Analysis of clinical judgment helps to improve agreement in the assessment of rheumatoid arthritis

J R KIRWAN, C G BARNES, P G DAVIES, AND H L F CURREY

From the Bone and Joint Research Unit and The Department of Rheumatology, The London Hospital and Medical College, London

SUMMARY Experienced rheumatologists differ widely in their assessments of rheumatoid arthritis even after extensive efforts to improve agreement by discussion and consensus. The use of computer feedback to provide an analysis of clinicians' judgment policies in a highly structured investigation has been shown to improve agreement, but this may not apply in normal clinical practice. Here the successful convergence of clinical agreements by three rheumatologists using computer assisted feedback over several months in a National Health Service outpatient department is reported. In the three months without feedback their pooled agreement for assessing the severity of rheumatoid arthritis was $r^2=0.62$. During the three months in which feedback was provided agreement rose to $r^2=0.92$. The principal component of all three judgment policies at the end of the feedback period was 'articular index'.

Key words: computer feedback, articular index, decision making.

There are many ways of assessing the severity of chronic disease. For rheumatoid arthritis (RA) even careful and experienced rheumatologists differ widely in their assessments when provided with identical patient data.1,2 Even after extensive efforts to improve agreement by discussion and consensus differences in judgment persist.3 This may be due to the difficulty faced by physicians in appreciating their own decision making processes.4-6 In a survey of 89 rheumatologists only 34% of the variance of their actual decisions could be explained by the descriptions they gave of their judgmental process.6 Clinical judgment analysis (CJA)1-4 has provided models of the way individual physicians' judgments relate to the available clinical data. Unlike their own descriptions, these 'judgment policy models' may explain up to 95% of the variation in rheumatologists' decision making.4 They show marked differences in the practice of clinical judgment, which will contribute to the important variations inherent in all clinical investigations of pathogenesis and therapy. If the consistency of clinical judgment could be improved this would benefit not only clinical investigations (such as the classification of disease groups and clinical trials) but also therapeutic decision making for individual patients and the rational use of patient investigations.

A major aim in studying variation in clinical judgment is therefore to reduce it. Recent studies have suggested that the insight provided by CJA into the reasons why physicians disagree may also provide a new opportunity to improve agreement about patient assessment.3,5,7,8 In one of these studies the provision of CJA feedback to two rheumatologists enabled them to achieve a level of agreement not possible even after extensive discussions held specifically for that purpose.3 This was in the context of a carefully constructed clinical investigation, however, and may not reflect the difficulties of normal clinical practice. Here we report the successful convergence of clinical assessment by three rheumatologists using computer assisted CJA feedback over several months in a National Health Service outpatient department.

Patients and methods

OVERALL DESIGN

Three rheumatologists took part and each recorded patient data and their assessment of the severity of disease on standard forms for patients with RA who attended their outpatient clinic during the 33 weeks

Accepted for publication 24 July 1987.
Correspondence to Dr J R Kirwan, Department of Medicine, Bristol Royal Infirmary, Bristol BS2 8HW.
of the study. All patients were examined by a
metrologist before they saw the rheumatologist, and
the patient data collected by the metrologist were
immediately entered into a microcomputer in the
outpatient examination room.

During the arbitrary nine week run-in period,
which familiarised all concerned with the pro-
cedures, no further use was made of these data.
Over the next three months no information was
supplied to the rheumatologists, but at the end of
that time the data collected were used to derive
dgment policy models for each one. In the final
three months, feedback was provided to each
rheumatologist about his judgment policies and
those of his two colleagues. Some of this feedback
resulted from computer calculations on the clinical
data entered by the metrologist and was displayed
for the rheumatologists to see during their consulta-
tion with the patient (see below). The data collected
were also used to calculate judgment policies
employed during this period of feedback.

The observed (calculated) assessment policies of
the rheumatologists were then compared for the
three months without feedback and the three
months with feedback to seek evidence of improved
agreement between them which might be attributed
to the provision of feedback.

RHEUMATOLOGISTS

The three rheumatologists were experienced and
had been working together for at least four years,
sharing outpatient clinics and participating in weekly
case conferences, at which patient assessment was
frequently discussed. At the outset they had been
informed of the purpose of the study.

PATIENTS

All patients with classical or definite RA\textsuperscript{9}
attending the outpatient clinics concerned during the study
were included. Because of the length of the study
many patients were included on two or more clinic
visits, but disease assessments were made independ-
ently each time.

PATIENT DATA

A nurse-metrologist\textsuperscript{10} interviewed each patient
before each consultation. The metrologist was
trained before the study to record the following
clinical data directly onto a microcomputer: dura-
tion of early morning stiffness (EMS) (in minutes);
patient’s own global assessment of the state of their
arthritis (GLOB) (0=no problem, 4=worst possible);
an articular index (AI) (the number of clinically
inflamed joints in the limbs); grip strength
(GRIP) (the mean of two measurements in each
hand); the modified Health Assessment Question-
naire Disability Index (HAQ-DI)\textsuperscript{11} (score
0.0–3.0); patient’s assessment of pain (PAIN)
taken from the HAQ visual analogue pain scale\textsuperscript{12}).

The rheumatologist also recorded these same data
during the consultation using standard forms except
that scores for GRIP were provided on the front of
the patient’s notes by the metrologist and disability
and pain were recorded on a 0–3 scale.

JUDGMENTS

Immediately after each consultation the rheuma-
tologist recorded his assessment of the patient’s
‘current disease activity’ on a 10 cm visual analogue
scale.\textsuperscript{4, 13}

JUDGMENT POLICY MODELS

All possible subsets regression analysis\textsuperscript{14} was used to
calculate policy models as previously described,\textsuperscript{1, 4}
taking the clinical data recorded by the rheuma-
tologist as the independent variables and the assess-
ments as the dependent variable. Policy models
were separately calculated for the periods without
and with feedback.

FEEDBACK

During the second three months of the study
feedback was provided in two ways. Each physician
was shown graphical representations of a model of
his own judgments of severity (his policy) and those
of his two colleagues (Fig. 1a). All three were visible
to him during his outpatient clinic sessions. In
addition, the policy models were also stored in the
microcomputer and used to calculate predicted
assessments by all three clinicians as soon as the
metrologist had entered the patient data. These
were then clearly indicated on the front of the
patient’s notes.

AGREEMENT

Clinical observations made on 240 further patients
not used in the calculation of judgment policy
models were used to test for similarities between the
way the rheumatologists assessed their patients. An
estimate of the assessment each rheumatologist
would have made had he seen the patients without
any feedback was obtained by applying his initial
policy model to each case. The three sets of
assessments were then compared for agreement
using the rank correlation coefficient. This process
was repeated on the same cases using the policy
models calculated when feedback was provided.

In addition, the relative contributions of the
clinical variables to each policy model\textsuperscript{4} were com-
pared without and with feedback.
Fig. 1  Judgment policy models of three rheumatologists (a) without and (b) with feedback. The relative contribution of each variable is shown on a percentage scale. EMS=early morning stiffness; GLOB=patient's own global assessment; AI=articular index; GRIP=grip strength; HAQ-DI=disability index; PAIN=pain score.
Results

Each rheumatologist took part in about 25 consultations during the run-in period, 40–50 consultations without feedback, and 40–50 while feedback was provided. Fig. 1a shows the graphical representations of the models derived from the first part of the study, when the rheumatologists knew that agreement between them was being sought and were familiar with the procedures for data collection, but

Fig. 2 Agreement between rheumatologists (a) without and (b) with feedback. Assessments are taken from the application of the judgment policy models to data from 240 outpatients not involved in policy model calculations and are scaled 0 (no current disease activity) to 100 (maximum possible current disease activity).
had been receiving no direct feedback about their assessments.

These histograms were supplied to the rheumatologists and were available to them during the second part of the study. After this the agreement between the three rheumatologists without and with feedback was calculated and is shown in Figs 2a and 2b respectively. Doctors L and M agreed closely. Overall agreement with feedback, indicated by the pooled value of the correlation coefficients, improved significantly from \( r=0.79 \) \( (r^2=62\%) \) to \( r=0.96 \) \( (r^2=92\%, \ p<0.001) \). The correlation between doctors L and M was slightly reduced, though it remained high, and that between doctors K and L and K and M improved significantly.

The contributions made by each clinical variable to the severity assessments based on the policy model equations (its weight) without and with feedback are shown in Fig. 1, together with a measure of the success of each modelling procedure in explaining the variance in each individual’s judgment, given as \( R^2 \). \( R^2 \) was lower for the with feedback models than for those without feedback. This may have resulted from the inclusion of all judgments made during feedback in the calculation of the with feedback policy models even though changes in policy probably occurred only gradually.

Doctors L and M showed agreement in the weight they attributed to articular index without feedback, but no variable appeared in all three models. With feedback ‘articular index’ was the principal component in each model and similarity was generally increased in that two of the five remaining variables, the patient’s own global assessment and HAQ disability index, made minor contributions to each model, and two, grip strength and pain, were omitted from all three models. Although early morning stiffness is often considered an important measure of disease,\(^{15,16} \) it was included in only one policy model on each occasion.

**Discussion**

Whereas previous studies of CJA have hinted at the possibility of increasing clinical agreement in routine practice, these results have shown that this end can in fact be achieved. Rheumatologists in general adopt a wide variety of patient assessment policies, though some show a similar approach.\(^1,2\) Two of those taking part in this study had already adopted similar assessment policies at the outset, making it inherently more difficult to demonstrate that feedback improved agreement. Nevertheless, agreement between these two remained high and that with the third participant was improved considerably.

It might be argued that the introduction of new procedures into the routine outpatient clinics and the knowledge that the experiment sought to improve agreement would, themselves, have increased the likelihood of finding common ground between the rheumatologists involved. This was taken into account by providing a run-in period before the first part of the experiment, ensuring that all were familiar with the procedures. Furthermore, all participants were aware throughout the study that the experiment was designed to measure agreement, and any effect this might have had operated both before and after the feedback was available.

Two kinds of feedback were provided simultaneously: graphical representations of the participants’ initial assessment policies and predictions of the assessments of each patient seen subsequently, based on these policies. It is not possible to separate the effects of the two methods in this experiment. In a previous study\(^3\) disagreement in RA patient selection was not reduced by simple ‘outcome’ feedback (the provision to each other of the participants’ judgments without information about their policy models) but was successful with process feedback, in which the policy model histograms were provided. Similar results were found when medical students were taught to diagnose urinary tract infections.\(^17\) In the present study predicted judgments were based on observations made by the nurse-medirologist, rather than on those made by the physician, and this probably reduced the accuracy of the predictions. It seems likely, however, that policy model feedback contributed significantly to the improved agreement.

Differences in patient assessment of the magnitude of those reported here and previously\(^1,2,4,5,18\) imply that the identification of patients with active or inactive current disease for treatment or research purposes may depend to a considerable extent on the clinician making the assessment as well as on the patient’s own clinical signs and symptoms. Though it might be thought that these differences relate to individual patient care, in practice they reduce the reliability of assessment and hence its validity.\(^19\)

One way of increasing uniformity between physicians is to encourage discussion and interchange of ideas, but the clinicians in this study had already been involved in regular joint meetings of this sort, and physicians are not good at explaining the process by which they reach decisions.\(^6\) This experiment and others\(^7,19\) suggest that feedback from judgment analysis achieves greater agreement and that increasing reliability of judgments both within and between physicians makes attainment of valid results more probable.\(^20\) Whether other types
of feedback, such as the Delphi method, would work in these circumstances is not known.

Before policy model feedback can be recommended for routine practice, however, unequivocal evidence that the procedure is worthwhile must be obtained. Several retrospective applications of CJA to completed clinical trials have suggested that their statistical power may indeed be increased. A prospective application of the technique in these circumstances is now warranted.

Perhaps the most important question raised by investigations of this kind is 'which is the correct policy?' The pathogenesis of RA is unknown, and the prognostic implications of its various clinical features obscure. Although tools now exist which may help to coordinate judgment, the major outstanding task for the rheumatologist in this area must be the elucidation of the clinical and prognostic significance of the features on which judgments are currently based. Only then will it be possible to educate ourselves to follow the most appropriate assessment policy.

We thank Mrs Mary Huggins SRN for her excellent technical assistance as a metrologist and Drs D M Chaput de Saintonge and C R B Joyce for helpful comments on the manuscript. J R K was supported by the Arthritis and Rheumatism Council for Research.

References

Analysis of clinical judgment helps to improve agreement in the assessment of rheumatoid arthritis.

J R Kirwan, C G Barnes, P G Davies and H L Currey

doi: 10.1136/ard.47.2.138

Updated information and services can be found at:
http://ard.bmj.com/content/47/2/138

These include:

Email alerting service
Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

Notes

To request permissions go to:
http://group.bmj.com/group/rights-licensing/permissions

To order reprints go to:
http://journals.bmj.com/cgi/reprintform

To subscribe to BMJ go to:
http://group.bmj.com/subscribe/