How can we design trials to detect clinically important changes in disease severity?

D. M. CHAPUT DE SAINTONGE¹, J. R. KIRWAN², S. J. W. EVANS³ & G. J. CRANE¹

¹Department of Clinical Pharmacology and Therapeutics, ²The Bone and Joint Research Unit and ³Department of Clinical Epidemiology, The London Hospital Medical College, Turner Street, London E1 2AD

1 Forty-eight British rheumatologists judged the change in disease activity in 50 sets of patient data drawn from life and presented as 'paper patients'. Each set comprised two values, recorded a year apart, for 10 commonly measured clinical variables. Doctors recorded the size of improvement or deterioration on a visual analogue scale (VAS) and whether the change was clinically important or not.

2 Clinical judgement policies were modelled using linear regression of the clinical variables on the VAS score.

3 Doctors showed little agreement over which patients had improved and which had not. Possible reasons could be discovered by inspecting their judgement policies.

4 The weights attributed to the clinical variables differed considerably between doctors. Furthermore weights the doctors believed they attached to the variables frequently differed from the weights in the regression models.

5 These models could be used to calculate the smallest change required in any clinical variable before it would be considered clinically important. However, the size of such changes was often outside the observed clinical range suggesting that the use of single outcome variables is unrealistic.

6 The modelling procedure described can be applied during the planning stage of the trial to participating physicians, patients, health economists or any other group having an interest in the results. The models themselves can then be used to reach a consensus policy for judging what is a successful outcome. This may be expressed as a linear combination of specific outcome measures. Its use may improve the power of clinical trials and the relevance of their results.

Keywords regression modelling clinical significance clinical judgement analysis clinical trials rheumatoid arthritis

Introduction

Clinical trials should be designed with adequate power to detect clinically important effects. To the patient an important effect is a useful reduction in the manifestations of disease which trouble him. To the clinician it means, in addition, that a worthwhile proportion of patients will achieve such an effect. To the health economist or public health doctor on the other hand even a tiny improvement may be worth having if the disease

is common enough and the treatment safe and cheap enough. In practice however, trials are frequently designed with little reference to how the results could be used. Investigators are encouraged to look for differences which 'might well exist', or 'maximally envisaged' improvements. This failure to distinguish between anticipated and clinically important goals leads to trials which may fail to address clinically relevant

issues (Spiegelhalter & Freedman, 1986). But what is meant by the terms 'clinically important' and 'worthwhile', and how is a successful treatment outcome judged in practice? This study shows how unbiased methods have been developed to answer these questions, how the data can be used to help choose outcome variables in clinical trials, and how such trials can be designed with adequate power. Rheumatoid arthritis was chosen as a model because it exemplifies the problem of assessing changes in disease severity when a wide variety of symptoms, signs and laboratory results must be integrated to form an overall judgement.

Although questionnaire or interview methods are often used to answer questions of this sort, there is evidence that doctors' answers to questions about how they make integrated judgements can differ significantly from what they do in practice. Studies both with general practitioners (Chaput de Saintonge & Hattersley, 1985) and experienced rheumatologists (Kirwan *et al.*, 1986) have shown that doctors are unable to provide adequate explanation for their clinical decisions. Judges' poor ability to recall and describe the basis for their decisions has been described in many situations (Hoepfl & Huber, 1970; Einhorn & Hogarth, 1978). Subjective information gained by interview method has been proposed as a way of defining clinically important effects since 'by definition adequate objective information does not exist' (Spiegelhalter & Freedman, 1986).

We have established that in response to written sets of clinical data 'paper patients' doctors' decisions about diagnosis, disease severity and treatment are similar to those they reach in real life (Kirwan *et al.*, 1983a). This has allowed us to compare doctor's decisions under standardised conditions which, although artificial, allow valid inferences to be drawn about clinical judgements.

Methods

Data on 10 standard clinical variables were collected on 30 patients who entered into a clinical trial, at the start and after a 1 year interval. The values for each variable 'before' and 'after a year' were presented on a single sheet of paper. Therefore, though 'paper patients', they were not fictitious or specially selected. Twenty cases were included in duplicate to check the consistency of doctors' responses, so there were 50 cases in all. These were presented to a sample of British rheumatologists chosen using random numbers from the membership of the Heberden Society. They were asked to decide the degree of change in each patient's condition by marking a

visual analogue scale (VAS). This bore at either end the labels 'greatest possible deterioration' and 'greatest possible improvement'. The respondents were also asked to state for each patient whether they regarded the change as clinically important or not. Finally each respondent was asked to award marks out of a total of 100 to indicate the relative importance of each variable. These marks were a measure of the variables' perceived relative importance and formed the expressed policy.

The judgements made by each rheumatologist about the progress of each patient were analysed to generate a modelled policy as previously described (Kirwan *et al.*, 1983b). In brief, all possible subset multiple regression analysis (BMDP 9R) was used taking the untransformed VAS judgement as the outcome variable and the log of the ratio of the final to the initial value of each clinical variable as the predictor variables. This transformation was chosen because it expresses improvement as a proportion of the patients' initial severity. Several alternative transformations exist, as do ways of choosing them. The regression equation derived in this way was taken as a mathematical model of the judgements and the importance of the clinical variables it included was expressed as their relative contribution to the total variance explained by the model. For each respondent, threshold values of clinical importance were obtained by examining his VAS scores for the set of paper patients on which he had made judgements. With increasing VAS scores judgements of clinical importance changed from 'not clinically important' to 'clinically important'. In all cases this change occurred over a 5 mm length of the scale or less. The midpoint of this transition zone was designated the threshold value of clinical importance. Values for threshold changes in single clinical variables which might be judged clinically important were obtained by substituting the threshold value of the VAS in each doctor's judgement model equation. Changes in all other clinical variables except one were set to zero and the equation solved for each variable in turn. This gave the threshold change required in any single clinical variable before the change reached 'clinical importance'.

Consistency of judgements was assessed by the Spearman rank correlation coefficient (r_s) for the duplicate cases. Agreement between doctors about clinical importance was measured using Cohen's weighted kappa (Cohen, 1968). Disagreement between clinically important deterioration and clinically significant improvement was weighted twice that between significant changes in either direction and no clinically significant change.

A consensus model was generated by weighting each individual's contribution by the measure of his consistency (r_s). The VAS scores for severity of each patient were multiplied by each doctor's r_s . The resulting products were summed for each patient and then divided by the sum of the r_s giving a single weighted VAS score for each patient. A regression analysis was then performed on these VAS scores to give a consensus model weighted by consistency.

Models of the consistent doctors ($r_s > 0.8$) were compared with a group of 15 selected at random from the total respondents. For each group weighted consensus models were derived from a random 80% subset of the paper patients. Predicted scores were generated for each group by applying the model equations to the remaining 20% of patients. The correlation between these predicted judgements was used as a measure of similarity between the models. To compare the relative importance of each variable for the consistent and random subgroups, the contribution of each variable to the total explained variance within the model was calculated.

Results

Forty-eight rheumatologists (86% of those approached) completed the study. In 15 doctors the correlation coefficients on duplicate patients exceeded 0.8 (that is, they were relatively reliable in their patient assessments) and their policies could be modelled in a way which accounted for more than 80% of the variance. Even these doctors showed little agreement ($\kappa = 0.3$) over which patients had shown clinically important improvement (Figure 1). Indeed, certain patients

were considered improved by some doctors and deteriorated by others. The reason for this could be seen from the policy models. There were marked differences in the weight that different doctors attached to the clinical variables (Table 1). However, some of the 10 cue variables appeared in most doctors' models (such as early morning stiffness and ESR) whereas some (such as pain score) appeared in none or few. This should be contrasted with the expressed policy weights (Table 2). Dr 1's opinion that erosions and functional capacity are the most important determinants of his judgements is supported by his modelled policy weights. However, agreement between the modelled and expressed policy weights is poor for most other doctors.

Strictly speaking, the doctors' modelled and expressed weights are not directly comparable. The former are derived from a multivariate analysis which minimises the number of contributing variables whereas the latter is probably similar to a univariate analysis which takes no account of the information content of other variables. Nevertheless, patients' global assessment appears as one of the most important variables in the expressed policies but one of the least important in the models.

Threshold values for changes in single variables showed wide variation. In a number of cases the changes required before the patient would be judged usefully improved are outside the observed clinical range. This means that no change in the variable concerned could, in itself, be sufficient to determine the verdict. It appears that Dr 2, for example, would need to see a 181% rise in the haemoglobin if that was the only variable which changed. In other cases the presence of a positive intercept term allowed even a small

Figure 1 Doctors' judgements of change in disease severity for each of 50 patients. ■ worse, ▨ no significant change, □ better.

Table 1 Modelled policy weights expressed as percentage of the explained variance

<i>Doctor</i>	<i>Variable*</i>									
	<i>EMS</i>	<i>ESR</i>	<i>FC</i>	<i>EROS</i>	<i>HB</i>	<i>ASP</i>	<i>AI</i>	<i>GRIP</i>	<i>GLOB</i>	<i>PAIN</i>
1			30	69						
2	14	22			14	5	45			
3	20	13	21	6		13	27			
4	41	13		5	3	31			6	
5	21	27	8	14	12	10	9			
6	54	13	6	4		13		4		
7	15		9	14	17	13		33		
8	27	32	18	4		15			4	
9	25	50					25			
10	25	13	38		13		11			
11	23	5	16	6	12	8			3	
12	16			6	50			18	10	
13	33	13	8				39		7	
14		14		22	40	6		8		
15	24		8				69		6	
Rank mean weight	1	2	3	4	5	6.5	6.5	8	9	10

*ESR = erythrocyte sedimentation rate; HB = haemoglobin; FC = functional capacity; AI = articular index; ASP = aspirin consumption; EROS = number of erosions on hand X-ray; EMS = early morning stiffness; GLOB = patient's global assessment; GRIP = grip strength; PAIN = pain score

Table 2 Expressed policy weights (%). For abbreviations see Table 1 footnote

<i>Doctor</i>	<i>EMS</i>	<i>ESR</i>	<i>FC</i>	<i>EROS</i>	<i>HB</i>	<i>ASP</i>	<i>AI</i>	<i>GRIP</i>	<i>GLOB</i>	<i>PAIN</i>
1			20	80						
2	9	11	5	17	11	9	17	12	7	7
3	11	14	13	7	6	8	12		11	11
4	2	12	12	5	8	15		8	12	12
5	6	15	7	13	15	11	13	9	6	6
6	16	9	11	5	9	16	9	7	13	7
7	14	9	11	11	7	12	11		14	14
8	5	15	15	15	10	10	5	5	5	10
9	14	11	7	8	5	14	14	11	14	7
10	15	15	15	5	5	15	15	5	15	10
11	9	14	12		12	9		14	12	16
12	12	11	8	11	13	12	6	4	16	11
13	13	17	8	2	13	13	8	8	8	8
14	8	12	10	15	12	8	12	4	10	6
15	12	15	6	10	13	12	12	13	8	5
Rank mean weight	2	1	5	7	6	9	4	10	3	8

deterioration to appear as an improvement.

Trial sample size calculations can be based on the observed proportion of patients exceeding the threshold for clinically important change. Such a change is regarded as a 'success'. Because the proportion of 'successes' seen in the trial varied between one doctor and another, so will the sample size estimate. An estimate is also required of the increase in the proportion of patients showing 'success' which is regarded as important. This is usually expressed as a per-

centage or as an absolute improvement on the currently observed rate. This study was not designed to make such estimates for the population but Table 3 shows calculated sample sizes for an arbitrary 50% increase where the spontaneous improvement is less than 50% and a 50% decrease in 'failures' where the spontaneous rate is greater than 50%. Trial sample sizes calculated on this basis range from 206 to 4134 because of the variation in how doctors judge a successful outcome.

Table 3 Calculated required total sample size for each doctor

Doctor	% spontaneous clinical improvement observed	Sample size required* to detect 50% spontaneous improvement in treatment group
1†		
2	6	4134
3	68	390
4	46	252
5	28	626
6	54	234
7	42	308
8	34	458
9	56	250
10	88	1244
11	76	560
12	34	458
13	66	360
14	50	206
15	36	414

* $\alpha = 0.05$, $\beta = 0.05$, $n_1 = n_2$

†Dr 1 gave no estimates of clinically important change

The consensus policy for the consistent group of doctors can be expressed as:

$$\begin{aligned} \text{VAS (disease severity)} = & -6.07f(\text{ASP}) - 15.25f(\text{ESR}) + 49.86f(\text{Hb}) - 5.79f(\text{EMS}) \\ & - 13.80f(\text{GLOB}) - 36.03f(\text{FC}) - 11.59f(\text{AI}) - 4.64 \end{aligned}$$

and for the randomly selected group as:

$$\begin{aligned} \text{VAS (disease severity)} = & -4.99f(\text{ASP}) - 14.67f(\text{ESR}) + 69.90f(\text{Hb}) - 33.38f(\text{EROS}) \\ & - 5.57f(\text{EMS}) - 7.03f(\text{GLOB}) - 33.22f(\text{FC}) \\ & - 15.32f(\text{AI}) - 0.21, \end{aligned}$$

where $f(\text{variable}) =$

$$\frac{\log(\text{final value of variable} + 1)}{\log(\text{initial value of variable} + 1)}$$

The correlation between predicted judgements derived using these policies was $r_s = 0.96$ ($P <$

0.01). The relative importance of the clinical variables in the models of these two groups is shown in Table 4.

Discussion

Claims that regression analyses necessarily give insight into the doctors' decision processes have been rightly criticised (Anderson & Shanteau, 1977). The models we have constructed by these methods are descriptive in a mathematical but not necessarily in a psychological sense. All that may be claimed strictly is that these doctors judge changes in disease activity as if they attributed the modelled weights to the clinical variables and adopted a linear combination rule. The possibility that they may indeed do so is not of course excluded, but is irrelevant to the aims of this study. Judgement models elicited in this way have been shown to be stable over time (Chaput de Saintonge & Hattersley, 1985; Kirwan & Curry, 1984), and are thus usable as reliable predictors of future judgements. Many of the rheumatologists we studied made inconsistent judgements. While this may simply reflect unfamiliarity with the task format or boredom, it may be that they were basically unable to make consistent decisions. Some degree of inconsistency is a feature of human decision taking from which doctors have never been noted to be immune. A few doctors appeared to be consistent yet their models accounted for less than 80% of the variance. It is possible our relatively simple models could have been refined to achieve a better fit for these respondents. However to do so would have added little to the arguments of this paper and would risk overfitting.

Trialists rarely enjoy the luxury of knowing they can call for assistance on a group of clinicians known to make consistent judgements. Fortunately this does not seem to be necessary. A consensus policy derived by weighting each judge's contribution by his consistency differs little in its

Table 4 Relative importance (% contribution to r^2) of clinical variables in the models of the consistent and random subsets of rheumatologists

Variable	Consistent subset of doctors	Random subset of doctors
Early morning stiffness	27	20
ESR	25	24
Articular index	12	23
Patients' global assessment	12	3
Functional capacity	9	6
Aspirin consumption	8	6
Haemoglobin	7	15
Erosions	—	3

effects between groups of doctors selected for high reliability and a randomly selected group. Although there appears to be some difference between the groups in the weights attached to some of the clinical variables it is not clear that these differences are significantly greater than differences within the groups. Because some of the variables are intercorrelated (as is usual in biological systems) inspecting the weights is a less good measure of difference between the policies than a comparison of the predicted judgements they generate. In the case of the consistent and the random group of doctors the correlation is very high suggesting it is unlikely there is a large difference between them.

Information derived from models of this sort can assist choice of outcome variables in clinical trials. Reliance on doctors' opinions alone could have led to an inappropriate selection of variables for use in the trial. For example, it is quite likely that global assessment would have been measured since the group as a whole considered it relatively important. However, it appears as the second least important variable in the modelled policies for the group. It is clear, both intuitively and from the results of this study, that it is often inappropriate to judge outcome in terms of changes in single variables. Nevertheless this is what many trial designs aim to do. Trials which fail to show significant change in any of the outcome variables are usually reported as failing to detect a treatment effect, a sometimes unrealistic conclusion. Combined modest changes in a number of variables may be considered clinically useful even though each alone might be insignificant. The value of weighted linear combinations of outcome variables as a criterion of success in pragmatic trials has been discussed before (Schwartz *et al.*, 1980). So also has the great difficulty in deciding on the values of the weights which is presumably why this approach has not hitherto been used in practice. In statistical terms the analysis of multiple outcome variables has been approached by Pocock *et al.* (1987) in a way which helps with this problem.

We have used an all possible subsets regression model incorporating a penalty for introducing variables into the equation. This minimises the numbers of predictor variables entering and gives some protection against overfitting. It has the general effect of excluding variables which contribute little. Economy in the use of clinical variables aids planning and improves the efficiency of the actual trial. However it may exaggerate differences between the modelled and the expressed policies since few doctors will admit they completely ignore any item of clinical data. There are several other regression methods which

could be used, though choice between them is, predictably, a matter of judgement. It is of interest that if the values of all changes in the policy equation are set to zero some doctors would still have judged the patient significantly improved. In other words, the intercept term in the regression equation was significantly different from zero. This suggests that the doctor's expectation was of significant deterioration over the year between measurements. It may be speculated that when this was not observed and the patient's condition remained stationary this appeared a useful improvement over the expected outcome. This finding is conditional on the observed improvements which were, in fact, always different from zero and is encouraged by relatively shallow regression slopes which are themselves correlated with positive intercepts.

It is possible that our respondents did not all interpret the visual analogue scales in the same way. Many rheumatologists feel that some of the disabilities in rheumatoid patients are irreversible. 'Greatest possible improvement' would have a different meaning to such doctors than to those who interpreted the extreme value of the scale as a return to complete normality. Our experiment did not distinguish these patterns of response, but this does not affect the results. The relative importance of the clinical variables and the clinical trial numbers remain the same whichever interpretation is used.

An analysis of clinical judgement, of the sort we have described, not only identifies the important clinical variables but also estimates improvement rates within the patient sample. This, in turn, permits calculation of a range of trial sample sizes covering the doctors' requirements. In some situations the number of variables and the range of sizes may be very great and some compromise will be necessary. The results of the present study show considerable variations between doctors in their judgement models. It is doubtful that in the highly consistent subset of doctors reported these can reasonably be attributed to random variation. Similar large variation appears when doctors' opinions about useful clinical benefit are sought (Freedman & Spiegelhalter, 1983).

A consensus policy may be achieved mathematically in the way we describe. However, the consensus is somewhat artificial since the participating doctors continue to operate their individual judgement policies. Fortunately there is some evidence that discussion of individuals' statistical models can actually alter judgement behaviour in ways which unaided discussion may not (Kirwan *et al.*, 1983c; Rohrbach, 1981). Analysis of clinical trials planned in this way will

require the application of the judgement policy (or policies) to classify each case as a success or a failure. Alternatively improvement scores can be calculated for each patient and analysed as a continuous variable in the usual way. It has already been argued that applying an explicit judgement policy (or policies) to patient's outcomes will improve consistency both within and between trial physicians. This, in turn, will improve the power of trials (Bech *et al.*, 1986; Stewart & Joyce, 1988). It is also clear that the use of realistic outcomes modelled in the way we describe must improve the relevance of the results to the real-life situations where they will be used. So this approach has potential advantages over current methods both in terms of sensitivity and realism. For the present however, there is no escaping the observation that rheumatologists differ in ways in which they judge treatment outcomes. Until some consensus is reached trials will always lack power and validity for some. It is likely that patients judge outcome in still different ways, being generally indifferent to changes in laboratory variables. No attempt has yet been made to model their judgements and examine

how they differ from the doctors who prescribe for them. This seems particularly worthwhile in situations where the benefits of treatment are largely subjective. The use of outcome modelling to define significant changes in life quality remains to be explored. Relatives and health care planners also have an interest in treatment outcomes and there is no reason why their expectations should not be modelled in the way we describe. Given human nature, it seems unlikely that full consensus will ever be reached. Nevertheless trials can be planned using group judgement models derived from a representative set of physicians (or patients). Trials designed in this way are more likely to satisfy the requirements of readers than those in which criteria have been imposed arbitrarily (Kirwan *et al.*, 1983c). Suggestions have however been made (Hilden, 1988) to report the results of trials in a way which allows patients' own trade-offs between benefit and complications to be considered.

We thank Professor C. R. B. Joyce and Dr. J. Hilden for their critical assistance, and allowing us to see pre-publication versions of their papers.

References

- Anderson, N. H. & Shanteau, J. (1977). Weak inference with linear models. *Psychol. Bull.*, **84**, 1155-1170.
- Bech, P., Haaber, A., Joyce, C. R. B. & Danish University antidepressant group (1986). Experiments on clinical observation and judgement in the analysis of depression: profiled videotapes and judgement analysis. *Psychol. Med.*, **16**, 873-83.
- Chaput de Saintonge, D. M. (1986). Clinical trials in otitis media—to whom do the results apply? *Proceedings of the 3rd World Conference on Clinical Pharmacology and Therapeutics*, Stockholm.
- Chaput de Saintonge, D. M. & Hattersley, L. A. (1985). Antibiotics for otitis media: can we help doctors agree? *Fam. Pract.*, **2**, 205-12.
- Cohen, J. A. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull.*, **70**, 213-9.
- Einhorn, H. J. & Hogarth, R. M. (1978). Confidence in judgement: persistence of the illusion of validity. *Psychol. Rev.*, **85**, 395-416.
- Freedman, L. S. & Spiegelhalter, D. J. (1983). The assessment of subjective opinion and its use in relation to stopping rules for clinical trials. *The Statistician*, **33**, 153-60.
- Hilden, J. (1987). Reporting clinical trials from the viewpoint of a patients' choice of treatment. *Stat. Med.*, **6**, 745-52.
- Hoepfl, R. T. & Huber, C. P. (1970). A study of self-explicated utility models. *Behav. Sci.*, **15**, 408-14.
- Kirwan, J. R., Chaput de Saintonge, D. M., Joyce, C. R. B. & Currey, H. L. F. (1983a). Clinical judgement in rheumatoid arthritis I. Rheumatologists' opinions and the development of 'paper patients'. *Ann. Rheum. Dis.*, **42**, 644-7.
- Kirwan, J. R., Chaput de Saintonge, D. M., Joyce, C. R. B. & Currey, H. L. F. (1983b). Clinical judgement in rheumatoid arthritis II. Judging 'current disease activity' in clinical practice. *Ann. Rheum. Dis.*, **42**, 648-51.
- Kirwan, J. R., Chaput de Saintonge, D. M., Joyce, C. R. B. & Currey, H. L. F. (1983c). Clinical judgement analysis—practical applications in rheumatoid arthritis. *Br. J. Rheumatol.*, **22** (suppl), 18-23.
- Kirwan, J. R., Chaput de Saintonge, D. M., Joyce, C. R. B., Holmes, J. & Currey, H. L. F. (1986). Inability of rheumatologists to describe their true policies for assessing rheumatoid arthritis. *Ann. Rheum. Dis.*, **45**, 156-61.
- Kirwan, J. R. & Currey, H. L. F. (1984). Clinical judgement in rheumatoid arthritis IV. Rheumatologists' assessments of disease remain stable over long periods. *Ann. Rheum. Dis.*, **43**, 695-7.
- Pocock, S. J., Geller, M. & Tsiatis, A. A. (1987). The analysis of multiple end-points in clinical trials. *Biometrics*, **43**, 487-98.
- Rohrbaugh, J. (1981). Improving the quality of group judgement: Social judgement analysis and the nominal group technique. *Org. Behav. Hum. Perf.*, **28**, 272-88.
- Schwartz, D., Flamant, R. & Lellouch, J. (1980). *Clinical trials*. London: Academic Press.
- Spiegelhalter, D. J. & Freedman, L. S. (1986). A pre-

dictive approach to selecting the size of a clinical trial, based on subjective opinion. *Stat. med.*, **5**, 1–13.

Stewart, T. R. & Joyce, C. R. B. (1988). Increasing the power of clinical trials through judgement an-

alysis. *Med. Decis. Making*, **8**, 33–8.

(Received 19 November 1987,
accepted 6 June 1988)