables recorded before and after one year's treatment. The rheumatologists were asked to estimate the size of any change in disease severity that had taken place and whether the change was clinically significant or not. Although all the rheumatologists were of consultant or senior registrar status they showed little agreement over which patients had shown important improvement ($\kappa=0.3$). Some patients were considered significantly improved by certain physicians and significantly worse by others! Inspection of their calculated judgment policy models showed that while some variables entered most doctors' models, others were unevenly represented and were a cause of major variation in judgement [28,30]. If every member of a collaborative research group is shown his or her policy in this way, it becomes possible to hold a truly informed discussion of differences, so that a uniform policy can be agreed. The value of judgment analysis in increasing the power and relevance of clinical trials is discussed in detail elsewhere [29, 30].

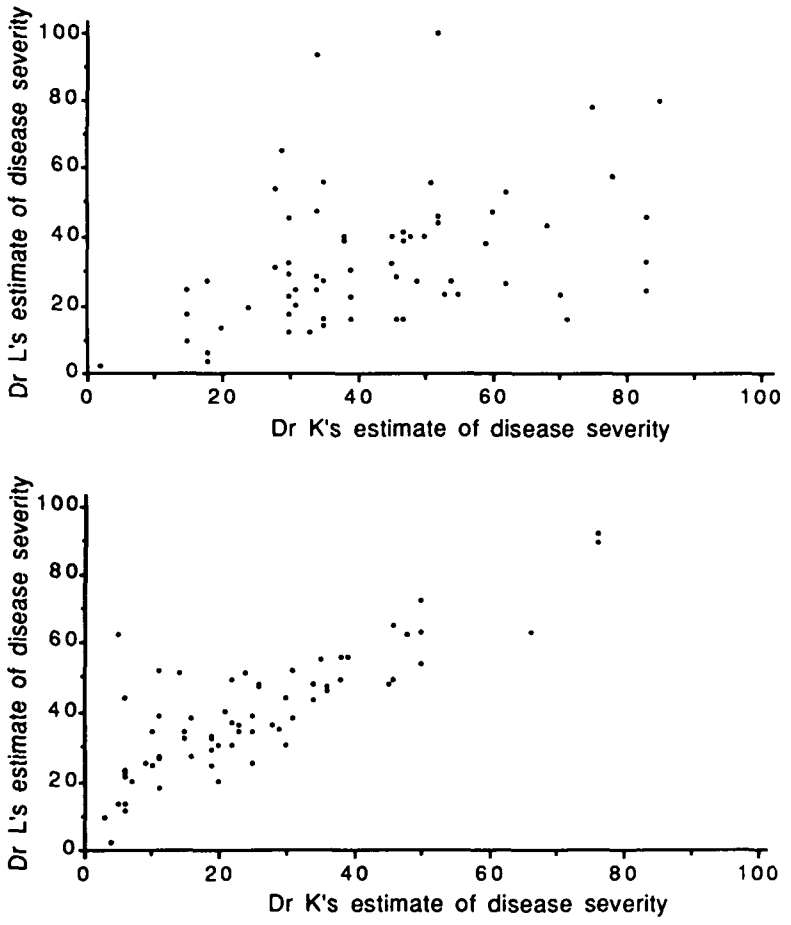

FIG. 6. Agreement between two rheumatologists in routine patient assessment. Predicted disease severity scores obtained by applying the policy models of two rheumatologists to new patient data (a) before feedback, (b) after feedback from clinical judgment analysis. (Reproduced by permission from ref. 24.)

Whether such judgment modelling would in fact aid a consensus about trial end points is as yet untested. However, the retrospective application of judgment analysis to trial data showed how valuable this approach might be. Fifty-six patients entered a double-blind placebo-controlled trial of a non-steroidal anti-inflammatory agent. The four physicians who assessed the patients were asked to make an overall judgment about the benefit their patients had derived from treatment (Table 1). The treatment was more effective than placebo ($\chi^2 = 42.5$, $P < 0.05$). Judgment analysis was performed on each of the physicians, relating overall benefit to the clinical measures recorded during the trial. Patients were then reclassified using the policy of the doctor who had assessed them. This eliminated the random variation and inconsistency of each clinician in applying his own policy. The new patient categories, of which the mean scores are shown in column B, indicate a greater distinction between treatment and placebo ($\chi^2 = 49.4$), showing that the reduced variability of the judgments has reduced the variance of the trial and hence increased its discriminating

TABLE 1. *Patients classified by doctor's judgment of overall efficacy (0–3).*

	Mean score	0	1	2	3	χ^2
Actual judgments						
Treatment	1·92	1	7	23	6	42·5
Placebo	1·05	7	6	4	2	
Judgments reclassified using CJA						
Treatment	2·03	0	6	24	7	49·5
Placebo	1·05	6	7	5	1	

power. Greater power can be obtained by applying the policy of a particular physician to all the observations, or by applying a consensual policy to all patients. These methods would (unlike the example given above) impose a uniform policy for all patients; the first has been used in a trial of antidepressive drugs [25] with predictable improvement in the power of the trial.

Several groups have considered what information a clinical trial report should contain [31]. However, there is little information about what data readers of clinical trial reports actually use when reaching decisions about the treatments under test. Clinical judgment analysis can help to identify those items of information that are selected and make their weighting explicit. To this end, 38 rheumatologists were asked to judge how helpful 50 trial reports were in deciding which (if any) of the treatments compared should be used. Each report described the amount of information available in 11 categories including design, conduct and analysis. Details of the experimental design and trial sampling procedure received greatest weight, whereas the number of withdrawals and the reasons for them appeared less important [32]. Studies of this kind can help trial producers improve the convincing power of their reports by including the information readers use.

EDUCATION

It is doubtful if the environment of clinical medicine is suitable for students to learn by 'osmosis' or by simply observing the decisions of experts and attempting to relate them to items of patient data. Simultaneous presentation of many items of data (many of them non-numerical), the need for rapid processing and the probabilistic relationship of clinical variables to the underlying event predispose to an intuitive rather than an analytical approach [33]. Evidence from many sources strongly suggests that learning in such environments may only occur effectively if students are given information not about the outcome of an expert's judgment processes but some knowledge of the process itself [1].

The success of this approach has now been shown in at least one medical undergraduate setting [34]. Allowing students to compare their own diagnostic policy models with those of an expert (cognitive feedback) enabled them to learn the appropriate decision rule more quickly than simple outcome feedback on the correctness of their diagnosis. Ideally, what is the appropriate decision rule should be discovered from analysis of a large clinical data base rather than by reference to expert weights. Where there is some 'gold standard' for diagnosis, such as laparotomy findings or pathology results, the 'correct weights' attributable to each data item can be generated (Fig. 1). When cognitive feedback was used to teach student health physicians the optimal policy for diagnosing streptococcal throat infections there

were significant improvements in their diagnostic abilities. These were found to persist beyond the experimental period into the doctors' clinical practice [35]. The same group, in a study on the diagnosis of pulmonary embolism, used pulmonary angiography as a 'gold standard' for establishing the diagnosis. This allowed 'ideal' or objective weights to be attributed to each of the clinical predictor variables. Comparison of these weights with the doctors' own weights derived by clinical judgment analysis showed that they attributed least weight to the heart rate whereas it was the most important variable in the objective analysis. Information of this sort allows doctors to conform their policies to one having established diagnostic power. The obvious advantage over simply learning the weights handed out by an 'expert' are illustrated by the finding that, in this study, the policies of expert faculty members were as variable as those of the students [36].

Judgment policy modelling offers a means of generating more uniform management policies, carefully chosen and consciously adhered to by clinicians, both individually and in groups. At a time when job rotations and duty rotas require patient care to be divided among increasing numbers of physicians (and others), any help in making management policy both explicit and consistent would be welcome.

HOW DO CLINICIANS REACT TO CLINICAL JUDGMENT ANALYSIS?

In the only study of this question of which we are aware, the reactions of general practitioners in the UK ranged from full acceptance through amused tolerance to a rather irritated dismissal of the study methods as 'totally artificial and invalid' [37]. However, even busy general practitioners did not consider the task of making 50–60 sets of judgments particularly arduous. One centre that had at first specifically asked to be involved later wished to be dissociated from the study after considerable differences in diagnostic and prescribing behaviour were revealed. In general, paper patients with otitis media were judged to be not very life-like, but nevertheless computed judgment models were rated as 'credible'. One large group practice, studied in detail, felt that the revealed differences in prescribing behaviour were probably a fair reflection of the truth; and although it was considered important that differences be resolved they believed this could only be done with difficulty. Nevertheless, several doctors felt that their own policies had changed as a result of seeing them in the perspective of judgment policies of other practice members, though whether this belief was justified was not tested.

Most general practitioners work in a highly intuition-inducing real-life environment in which large amounts of data, often pictorial, are presented simultaneously or over a few minutes. Many diagnostic and therapeutic decisions have to be made within a short period of time if the day's work is to be accomplished. Furthermore, it is generally accepted that these decisions have to be made with a high degree of confidence and accuracy if professional appearances and status are to be maintained. Under these circumstances it is understandable that an approach which allows disclosure of inconsistency or conflict of judgments may be unwelcome [38].

Rheumatologists on the other hand work more often with numerical data and to them printed simulations may differ little from standard hospital records and summaries. Many are associated with undergraduate or postgraduate education and so are used to exposing their decision policies to scrutiny. Most were content to participate in the studies for their own sake and, unlike the general practitioners, required little convincing of their ultimate value in terms of improved patient care. Psychiatrists on the other hand, though generally lacking reliable (or valid) methods of measurement that can be exteriorized are generally

aware of their need for help in reducing sources of variation. Consequently they have generally accepted the help provided by judgment analysis when it has been offered.

WHERE NEXT?

Judgment analysis has been used to make statistically firm models of doctors' diagnostic and therapeutic decisions. However the promise of this approach has not yet been translated into measured improvements in the standard of patient care. Prejudice against the use of computer technology is not an obvious cause – the use of desk-top microcomputers as information sources seems widely acceptable amongst general physicians and hospital doctors – yet many have expressed serious reservations about the analyses involved in clinical judgment analysis. This may be justified if the regression models are presented as representations of how the doctor making the judgments actually thinks, instead of, as most Brunswikians probably believe, as paramorphic models of 'the doctor thinks as if . . .'. The implication remains that these models may nevertheless help decision-takers gain insight into what is really happening. Whether such insight is ever revealed is, of course, unknown since what 'really happens' remains inaccessible. The unfamiliar form of the regression models does however psychologically distance the doctors whose judgments have been modelled. Models that suggest unpalatable inconsistency or discordance can easily be discarded on grounds of irrelevance. Unfortunately it is in just these situations that doctors need help with the resolution of conflicts. The superiority of social judgment analysis over Delphi and the nominal group technique at achieving consensus among groups has already been demonstrated [39]. Whether clinical judgment analysis will prove equally successful with groups of doctors remains to be seen. It will depend in part on whether participants can admit the possibility of inconsistency and even error and how committed they are to reaching a consensus. KR Hammond has proposed that unless they admit to desperation, judgment analysis is unlikely to help resolve conflicts of judgment between participants. Where decisions concern matters of ethics rather than matters of science judges often seem most reluctant to abandon their prepared positions. Even here clinical judgment analysis may help reveal some of the underlying factors. In a study of decisions over the use of tube feeding in seriously ill patients, physicians could be classified as more or less paternalistic according to the weights attributed to patient preferences [40].

Expert systems that do not draw on statistical databases must use knowledge elicited from experts. The finding that experts have difficulty describing their judgment processes suggests that more objective procedures may be required to obtain this knowledge other than simply asking the expert to explain his reasoning. Clinical judgment analysis offers a method of knowledge elicitation which has wide applicability [41]. However, because the procedure aims to create statistically firm models, it is considerably more time consuming than traditional interview methods. So far, it has only been used in situations where decisions are simple one-step affairs in which all the relevant data items can be considered simultaneously. Whether clinical judgment analysis would remain a practical alternative in the multi-stage discriminatory tasks usually associated with medical diagnosis remains to be seen. A potential disadvantage of mathematical models lies in their very mathematical form. This may be a highly efficient description of their function and an aid to their application, but it bears little resemblance to the generally accepted medical model of disease and its treatment. This is of little importance in expert systems that are not designed to be interrogated, but will need disguising in knowledge-based systems designed to have explanatory facilities.

A major attraction of the lens-model paradigm and its regression modelling approach lies in its general applicability to many problems of human judgment and these have recently

been reviewed [42]. Its ability to operate outside biological or sociological areas of expertise permits involvement of the scientifically naive. Several studies have shown how Brunswikian methods involving lay persons have improved experts' understandings of public needs in the field of social policy formation [43–45]. Patients' needs and expectations of treatments are obvious targets within the medical field. So far we know little about how patients weight the efficacy and toxicity in judging the acceptability of treatments. A series of studies examined doctors' and patients' decisions to recommend or accept hormone replacement therapy. Judgment analysis was used to examine the weights attached to the benefits and risks of this treatment. Patients' policies clustered into four groups depending on the importance they attached to relief of hot flushes, the risk of osteoporosis, resumption of cyclical bleeding and the risk of cancer. Cancer risk featured prominently in the physicians' models while many patients appeared more concerned with the mortality and morbidity of fractures [46]. This information should help the development of consensus management plans incorporating patients' and possibly relatives' models as well as those of the medical team.

If rational analysis of judgment policies (rather than intuitive guessing) proves as helpful, for instance, as measuring haemoglobin concentrations (rather than guessing their value from the colour of the nailbed) we can look forward to many developments in this area. These advances, using modern technology as an aid rather than a replacement, would expand rather than contract the physician's role in treating his patients, just as the pathology laboratory and other technical developments have done in the past.

REFERENCES

1. Hammond KR, Stewart TR, Brehmer B, Steinmann DO. Social judgment theory. In: Arkes HR, Hammond KR (eds), Judgment and decision making. Cambridge, Cambridge University Press, 1986, pp. 56–76.
2. Brunswik E. The conceptual framework of psychology. In: International Encyclopedia of Unified Science (Vol 1 No 10). Chicago, University of Chicago Press. 1952.
3. Hammond KR, Stewart TR, Brehmer B, *et al.* Social judgment theory: In: Kaplan MF, Schwartz S (eds), Human Judgment and Decision Processes. New York, Academic Press, 1975, pp. 271–312.
4. Dawes RM. The robust beauty of improper linear models in decision making. Am Psychol 1979; 34: 571–582.
5. Chaput de Saintonge DM, Hathaway HR. Antibiotic use in otitis media: patient simulations as an aid to audit. Br Med J 1981; 283: 883–884.
6. Kirwan JR, Chaput de Saintonge DM, Joyce CRB. Clinical judgment in rheumatoid arthritis. I. Rheumatologists' opinions and the development of 'paper patients'. Ann Rheum Dis 1983; 42: 644–647.
7. Rovner DR, Rothert ML, Holmes MM, Schmitt N, Given CW, Ialongo NS. Validity of structured cases to study clinical decision making in urinary tract infection. (Unpublished).
8. Kirwan JR, Chaput de Saintonge DM, Joyce CRB. Clinical judgment in rheumatoid arthritis. III. British rheumatologists' judgments of 'change in response to therapy'. Ann Rheum Dis 1984; 43: 686–694.
9. Kirwan JR, Currey HLF, Brooks PM. Measuring physicians' judgment – the use of clinical data by Australian rheumatologists. Aust NZ J Med; 1985; 15: 738–744.
10. Kirwan JR, Bellamy N, Condon H. Judgment of 'current disease activity' in rheumatoid arthritis – an international comparison. J Rheumatol 1983; 10: 901–905.
11. Fisch H-U, Hammond KR, Joyce CRB and O'Reilly M. An experimental study of the clinical judgment of general physicians in evaluating and prescribing for depression. Br J Psychiatr 1981, 138: 100–109.
12. Fisch H-U, Hammond KR, Joyce CRB. On evaluating the severity of depression: an experimental study of psychiatrists. Br J Psychiatr 1982; 140: 378–383.
13. Slovic P, Rovner LG, Hoffman PJ. Analysing the use of diagnostic signs. Invest Radiol 1971; 6: 18–26.

14. Chaput de Saintonge DM, Hattersley LA. Antibiotics for otitis media: can we help doctors agree? Fam Pract 1985; 2: 205–212.
15. Kirwan JR, Currey HLF. Clinical judgment in rheumatoid arthritis. IV. Rheumatologists' assessments of disease remain stable over long periods. Ann Rheum Dis 1984; 43: 695–697.
16. Kirwan JR, Chaput de Saintonge DM, Joyce CRB. Inability of rheumatologists to describe their true policies for assessing rheumatoid arthritis. Ann Rheum Dis 1986; 45: 156–161.
17. Hoffman PJ. The paramorphic representation of clinical judgment. Psychol Bull 1960; 57: 116–131.
18. Lane DM, Murphy KR, Marques TE. Measuring the importance of cues in policy capturing. Org Behav Hum Perform 1982; 30: 231–240.
19. Kirwan JR. Social Judgement Theory and the Rhuematic Diseases: Application of Clinical Judgement Analysis to Rheumatoid Arthritis. MD Thesis (Chapter 8, pp 106–116). University of London, 1983.
20. Wigton RS. Use of linear models to analyse physicians' decisions. Med Decision Making 1988; 8: 241–52.
21. Kirwan JR, Chaput de Saintonge DM, Joyce CRB. Clinical judgment in rheumatoid arthritis. II. Judging 'current disease activity' in clinical practice. Ann Rheum Dis 1983; 42: 648–651.
22. Kirwan JR, Chaput de Saintonge DM, Joyce CRB. Clinical judgment analysis – practical application in rheumatoid arthritis. Br J Rheumatol 1983; 22(suppl): 18–23.
23. Hammond KR, Summers DA, Deane DH. Negative effects of outcome feedback in multiple cue probability learning. Org Behav Hum Perform 1973; 9: 30–34.
24. Kirwan JR, Barnes CG, Davies PG, *et al.* Analysis of clinical judgment improves agreement in disease assessment. Ann Rheum Dis 1988; 47: 138–143.
25. Bech P, Haaber A, Joyce CRB, and the Danish University Antidepressant Group. Experiments on clinical observation and judgment in the assessment of depression: profiled videotapes and judgment analysis. Psychol Med 1986; 16: 873–883.
26. Chaput de Saintonge DM. Clinical trials in otitis media – to whom do the results apply? Proceedings of 3rd World Conference on Clinical Pharmacology and Therapeutics, Stockholm 1986.
27. Jachuck SJ, Brierly H, Jachuck S, *et al.* The effect of hypotensive drugs on the quality of life. J R Coll Gen Pract 1982; 32: 103–105.
28. Chaput de Saintonge DM, Kirwan JR. How can we discover what treatment responses are clinically important? Br J Clin Pharmacol 1985; 20: 530–531.
29. Stewart TR, Joyce CRB. Increasing the power of clinical trials through judgment analysis. Med Decision Making 1988; 8: 33–38.
30. Chaput de Saintonge DM, Kirwan JR, Evans SJW, *et al.* How can we design trials to detect clinically important changes in disease severity? Br J Clin Pharmacol 1988; 26: 385–362.
31. Blum AL, Chalmers TC, Deutsch E, *et al.* The Lugano statement on controlled clinical trials. J Int Med Res 1987; 15: 2–22.
32. Chaput de Saintonge DM, Hattersley LA, Kirwan JR. What makes a clinical trial report helpful? A preliminary analysis of rheumatologists' judgments. Br J Rheumatol 1983; 22(Suppl): 59–66.
33. Hammond KR, Hamm RM, Grassia J, Pearson T. Direct comparison of the efficacy of intuitive and analytical cognition in expert judgment. IEEE transactions on systems, man and cybernetics SMC-17:753-770, 1987.
34. Wigton RS, Patil KD, Hoellerich VL. The effect of feedback in learning clinical diagnosis. J Med Educ 1986; 61: 816–822.
35. Poses RM, Cebul RD, Wigton RS, Collins M. Feedback on stimulated cases to improve clinical judgment. Med Decision Making 1986; 6: 274.
36. Wigton RS, Hoellerich VL, Patil KD. How physicians use clinical information in diagnosing pulmonary embolism. Med Decision Making 1986; 6: 2–11.
37. Chaput de Saintonge DM, Hattersley LA (unpublished).
38. Katz J. Why don't doctors disclose uncertainty? The Hastings Centre Report February 1984; 35–44.
39. Rohrbaugh J. Improving the quality of group judgment: social judgment analysis and the nominal group technique. Org Behav Hum Perform 1981; 28: 272–288.
40. Smith DG, Wigton RS. Modelling decisions to use tube feeding in seriously ill patients. Arch Intern Med 1987; 147: 1242–1245.
41. Chaput de Saintonge DM, Cookson MJ. The role of clinical judgment analysis in the development of medical expert systems. In Hunter J, Cookson J, Wyatt J (eds) Lecture notes in Medical

Informatics No. 38: AIME 89, Second European Conference on Artificial Intelligence in Medicine. Springer-Verlag 1989, pp. 3–14.
42. Brehmer B, Joyce CRB (eds). Social judgment – The SJT view. Amsterdam, North-Holland. 1988.
43. Earle TC. Risk judgment, risk communication and conflict management. In: Brehmer B, Joyce CRB (eds), Human Judgment: The SJT view. Elsevier, North Holland 1988, p. 361–400.
44. Mumpower J, Veirs V, Hammond KR. Scientific information, social values, and policy formation: the application of simulation models and judgment analysis to the Denver regional air pollution problem. IEEE transactions on systems, man and cybernetics 1979; 9: 464.
45. Rohrbaugh J, Wehr P. Judgment analysis in policy formation. Report no. 201, Centre for Research on Judgment and Policy, Institute of Behavioural Science, University of Colorado, 1978.
46. Holmes MM, Rovner DR, Rothert ML, *et al.* Women's and physicians utilities for health outcomes in oestrogen replacement therapy. J Gen Intern Med 1987; 2: 178–182.