Psychological Medicine, 1986, **16**, 873–883
Printed in Great Britain

Experiments on clinical observation and judgement in the assessment of depression: profiled videotapes and Judgement Analysis

P. BECH, A. HAABER, C. R. B. JOYCE[1]
AND THE DANISH UNIVERSITY ANTI-DEPRESSANT GROUP (DUAG)[2]

From the Psykiatrisk Afdeling, Frederiksborg Amts Centralsygehus, Hillerød, Denmark

SYNOPSIS Variations within and between observer-judges reduce the accuracy of clinical research. Judgement Analysis allows strategies to be developed and applied which reduce variation in judgement. The prediction that the removal of important sources of error variance by this means would reduce the likelihood of committing a Type 2 Error was supported by the application of Judgement Analysis to assessments by 15 psychiatrists of 92 patients in a clinical trial of 2 antidepressive treatments. The statistical significance of differences between the effect of the treatments on the severity of depression was increased, and significant differences appeared earlier. Ten stimulated patient profiles were also converted into narrative case histories, enacted by experienced psychiatrists or psychologists and videotaped. The participants' judgements of the overall severity of the depression were in good agreement with those they had made on the original cases. Videotapes so prepared help training to reduce variation in observation, just as Judgement Analysis can lead to reductions in the variation of judgement.

INTRODUCTION

In medicine, as in many activities, two main tasks are to observe and to make judgements on the basis of the observations. Observation can often be taken over by apparatus (electrocardiographic, spirometric, haematologic etc.) or standardized in other ways. In psychiatry, however, much information is still collected by direct human observation, whether or not recorded quantitatively with the aid of rating-scales. These observations are then combined to arrive at diagnoses, choice of treatment, assessment of outcome, etc. Often, observations themselves contain a large element of judgement, and the two tasks therefore appear inextricably entwined. In consequence, accuracy and consistency are lost, and the use of the information obtained is impaired. Because of the resulting 'noise', the rate of Type 2 Errors[3] will increase (Stewart *et al.* 1975; Joyce, 1984) and patient treatments will, in consequence, be less accurately chosen. This kind of error is more likely to arise when, as is usual in clinical trials, several investigators are involved. (Even in single centre trials, it is common for more than one investigator to be involved, not always explicitly.)

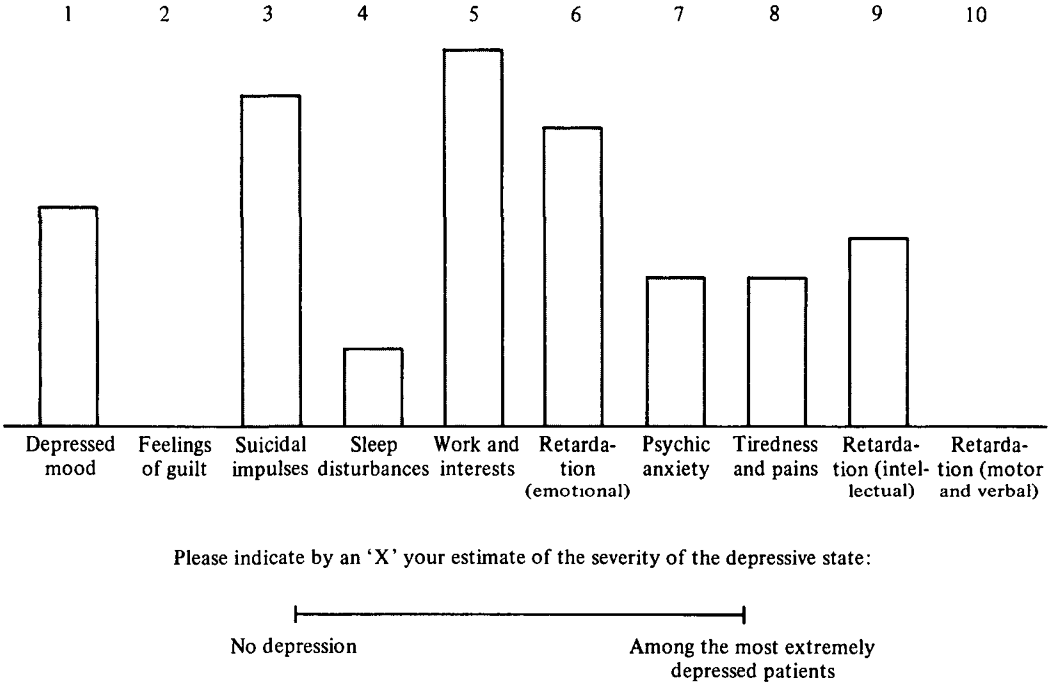
Pre-clinical as well as clinical tests of new drugs are frequently carried out on samples that are too small to allow either Type 1[4] or Type 2 Error to be avoided with a high degree of probability (Freiman *et al.* 1978). The avoidance of Type 1 Error is normally given priority – indeed, the Type 2 Error is seldom stated explicitly (Vere, 1984) – so an unknown, but perhaps high, proportion of useful treatments are incorrectly evaluated because of Type 2

[1] Address for correspondence: Dr C. R. B. Joyce, Project Innovation Group, Medical Department, Ciba-Geigy AG, 4002 Basel, Switzerland.

[2] Members: Drs J. Aagaard, J. Andersen, R. Bang-Olsen, M. Bjerre, S. Bøjholm, P. Christensen, A. Gjerris, L. F. Gram, H. Hansen, W. Hansen, L. Hansted, E. H. Høstrup, E. Jensen, P. W. Jepsen, L. Jørgensen, M. Kastrup, B. Kijne, P. Kragh-Sørensen, C. B. Kristensen, D. Loldrup, K. W. Maarbjerg, O. F. Madsen, E. B. Østergaard, K. R. Pedersen, O. L. Pedersen, O. J. Rafaelsen, S. Rasmussen, N. Reisby, J. Schmid, F. Sevaj, K. Sørensen, E. Thøgersen, H. Y. Thomsen, P. Vestergaard, F. Zierau.

[3] Type 2 Error: failure to detect a true difference between treatments (false negative).

[4] Type 1 Error: acceptance of a difference between treatments as true when it is in fact false (false positive). See Mainland (1963) for a full discussion of these well-known problems.



Please indicate by an 'X' your estimate of the severity of the depressive state:

FIG. 1. Example of simulated case record (MES profile) for Case No. 14.

Errors. It should be noted that although a Type 2 Error may, in theory, cause *either* of the treatments being compared to be dropped, in practice it is more likely that a new treatment will be discarded, whether the comparison is with a placebo or a standard.

As a first step in attacking this problem, by reducing variation due to observation and judgement, it is desirable to separate them, as far as possible, for separate study. There are various means for doing this. Fisch and his colleagues have used a form of multivariate regression analysis (Hammond, 1975) to study the judgement of physicians (Fisch *et al.* 1981, 1982). This makes explicit the bases of such differences between judges (as well as the almost invariably neglected variation between occasions within the same judge), allows them to be resolved, and permits the observations to be judged according to virtually any strategy desired. The purpose of the present study is to demonstrate that the method, by removing variation arising from judgement 'error' in regard to disease severity, allows the size of differences between treatments in a real clinical trial to be more accurately determined.

METHODS AND PROCEDURE

Judgement Analysis (JA) has been fully described by Hammond (1975), Fisch *et al.* (1981) and others. Important differences from these accounts are described below, but the basic procedure consists in presenting the judge with a series (frequently 50) of either real or simulated 'cases', each of which is represented as a set of numerical or pictorial scores on a number (usually 10 or fewer) of items previously agreed to contain most, or all, of the information required to make an appropriate judgement (diagnosis, assessment of progress, choice of treatment, etc.) about the case. The resulting judgements are submitted to a form of multiple regression analysis (JA) to determine the weights applied to each item of information by the individual in reaching a judgement.

Simulated cases

Patients with depressive states were simulated with numbers (generated by a standard method) to represent scores on the items contained in the Melancholia Scale (MES) (Bech & Rafaelsen, 1980; Bech, 1981; Bech *et al.* 1983). The MES consists of 11 items, selected from the Hamilton Depression Scale (HDS) (Hamilton, 1960) and the Cronholm–Ottosson Scale (Cronholm & Ottosson, 1960) for their validity in evaluating the severity of depressive states. In the present study two MES items (motor and verbal retardation) were combined into one, because verbal retardation would have interfered with the preparation and usefulness of the video representations.

Fifty cases were each represented by a sheet containing a set of scores for the 10 MES items as vertical bars, the heights of which were proportional to numbers generated by the computer (Fig. 1 illustrates the display for Case No. 14). Nine were chosen at random for duplication, to provide a check on within-observer reliability. This number of cases and replicates has repeatedly been found adequate to model judgemental policies and to estimate individual reliability. The task of each psychiatrist was to judge the severity of depression on a 10 cm visual analogue scale (VAS), labelled 'No depression' at the left-hand end, and 'Among the most extremely depressed patients' at the right.

Videotape cases

The scores of the duplicate simulated cases plus one other were converted into narrative case histories by one of us (P.B.). One case was enacted by each of 8 experienced psychiatrists and 2 experienced clinical psychologists in such a way as to represent a patient undergoing a clinical interview with P.B. Each recording lasted for 5–8 minutes. A narrative account of the interview for Case No. 14 (compare Fig. 1) is given in the Appendix.

Psychiatrists

Twenty-eight psychiatrists, all members of the Danish University Anti-depressant Group (DUAG) and working in the Departments of Psychiatry of 4 hospitals (Rigshospitalet, Copenhagen; Frederiksborg General Hospital, Hillerød; Odense Hospital, Odense; Aarhus University, Aarhus), were invited to participate in a comparative clinical trial (see below). They are described in Table 1 in terms of years of practice in psychiatry, sex, treatment orientation, attitude to the use of rating scales, experience with the MES, Hamilton Depression Scale (HDS) or Newcastle Scales. Fifteen took part in the clinical trial. The remaining 13 psychiatrists were working in the same 4 departments and had similar professional status, etc., but admitted no patients to the trial. Information about the latter (including their judgemental policies) is presented as evidence of the representativeness of the sample of those participating in the clinical trial. The 'expert' judgements of P.B. and of another senior colleague (Professor O. J. Rafaelsen) were obtained separately in a similar way.

Presentation sessions

On the first occasion the participants, who had previously been contacted informally to secure their cooperation, were brought together in their respective hospitals for about 1 hour. P.B. outlined the intentions and the methods to be used, answered questions and handed out questionnaires about training experience, etc. Each psychiatrist completed a record form to specify the importance which he or she considered each item of the MES should be given when estimating the severity of depressive illness, and when selecting a treatment for depression. Each then received a booklet containing the 50 simulated case profiles in standard order, for assessment at home. Judgements were made at the physician's own preferred pace; no record of time taken was required.

The case information and judgements were entered into the POLICY programme (Hammond *et al.* 1975) to determine the multivariate regression of judgements on the MES items and the 30 judgemental 'policies' (the weights used when each participant combined the information from the 10 items in arriving at a judgement of severity) were extracted.

On the second occasion the videotapes were shown by P.B. and A.H. to the participants of two centres at a time (Copenhagen and Hillerød; Aarhus and Odense). Participants were asked to score each videotape case on the items of the MES, before making the judgement of severity in the same way as they had for the simulated cases.

The clinical trial

Fifteen psychiatrists contributed patients to a double-blind, 35-day between-group comparison of clomipramine (Anafranil® Geigy) and citalopram in the treatment of 92 hospitalized patients

Table 1. *Information on place of work, sex, experience, attitudes of participating psychiatrists*

					Experience with†			Balance of
Doctor	Department	Participant in clinical trial	Sex	Years in psychiatry	MES (HDS)	Newcastle Scales	Attitude to rating scales‡	preference for treatment§
101	Copenhagen	+	M	10	3	3	+	72
102	Copenhagen	+	F	5	3	3	+	82
103	Copenhagen	–	F	8	2	2	–	60
104	Copenhagen	–	M	9	2	2	+	39
105	Odense	+	M	7	3	3	+	76
106	Odense	+	M	6	3	3	+	49
107	Odense	+	M	7	3	3	+	50
108	Odense	–	M	4	2	2	+	53
109	Odense	–	M	4	0	0	+	26
110	Odense	–	M	5	0	0	+	50
111	Hillerød	+	F	4	3	3	+	57
112	Hillerød	+	M	8	3	3	+	67
113	Hillerød	+	M	4	3	3	+	53
114	Hillerød	+	M	5	3	3	+	26
115	Hillerød	+	M	3	3	3	+	77
116	Hillerød	–	F	9	2	0	?	72
117	Hillerød	–	F	4	3	3	+	78
118	Hillerød	–	M	7	2	2	+	50
120	Hillerød	–	M	5	2	2	?	53
131	Hillerød	–	F	5	2	2	–	74
132	Hillerød	–	F	4	3	3	+	50
122	Aarhus	+	M	4	3	0	+	37
124	Aarhus	+	F	5	2	1	+	50
126	Aarhus	–	M	4	1	0	–	50
127	Aarhus	+	M	7	3	1	–	41
128	Aarhus	+	F	5	2	2	+	51
129	Aarhus	–	M	3	2	1	+	50
130	Aarhus	+	M	5	2	0	+	92

† 0 = none; 3 = great.
‡ – = little value; + = high value; ? = doubtful.
§ 0 = psychotherapy only; 100 = pharmacotherapy only.

suffering from depressive disorders. The methodology and results were fully described at the CINP Symposium on 5-HT Re-uptake Inhibitors, held in Florence in June 1984 (Bech, 1984).

The judgement policies previously obtained were applied to the trial observations on the 10 items of the modified MES to obtain the sets of computed judgements described below. The appropriate derived weight was applied to each item score and the products were summed to give the predicted judgement for the case.

RESULTS

Differences between psychiatrists

The place of work, sex, years of experience in psychiatry, experience with and attitude to rating scales and various forms of therapy are summarized for those who did ($N = 15$) and did not ($N = 13$) participate in the trial (Table 1). Participants and non-participants were about equally represented in all centres except Aarhus. There were 9 women and 19 men. Attitudes to rating scales were positive (22/28), and experience with rating scales was greater in those contributing patients to the clinical trial; it was not related to years in psychiatry.

Psychiatrists in Copenhagen had had more (8·0) and in Aarhus (4·7) less than the mean experience (5·5 years). As measured on a 10 cm visual analogue scale (0 = maximal preference for psychotherapy; 100 = maximal preference for pharmacotherapy), participants perhaps had a slightly more positive attitude towards the use of pharmacotherapy than non-participants.

Differences in observations and judgements

Correlations between the original computer-simulated mean total scores on the MES items

Table 2. *Correlations between measures for* 10 *replicate patients*

	VMES	SS	VS
SMES	0·78**	0·73*	0·71*
VMES	—	0·68*	0·98***
SS	—	—	0·62*

SMES: simulated cases – total MES item score (computer generated).
VMES: video cases – mean total MES ($N = 28$ scores).
SS: simulated cases – severity ($N = 28$ judgements).
VS: video cases – severity ($N = 28$ judgements).
* $P < 0{\cdot}05$; ** $P < 0{\cdot}01$; *** $P < 0{\cdot}001$.

(SMES), the mean total MES scores given the 10 video cases by the 28 psychiatrists (VMES), and the corresponding severity judgements for the simulated (SS) and video cases (VS) are presented in Table 2. Although all correlation coefficients were statistically significant, only one (VS/VMES) accounted for more than 61% of the variance. Inter-observer reliability, as estimated by kappa (Bartko & Carpenter, 1976), was low (0·38, 0·57, 0·58), although significant ($P < 0{\cdot}01$), for the 3 variables where such a measure was possible (VMES, SS, VS: excluding the original set of standard HMES scores generated by the computer).

Policy differences within and between judges

The weights which the psychiatrists considered that they usually attached to each of the 10 MES items (the 'specified' weights) are shown in Table 3. The average marks of each judge ($\bar{x}$) on the 10 analogue scales representing the specified weights ranged from 39 to 84 (mean 63·2).

The departures of the weights specified from those expected had the judge given equal weight (i.e. 10%) to each item are expressed by χ^2. The specified weights of only 9 judges departed significantly from equality ($\chi^2 > 16{\cdot}92$, df = 9, $P < 0{\cdot}05$). Only 2 (Nos. 112 and 130) gave a weight of more than 20% to a single cue.

The actual weights obtained from multiple regression analysis of the judgements (Table 4)

Table 3. *Specified relative weights for* 28 *investigators using* 10 *MES cues to severity of depression*

	Cue											
Psychiatrist	1	2	3	4	5	6	7	8	9	10	χ^2	$\bar{x}$
101	13	7	12	8	8	13	13	4	13	10	9·3	63·1
102	14	16	13	4	4	15	8	3	8	15	24·0**	59·1
103	15	14	8	14	15	8	8	4	7	7	14·8	65·4
104	17	18	11	11	7	6	3	5	6	16	26·6**	44·4
105	12	12	6	6	6	12	12	12	12	12	7·6	76·3
106	13	9	9	5	5	16	5	12	9	17	17·6*	52·2
107	11	11	9	9	10	10	8	8	11	11	1·4	79·9
108	13	15	7	10	9	10	9	10	7	9	5·5	54·0
109	7	13	6	10	5	13	13	10	5	10	10·2	48·2
110	9	17	4	9	6	16	3	5	16	16	28·5***	45·5
111	15	8	15	8	4	13	6	9	13	9	13·0	61·3
112	23	7	19	4	19	6	6	6	5	6	46·5***	39·3
113	13	13	13	9	7	9	10	8	6	12	6·2	63·5
114	6	12	11	11	12	13	3	3	12	16	17·3*	63·6
115	12	11	11	11	7	12	7	5	12	11	5·9	79·6
116	12	12	13	7	8	11	7	8	9	13	5·4	67·6
117	17	11	10	11	14	7	4	9	10	7	12·2	40·9
118	8	12	12	12	8	10	10	7	10	11	3·0	81·7
120	14	11	11	7	4	12	8	7	12	12	8·8	66·6
122	13	9	13	5	11	11	6	7	12	12	7·9	73·0
124	13	11	13	6	12	12	10	4	6	13	10·4	71·4
126	14	14	14	7	14	14	7	0	0	14	21·4*	68·7
127	12	12	12	11	10	10	7	8	8	12	3·4	80·1
128	12	10	10	10	12	12	6	6	10	12	4·8	82·1
129	19	10	15	5	5	16	2	5	7	17	33·9***	43·1
130	23	1	11	11	15	4	15	15	1	5	46·9***	40·0
131	11	11	12	9	9	12	10	7	10	10	2·1	83·6
132	12	12	12	10	9	10	5	8	11	12	4·7	75·5
											Mean =	63·2

χ^2 = S(O-E)^2, where O = observed weight, E = expected weight (10 for all cues).
$\bar{x}$ = Mean of original (untransformed) score on 100 mm scale.
* $P < 0{\cdot}05$; ** $P < 0{\cdot}01$; *** $P < 0{\cdot}001$.

Table 4. *Actual relative β-weights* (× 10^{-2}), *multiple regression coefficients* (*R*) *and correlations between replicates* (*r*)

	Cue												
Psychiatrist	1	2	3	4	5	6	7	8	9	10	χ^2	*R*	*r*
101	22	12	26	3	−1	12	6	9	6	−3	78·0	0·88	0·99
102	20	12	17	11	4	7	−5	5	8	11	45·4	0·82	0·96
103	25	19	17	13	2	−2	−3	4	5	10	80·2	0·90	0·99
104	17	9	27	31	12	−1	3	0	1	0	123·5	0·88	0·98
105	9	10	11	9	10	10	11	10	11	10	0·5	1·00	1·00
106	24	14	16	5	9	6	3	−1	12	11	46·5	0·94	0·99
107	12	14	23	8	6	13	4	10	6	3	31·9	0·85	0·99
108	14	18	10	17	−5	10	7	−2	−8	−9	119·2	0·81	0·95
109	7	19	30	2	−8	8	7	1	−2	16	115·2	0·86	0·96
110	17	19	8	8	−3	25	1	9	3	−8	98·7	0·83	0·99
111	40	5	11	−4	7	2	0	10	9	11	129·7	0·95	0 99
112	37	7	11	−3	16	11	5	0	4	6	112·2	0·95	0·99
113	15	17	24	8	11	9	6	3	5	−1	48·7	0·94	0·99
114	13	11	5	4	4	5	−3	6	11	38	110·2	0 71	0 85
115	43	19	−3	0	−9	12	−1	0	−4	−10	262·1	0·86	0·98
116	25	1	17	12	1	3	0	10	7	23	76·7	0·82	0·98
117	28	10	4	2	27	10	13	4	0	3	90·7	0·91	0 97
118	20	29	−1	16	4	9	10	−6	3	−2	110·4	0·85	0·98
120	32	11	25	10	4	−5	−2	0	3	7	145·3	0·94	0·97
122	36	12	21	−6	5	1	2	2	11	5	131·7	0·93	0·99
124	22	26	20	1	3	−1	−3	8	9	7	105·4	0·83	0·97
126	42	19	3	−1	−1	9	2	9	3	10	151·1	0·88	0 94
127	43	12	20	2	6	2	0	2	8	−6	176·1	0·95	0·98
128	22	12	19	9	8	19	3	2	3	2	54·1	0·92	0·98
129	50	2	15	−6	5	9	3	−6	−1	−2	254·1	0 93	0·98
130	55	3	−6	−2	−7	9	0	−3	−9	−7	368·3	0·84	0·83
131	11	12	43	−2	1	−4	9	−1	8	8	164·5	0·86	0·98
132	32	17	13	12	4	3	1	6	11	−1	85·0	0·91	0·99

χ^2 = see Table 3. $P < 0{\cdot}001$ for all except No. 105.

showed considerable variation within as well as between judges. The squared multiple regression coefficients (R^2) estimate the consistency with which each applied his or her judgemental policy. An *R* of 0·8 or higher ($R^2 \geqslant 0{\cdot}64$) indicates that at least 64% of the judgemental variation was satisfactorily accounted for by the linear and additive regression model. The correlation between replicates (*r*) suggests the extent to which inconsistency within the observer may be responsible for such deficiencies as are observed in the model. Nos. 114 and 130 differed to the most marked degree from equality. Weights of 20 or more were awarded on 35 occasions, 25 judges doing so at least once. The average mean rank order of the actual β-weights correlated significantly with that of the specified weights (rho = +0·703; $0{\cdot}05 > P > 0{\cdot}01$).

Items 1–3 (see Fig. 1) were most heavily weighted by almost half the judges, although not necessarily in that order; 6 other items were judges of first or second importance by at least one psychiatrist. Items 7, 8 and 9 (Fig. 1) only once received a weight of as much as 13%, a value that can probably be attributed to chance. All negative weights (which never exceeded 10%) can also be neglected, since the scoring of each item was arranged to show a positive association with depression. *R* and *r* were, with one exception, high (> 0·8), indicating that the rectilinear regression probably accounted satisfactorily for most of the variance in the individual judgements. The relatively low replicate reliability ($r = 0{\cdot}85$) of No. 114 suggested that the poor fit of the model ($R = 0{\cdot}71$, $R^2 = 0{\cdot}504$) was due to some unreliability of judgement.

The clinical trial

The conventional results of the clinical trial have previously been presented in terms of the Hamilton Depression Rating Scale (HDRS) (Bech, 1984). In the present study, investigators'

Table 5. *Severity of endogenous and non-endogenous depression under treatment with citalopram (C) and clomipramine (A) using alternative judgemental policies (see text for explanation)*

	Endogenous								Non-endogenous								All patients							
	Equal		Appro-priate		Expert		Atypical		Equal		Appro-priate		Expert		Atypical		Equal		Appro-priate		Expert		Atypical	
Week	C	A	C	A	C	A	C	A	C	A	C	A	C	A	C	A	C	A	C	A	C	A	C	A
0	23	22	197	202	186	182	194	194	20	21	157	184	161	176	179	185	22	22	188	197	180	180	190	191
1	19	20	156	181	147	159	160	148	19	16	138	129	148	126	163	125	19	19	152	165	147	149	161	141
2	17	16	141	142	131	126	145	115*	16	12	105	99	108	88	120	94	17	14	132	129	126	114	139	109**
3	15	11*	125	97*	115	80*	132	79**	14	11	100	103	95	96	91	88	15	11*	119	99*	110	85	122	82**
4	13	10*	106	84	94	74	115	72**	14	9	116	73	99	60	91	68	13	9*	109	81*	96	69	109	71**
5	11	10	98	89	94	79	96	82	14	9	103	80	91	63	83	75	12	10	99	86	93	74	93	80
N	35	32	35	32	35	32	35	32	11	14	11	14	11	14	11	14	46	46	46	46	46	46	46	46

* $P < 0.05$; ** $P < 0.01$.

policies were obtained from their weighting of the MES items, and the description of the results that follows also differs in some other respects from that of Bech (1984).

Results from 46 patients on each treatment were available for this analysis. The two groups were broadly comparable in sex ratio, age distribution, duration of illness and current episode, and diagnosis by the Newcastle 1965 and 1971 Scales and DSM-III. Sixty-five and 67 respectively were diagnosed with the aid of the two Newcastle Scales as suffering from endogenous and either 27 or 25 from non-endogenous depression. The number of patients seen by the 15 participating physicians varied from 1 to 17, with a median of 6. Such a distribution is by no means atypical for trials with many investigators.

Application of different judgemental policies to clinical trial observations

Reasons for the choice of the actual judgement policies applied to the MES scores of the 92 patients in the clinical trial are considered in the Discussion. Four were chosen (Table 5): first, the equal weights policy (which left the raw scores unchanged); secondly, the policy of the appropriate psychiatrist – i.e. the one who had seen the patient on the occasion in question; thirdly, an 'expert' policy (see section on 'Psychiatrists', above); fourthly, that of a single representative of the group (No. 104), atypical in that he gave a particularly high weight to cue 4 (the sleep item of the MES). These are referred to in the column headings of Table 5 as 'Equal', 'Appropriate', 'Expert' and 'Atypical', respectively.

For the 'Equal' weight policy, statistically significant differences between treatments ($P < 0.05$) were apparent at weeks 3 and 4 for the endogenous group as well as for the patients as a whole. No significant differences appeared at any time in the smaller group with the diagnosis of non-endogenous depression.

There are striking contrasts between the results of applying the other policies. None gave rise to significant differences in the non-endogenous group; but, whereas the 'Appropriate' policy scored one fewer significant difference than that of 'Equal' (endogenous at week 4), application of the 'Expert' policy lost 3 of the 4 (endogenous at week 4 and all patients at weeks 3 and 4). However, not only did the

'Atypical' policy give rise to 2 additional significant differences (at week 2 in the endogenous and all patient groups), but 5 of the 6 differences were significant at $P < 0{\cdot}01$, a level not attained by any policy on any occasion. Thus, with the use of this policy, differences between treatments appeared earlier and reached greater levels of statistical significance.

DISCUSSION

An assumption almost always made in clinical trials involving more than one investigator is that the judgements of all investigators, or at least of those within a given centre, can be treated as equivalent, whether or not inter-observer agreement has been measured and whether or not this turns out to be 'satisfactory'. If the degree of agreement is considered to be 'unsatisfactory', training to reduce the differences may be carried out. But, even if measures of inter-observer agreement are made (e.g. in terms of the kappa statistic), they do not lend themselves to improvements in the sensitivity of the trial.

Judgement Analysis, however, makes explicit the *reasons* for the differences observed, and the resulting information may be used in several ways: first, knowledge and use of actual individual policies enables their owners to carry out their intentions more accurately – the computed policy can be adjusted if necessary and then applied with complete consistency; secondly, shared knowledge of policy differences can accelerate mutual agreement and the development of consensus; thirdly, a 'best' policy, derived from consensus, authority or in some other way can be applied to the judgements of all.

If investigators in a clinical trial, especially those working in different centres, do not even observe the same patients, it is surely even more important to bring their judgements into a common frame of reference. By this means, within- and between-observer variation can be removed from the error variance. The power of the design is thereby increased; and, in consequence, the likelihood of committing a Type 2 Error (Stewart *et al.* 1975; Joyce, 1984) will be decreased. Alternatively, the sample size can be reduced – a choice that is always defensible on ethical (Montgomery & Åsberg, 1979) as well as usually on economic grounds.

The present study was intended to test the prediction that the application of one or more judgement policies, rendered consistent by the removal of unreliability, to the assessment made in a real-life clinical trial would improve its specificity. Any, or all, or some combination, of the policies available might have been used. One psychiatrist used an equal weights policy; many others resembled each other. There would have been no point in testing all of them. In the event, it seemed rational to apply the policy of the 'Appropriate' psychiatrists to the observations on the patients for whom they were responsible. An 'Expert' policy was also chosen, for obvious reasons. The 'Atypical' policy was chosen because it gave a particularly high weight to an MES item otherwise seldom selected; except for several policies that, like the 'Expert' policy, gave very high weights to one or other of the first three cues, it departed most widely from one of equal weights.

Although none of the three examples tested affected the judgements of severity in patients with non-endogenous depression to the point where the differences between treatments became statistically significant, the 'Atypical' policy improved the discrimination between treatments in most of the time/diagnosis cells for endogenous patients and for the group as a whole (Table 5), usually increasing the level of statistical significance. On the other hand, the use of the 'Appropriate' and 'Expert' policies reduced the discriminations below those achieved by the use of 'Equal' weights. Untransformed (i.e. equally weighted) scores are frequently used by those who employ rating scales, although this is not the case in, for example, the Hamilton Depression Scale (HDS) where the range of possible scores is not the same for all items. However, even if the weights are not entirely arbitrary, they are seldom submitted to sensitivity analysis. Judgement Analysis in effect does this and so provides a rational or at least an explicit basis for weight selection.

As the use of Judgement Analysis to decrease error variance increases the sensitivity of the trial design, another obvious explanation for the failure of the 'Atypical' policy to establish any significant treatment difference in the non-endogenous group is that there was no real difference between the two treatments for such patients. Both drugs in the trial acted on endogenous and non-endogenous depression, clomipramine producing a significantly faster change than citalopram on total HDS score (Bech, 1984). However, when the sleep item was omitted, as in the Bech *et al.* (1975) sub-scale of the HDS, the difference between the two treatment groups became less pronounced. (The group of 5-HT uptake inhibitors, to which citalopram belongs, seems in general to have little effect on sleep disturbances during the early weeks of treatment (d'Elia *et al.* 1981).) Although other evidence suggests that clomipramine is effective in non-endogenous depression (Rack, 1973), no studies with citalopram have been published so far, so the choice between these explanations is not clear.

Surprisingly, neither the 'Appropriate' nor the 'Expert' policies improved on 'Equal' weights. The 'Appropriate' policy was really less appropriate than its title indicates: in spite of considerable efforts to ensure that individual patients were seen by the same psychiatrist, this was often not possible in practice. On such occasions, the policy of the intended rather than of the actual psychiatrist was used in the analysis. The relative failure of the 'Expert' policy is harder to explain, given the success of the 'Atypical' policy. However, the failure of the 'Expert' policy was only relative. It led to a better discrimination of the treatments than did the 'Appropriate' policy on 12 of the 15 possible comparisons, 4 occurring in the non-endogenous group.

Other procedures would have been possible. For example, policies based on mean or modal weights could have been used. These would probably have given rise to policies similar to those based on equal weights, and therefore results resembling those obtained in the traditional way. A number of studies (e.g. Deane *et al.* 1972; Kirwan *et al.* 1983*c*) suggest that the optimal procedure is to use the 'captured' individual policies to provide feed-back to the participants about their judgemental processes and so facilitate the development of an informed consensual policy. This was not attempted, but is planned for a future study.

Progress can also be made in the improvement of observation by using standardized simulated video performances (Russell *et al.* 1983); for example, by having the MES scores of 10 patients enacted and videotaped. This establishes control over the observational, as distinct from the judgemental, task by creating external criteria; and its use therefore deserves closer study in psychiatric training. No doubt because the scores were, in effect, twice transformed (by the actors and the members of the audience), the resulting judgements of severity did not agree very closely with those on the cases to which they corresponded ($r = 0{\cdot}62$, $P < 0{\cdot}05$: Table 2); the ratings on the 10 individual scale items were in better agreement ($r = 0{\cdot}78$, $P < 0{\cdot}01$: Table 2). Customary methods (e.g. Tiplady & Loudon, 1980) do not separate observation and judgement and so do not make explicit the causes of disagreement.

The results of analysis of actual judgements were very similar, both in the general distribution of weights and as to the levels of consistency and reliability achieved, to those found by Fisch *et al.* (1981, 1982) with general physicians and psychiatrists in Switzerland, by Joyce *et al.* (1977) and by Kirwan and his colleagues with rheumatologists in the United Kingdom (Kirwan *et al.* 1983*a*, *b*, *c*, 1984). Although, as expected, there were wide differences in the policies of individual judges, as well as between the policies they believed themselves to be using and the actual policies elicited from their judgements about simulated patients, the rectilinear multiple regression model fitted the actual policies of most individuals very closely. The equal weights model did not (with the exception of that of No. 105), although, when asked to specify their policies, the majority of judges also expressed weights that tended to equality. This is noteworthy, because the assumption is very often tacitly made in the construction of rating scales that all items contribute equally to the total. In clinical trials, too, the global evaluation of treatment, if defined at all, often seems to represent the total of individual items that are equally weighted, but not explicitly so.

The disparity between specified and actual policies illuminates current conflicts of opinion about rating scales in clinical psychiatry. Asked to estimate the weights they attributed to the items of the MES, most psychiatrists (19/28) specified a policy of weights that were close to equal (i.e. corresponding to traditional rating-scale practice); their behaviour in a task corresponding more closely to clinical practice, however, was quite different. Here, the use of an equal weights policy was exceptional (1/28).

The high average mark ($\bar{x}$: Table 3) placed on the 10 analogue scales during weight specification presumably indicated that the items in the MES scale were considered by most psychiatrists to be relevant to the assessment of severity of depression – a low mean weight would have suggested that some items at least were considered to be irrelevant. It seems that giving the scale items of the MES equal weights 'feels' unsatisfactory. The feeling may be correct; it is evidently possible to obtain a more specific measure of the severity of depression than that given by traditional scoring by weighting the items according to other considerations. In the absence of an *external* criterion for assessment, it may be objected that it is not possible to decide which method has the highest validity. This situation is, of course, typical in psychiatry, where, as elsewhere, the type of validity in question often also lacks definition. In the present case, however, it has been demonstrated that the application of Judgement Analysis to a clinical trial can improve the power of the trial design to discriminate between treatments. On the other hand, the use of trials to study clinical judgement provides a rational criterion for choosing between alternative ways of estimating the severity of illness.

CONCLUSIONS

Twenty-eight Danish psychiatrists, 15 of whom were participants in a subsequent clinical trial, first specified the relative importance they attached to each item of the Melancholia Scale (MES) in assessing the severity of depression. Their actual behaviour in judging depression was examined by multiple regression analysis of their evaluation of 50 simulated cases. Large differences were found between individuals' specified and actual judgemental policies, and between the actual policies of individuals, although most of the latter were applied with high consistency.

Such variations impair the accuracy of a clinical trial. Judgement Analysis allows a policy to be developed that can be applied to all judgements. The prediction that the removal of important sources of error variance by this means would reduce the likelihood of committing a Type 2 Error was supported by the application of actual policies to observations made in a clinical trial of 2 antidepressive treatments, in which the significance of differences in their effects on the severity of depression was increased and appeared earlier.

The MES profiles of 10 cases were also converted into narrative case histories enacted by experienced psychiatrists or psychologists and videotaped. The participants scored the items of the MES scale for these videocases and judged the overall severity of the depression. These judgements were in good agreement with those for the original patients. Thus, patients with quantitatively defined characteristics were satisfactorily simulated by this method. Videotapes so prepared can help to reduce variation in observation, just as Judgement Analysis can lead to reductions in the variation of judgement.

Kurt Jacobsen, Jan Trøst and Bodil Winter of Ciba-Geigy (Denmark) gave invaluable administrative and clerical assistance. Statistical advice and/or computing assistance was kindly given by Max Stubgaard (Frederiksborg Amt Centralsygehus, Hillerød), Peter Allerup (Danish Institute of Educational Research), Ove Aaskoven (Lundbeck) and Claus Bech. Teresa Harnisch, Sonja T. Pedersen and Rosemary Szente efficiently and patiently typed the several drafts. Several colleagues, notably Hans-Ueli Fisch, provided valuable criticism.

APPENDIX

Narrative for Simulated Case No. 14

The patient was hospitalized a week ago after an attempted suicide. The interviewer begins by focusing on his experience of the previous 3 days.

The patient communicates a lowered mood, has no interest in usual activities, feels sad and miserable, nothing in the surroundings can excite him. During the interview he gives non-verbal signs of depression, is expressionless, 'greyish'. There are still suicidal impulses. There is no feeling of guilt. During the last 3 nights there has been some difficulty in falling asleep (1–1½ hours) but the sleep length is only slightly reduced. In the ward the patient has not been able to participate in the daily activities. He shows no need or ability to contact other patients or ward personnel, even the person especially charged with watching the patient due to the risk of suicide. He feels emotionally indifferent, even to near friends or his family. He experiences an unpleasant state of anxiety or fear without apparent reason, a vague and uneasy feeling which he can, however, just control. Hence, there have been no panic attacks.

The patient is unable to relax, feels tensed up and becomes tired easily. He has less energy than usual and there are sometimes painful sensations in his back, but no real pains. He has difficulties in concentrating; it is more difficult for him than usual to read a book or even a newspaper. During the interview, however, there are no signs of difficulties in concentration.

REFERENCES

Bartko, J. J. & Carpenter, W. T. (1976). On the methods and theory of reliability. *Journal of Nervous and Mental Disease* **163**, 307–317

Bech, P. (1981). Rating scales for affective disorders: their validity and consistency. *Acta Psychiatrica Scandinavica* **64**, Suppl. 295, 1–101.

Bech, P. (1984). Citalopram versus clomipramine: a controlled clinical study. *Clinical Neuropharmacology* **7**, Suppl. 1, 5471–5472.

Bech, P. & Rafaelson, O. J. (1980). The use of rating scales exemplified by a comparison of the Hamilton and the Bech–Rafaelsen Melancholia Scale. *Acta Psychiatrica Scandinavica* **62**, Suppl. 285, 128–132.

Bech, P., Gram, L. F., Dein, E., Jacobsen, O., Vitger, J. & Bolwig, T. G. (1975). Quantitative rating of depressive states. *Acta Psychiatrica Scandinavica* **51**, 161–170

Bech, P., Gjerris, A., Andersen, J., Bøjholm, S., Kramp, P., Bolwig, T. G., Kastrup, M., Clemmesen, L. & Rafaelsen, O. J. (1983). The Melancholia Scale and the Newcastle Scales. Item-combinations and inter-observer reliability. *British Journal of Psychiatry* **143**, 58–63.

Cronholm, B. & Ottosson, J.-O. (1960). Experimental studies of the therapeutic action of electroconvulsive therapy in endogenous depression. The role of the electrical stimulation and of the seizure studied by variation of stimulus and modification by lidocaine of seizure discharge. *Acta Psychiatrica Neurologica Scandinavica* Suppl. **145**, 69–97.

Deane, D. H., Hammond, K. R. & Summers, D. A. (1972). Acquisition and application of knowledge in complex inference tasks. *Journal of Experimental Psychology* **92**, 20–26.

d'Elia, G., Hällström, T., Nyström, C. & Ottosson, J.-O. (1981) Zimelidine versus maprotiline in depressed outpatients. A preliminary report. *Acta Psychiatrica Scandinavica* **63**, Suppl. 290. 225–235.

Fisch, H.-U., Hammond, K. R., Joyce, C. R. B. & O'Reilly, M. (1981). An experimental study of the clinical judgment of general physicians in evaluating and prescribing for depression. *British Journal of Psychiatry* **138**, 100–109.

Fisch, H.-U., Hammond, K. R. & Joyce, C. R. B. (1982). On evaluating the severity of depression: an experimental study of psychiatrists. *British Journal of Psychiatry* **140**, 378–383.

Freiman, J. A., Chalmers, T. C., Smith, H. J. & Kuebler, R. R. (1978). The importance of β, the type II error and sample size in the design and interpretation of the randomized control trial: survey of 71 'negative' trials. *New England Journal of Medicine* **299**, 690–694.

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery and Psychiatry* **23**, 56–62.

Hammond, K. R. (1975). Social Judgment theory: its use in the study of psychoactive drugs. In *Psychoactive Drugs and Social Judgment* (ed. K. R. Hammond and C. R. B Joyce), pp. 69–105. Wiley Interscience: New York.

Hammond, K. R., Stewart, T. R., Brehmer, B. & Steinmann, D. O. (1975). Social Judgment Theory. In *Human Judgment and Decision Processes* (ed. M. F. Kaplan and S. Schwartz), pp. 271–312. Academic Press: New York.

Joyce, C. R. B. (1984). Judgment analysis and clinical trials. How to save the baby from the bathwater. *Controlled Clinical Trials* **5**, 307–308.

Joyce, C. R. B., Berry, H., Chaput De Saintonge, D. M., Domenet, J., Fowler, P. & Mason, R. M. (1977). Judgment analysis of investigators' assessments: a way to reduce one important source of error in multi-centre trials. In *International Co-ordination of Drug Trials* (ed. K. Fehr, E. C. Huskisson and E. Wilhelmi). Eular Bulletin Monograph No. 1.

Kirwan, J. R., Chaput de Saintonge, D. M., Joyce, C. R. B. & Currey, H. L. F. (1983*a*). Clinical judgement in rheumatoid arthritis. I. Rheumatologists' opinions and the development of 'paper patients'. *Annals of the Rheumatic Disease* **42**, 644–647.

Kirwan, J. R., Chaput de Saintonge, D. M., Joyce, C. R. B. & Currey, H. L. F. (1983*b*). Clinical judgement in rheumatoid arthritis. II. Judging 'current disease activity' in clinical practice. *Annals of the Rheumatic Diseases* **42**, 648–651.

Kirwan, J. R., Chaput de Saintonge, D. M., Joyce, C. R. B. & Currey, H. L. F. (1983*c*). Clinical judgment analysis – practical application in rheumatoid arthritis. *British Journal of Rheumatology* **22**, (Suppl.), 18–23.

Kirwan, J. R., Chaput de Saintonge, D. M., Joyce, C. R. B. & Currey, H. L. F. (1984). Clinical judgement in rheumatoid arthritis. III. British rheumatologists' judgments of 'change in response to therapy'. *Annals of the Rheumatic Diseases* **43**, 686–694.

Mainland, D. (1963). *Elementary Medical Statistics.* W. B. Saunders: Philadelphia.

Montgomery, S. A. & Åsberg, M. (1979). A new depression scale designed to be sensitive to change. *British Journal of Psychiatry* **134**, 382–389.

Rack, P. H. (1973). A comparative clinical trial of oral clomipramine (Anafranil) against imipramine. *Journal of International Medical Research* **1**, 332–337.

Russell, M. L., Ghee, K. L., Probstfield, J. L. & Insull, W. (1983). Development of standardized simulated patients for quality control of the clinical interview. *Controlled Clinical Trials* **4**, 197–208.

Stewart, T. R., Joyce, C. R. B. & Lindell, M. K. (1975). New analyses: application of judgment theory to physicians' judgments of drug effects. In *Psychoactive Drugs and Social Judgment* (ed. K. R. Hammond and C. R. B. Joyce), pp. 249–262. Wiley Interscience: New York.

Tiplady, B. & Loudon, J. B. (1980). Assessment of depression and the non-psychiatrist. In *Abstracts of the 12th CINP Congress*, 340. Pergamon: Oxford.

Vere, D. W. (1984). Logic and valid inference. In *Current Problems in Clinical Trials* (ed. D. M. Chaput de Saintonge and D. W. Vere), pp. 35–40. Blackwell: Oxford.