STATISTICS IN MEDICINE, VOL. 15, 1421-1434 (1996)

AN INDEX FOR ASSESSING BLINDNESS IN A MULTI-CENTRE CLINICAL TRIAL: DISULFIRAM FOR ALCOHOL CESSATION—A VA COOPERATIVE STUDY

KENNETH E. JAMES

Health Services Research and Development (152), VA Medical Center, Oregon Health Sciences University, 3710 SW US Veterans Hospital Road, Portland, OR 97201, U.S.A.

DANIEL A. BLOCH

Department of Health Research and Policy, Stanford University Medical School, Stanford, CA 94305, U.S.A

KELVIN K. LEE

Cooperative Studies Program Coordinating Center (151K), VA Medical Center, 3801 Miranda Avenue, Palo Alto, CA 94304, U.S.A.

HELENA C. KRAEMER

Department of Psychiatry and Behavioral Sciences, Stanford University Medical School, Stanford, CA 94305, U.S.A.

AND

RICHARD K. FULLER

Department of Health and Human Services, National Institute of Alcohol Abuse and Alcoholism, 6000 Executive Boulevard, Rockville, MD 20892, U.S.A.

SUMMARY

This paper considers an index to assess the success of blinding with application to a clinical trial of disulfiram. The index increases as the success of blinding increases, accounts for uncertain responses, and is scaled to an interval of 0·0 to 1·0, 0·0 being complete lack of blinding and 1·0 being complete blinding.

INTRODUCTION

One of the accepted standards for the conduct of randomized clinical trails is that they be performed double-blind, that is, neither subjects nor researchers and clinicians potentially influencing or evaluating subjects' responses to treatment know to which treatment group the subject has been randomly assigned. Only in this case can one have certainty that any differential effects between groups stem from the treatment rather than the subjects' or researchers' biases. Such blindness, however, is difficult even to attempt, especially in studies where the treatment is obvious, such as those that compare surgery versus medical treatments or different modes of

*Supported by the Department of Veterans Affairs Cooperative Studies Program.

CCC 0277-6715/96/131421-14

Received April 1995
Revised June 1995

health care. When attempted, such as in drug comparison studies, either different side-effects of the treatments or simply slips in communication can compromise the blindness. Byington *et al.*[1] reported that in the Beta Blocker Heart Attack Study trial, 80 per cent of the patients who received propranolol correctly identified their treatment group assignment and 57 per cent of the patients who received placebo incorrectly guessed that they were receiving propranolol. Clinic personnel correctly identified the group assignment an even higher proportion of the time. Heart rate and heart rate change seemed to be major factors in this identification. Moscucci *et al.*[2] reported the assessment of patient blinding in a double-blind trial of phenylpropanolamine (PPA) versus placebo in mild obesity. They found that 74 per cent of the placebo patients and 43 per cent of the PPA patients guessed their treatment correctly. Appetite control was the most frequently reported basis for guessing PPA, even by placebo patients. Lack of adverse reactions was the most frequently reported basis for guessing placebo, also by PPA patients. The results of these and similar trials suggest that in double-blind studies, differences in outcome or incidence of adverse drug reactions can act as unblinding factors. Assessment of study blindness has generally used data pooled across the hospitals and ignored the 'don't know' or 'uncertain' responses. Such assessments usually describe the proportion of correct reports by actual treatment, which includes both correct guesses as well as those known one way or another by the reporter. Howard *et al.*[3] evaluated the study blindness in the Aspirin for Myocardial Infarction trial by computing a proportion that ignored the 'don't knows' and extraneous responses in determining the numerator but included such responses in the denominator, thereby computing the proportion of informed guesses. Yet the 'don't know' responses, if honestly reported, are the strongest indicator of success of the blinding procedures. Hughes and Krahn[4] assessed the success of blindness in a study that compared the effect of placebo versus nicotine on withdrawal symptoms for subjects who stopped smoking. They compared the proportion of patients who correctly identified the drug they were taking to the proportion who did not and assumed that the study was unblinded if the first proportion significantly exceeded the second. They also analysed the drug effect across the patients who correclty identified, incorrectly identified or were uncertain of the medication they received and found no difference across the three groups. Blindness has been considered by several other authors,[5-8] but there has been no previous discussion of the construction of an index to assess the success of blinding.

This paper describes the construction of an index of blindness, which takes into account the 'don't know' responses, and assesses the blindness of clinic personnel to treatments administered in a blinded clinical trial. We examine the effect of the 'don't knows' and the homogeneity of correct guesses across participating centres. We apply the analytic technique to data from VA Cooperative Study Number 107, Disulfiram (Antabuse) in the Treatment of Alcoholism, the primary results of which have been previously reported.[9]

METHODS

Conduct of the Trial

During the period July 1979 to September 1983 the VA Cooperative Studies Program supported the conduct of Cooperative Study Number 107, a clinical study involving 605 patients in nine Veterans Affairs medical centres to test the efficacy of the drug disulfiram in helping alcoholic patients to stop drinking. Patients were randomly assigned with equal probability to one of three treatments: 250 mg disulfiram (202 patients); 1 mg disulfiram (204 patients); or riboflavin (a vitamin with an inert marker to test compliance) (199 patients), to be taken during the one year study period. The primary endpoints related to alcohol cessation: complete abstinence during the study period; time to first drink, and the total number of drinking days during the year. The rationale and details of the study design appear elsewhere.[10]

At the time of randomization, the participating investigator opened an envelope that contained one of two drug assignments: 'disulfiram' or 'no disulfiram'. If the assignment was to 'disulfiram', the envelope did not reveal the disulfiram dosage. The participating investigator communicated the randomization outcome to the patient and destroyed the envelope to prevent unblinding of the study co-ordinator or the alcoholic treatment programme therapist. Thus, subjects were not 'blinded'. The study was designed in this manner so as to measure the psychological effect of disulfiram.

Study co-ordinators, hired by the study, were responsible for the conduct of Cooperative Study Number 107 at their respective participating hospitals, including patient recruitment, assessment and follow-up. They communicated with study patients regularly and collected data related to abstinence from alcohol and compliance in taking study medication. They had, however, no clinical assignments. They were aware that the study medications consisted of one of two doses of disulfiram (1 mg or 250 mg) or a dose of riboflavin but they were 'blinded' to the medication that the patient received. All clinical responsibilities were assumed by one or more programme therapists.

Programme therapists dispensed the study medication, which was identical in appearance for the three treatments, and provided counselling and treatment as prescribed by the alcoholic treatment programme at their hospital. They knew that the study design called for the administration of disulfiram and riboflavin but they did not know that there were two doses of disulfiram. They were also 'blinded' to the medication that the patient received. Patients were asked not to discuss the randomization outcome with anyone, especially not the study co-ordinator or the programme therapist.

The effectiveness of the blinding procedures for the study co-ordinator and programme therapist was evaluated after each patient had completed the one year follow-up. Each respondent was asked to guess which of the study medications the patient took during that year, and, in the case of the study co-ordinator (who knew that disulfiram was being used in two dosages), which was the dosage level of disulfiram used. Both the study co-ordinator and the therapist could say that they did not know which medication the patient received. The responses were then compared to the actual treatment group. If the blinding procedures were completely effective, one would expect that the respondent would either report 'don't know' or would in essence randomly choose a medication.

We can display the response data in such a situation in a $(k + 1) \times k$ frequency table, where k is the number of treatments, and we denote the actual treatment groups and responses by the columns and rows, respectively. Table I shows the configuration for the disulfiram study where $k = 3$. The first three rows of Table I indicate specific responses made by the respondent. The diagonal and off-diagonal cells represent correct and incorrect responses, respectively. The fourth row records the frequency of the 'don't know' responses. One measure of agreement is obtained by the traditional kappa coefficient[11] ignoring the 'don't know' responses:

$$\kappa_A = (p_{Ao} - p_{Ae})/(1 - p_{Ae}) \tag{1}$$

where

$$p_{Ao} = \sum_{i=1}^{3} p_{ii}, \quad p_{Ae} = \sum_{i=1}^{3} p_{.i} p_{i.}$$

$p_{ii} = n_{ii}/L$, $p_{.i} = n_{.i}/L$, $p_{i.} = n_{i.}/L$, and L is the total number of specific responses excluding the 'don't know'. κ_A is a chance corrected index. If there is complete agreement, $\hat{\kappa}_A = +1$. If the

Table I. Data for measuring responses between guesses and actual treatment

Guessed treatment	Actual treatment			
↓	Disulfiram 1 mg	Disulfiram 250 mg	Riboflavin	Total
Disulfiram 1 mg	n_{11}	n_{12}	n_{13}	$n_{1.}$
Disulfiram 250 mg	n_{21}	n_{22}	n_{23}	$n_{2.}$
Riboflavin	n_{31}	n_{32}	n_{33}	$n_{3.}$
Don't know	n_{01}	n_{02}	n_{03}	$n_{0.}$
Total sample	$n_{.1}$	$n_{.2}$	$n_{.3}$	N

N = total number of patients

observed agreement is greater than or equal to chance, $\hat{\kappa}_A \geqslant 0$, and if the observed agreement is less than or equal to the chance agreement, $\hat{\kappa}_A \leqslant 0$.

There are, however, several problems with κ_A when used to measure the success of blinding procedures. First, the index is designed to measure agreement rather than disagreement. It is a measure sensitive to where between random guessing ($\kappa_A = 0$) and total agreement ($\kappa_A = 1$) a particular situation lies. Values of kappa below zero indicate less than random agreement, but the index is not scaled to provide meaningful indicators at that end of the scale. In fact, this form of kappa has a lower bound that changes with the number of response categories and the marginal probabilities, and thus negative values are essentially uninterpretable. Second, the index ignores the 'don't knows', which is the response most indicative of blinding. To resolve these problems, we propose a variation of a kappa coefficient that is sensitive not to the degree of agreement, but to the degree of disagreement, and places more appropriate weight on the desirable 'don't know' responses. An important assumption in the construction of such an index is that when a respondent says that s/he does not know, that represents an honest response, not just a socially desirable response or one that avoids making an assessment. It is therefore important that in obtaining the reports that one encourages the respondents to make their best effort to report their suspicions when such suspicions exist.

An Index of Success of Blinding

The traditional kappa (equation (1)) assigns a weight of 1 to the diagonal cells (complete agreement) and a weight of 0 to the off-diagonal cells (disagreement) in the $k \times k$ table excluding the 'don't knows'. Kappa then relates this score to what one would obtain if the responses were random and to what is ideal attainment (complete agreement).

Here the principle is the same. However, correct guesses are least supportive of blinding and we assign them a weight of 0, while 'don't know' responses are most supportive and we assigned them a weight of 1. Other responses in which the respondent thinks s/he knows but guesses incorrectly are intermediate and we assign them intermediate weights.

We define a blinding index score, BI, as follows:

$$\mathrm{BI} = [1 + P_{\mathrm{DK}} + (1 - P_{\mathrm{DK}})\kappa_{\mathrm{D}}]/2,$$

where

$$\kappa_{\mathrm{D}} = (p_{\mathrm{Do}} - p_{\mathrm{De}})/p_{\mathrm{De}}.$$

Table II. Six hypothetical configurations of actual versus guessed treatment identifications and the resulting $\widehat{BI}$'

Guessed treatment	Actual treatment								
	Dsf 1 mg	Dsf 250 mg	Rbf	Dsf 1 mg	Dsf 250 mg	Rbf	Dsf 1 mg	Dsf 250 mg	Rbf
Dsf 1 mg	29	0	0	24	5	0	13	5	5
Dsf 250 mg	0	29	0	0	24	5	5	13	5
Rbf	0	0	29	5	0	24	5	5	13
DK	4	4	4	4	4	4	10	10	10
$\widehat{BI}$		0·121			0·235			0·530	
Dsf 1 mg	5	14	14	3	0	30	0	0	5
Dsf 250 mg	14	5	14	0	0	0	0	0	0
Rbf	14	14	5	30	33	3	5	5	0
DK	0	0	0	0	0	0	28	28	28
$\hat{B}I$		0·637			0·746			0·971	

Dsf Disulfiram
Rbf Riboflavin
DK Don't know

The weighted proportion of observed guesses is

$$p_{\mathrm{Do}} = \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}p_{ij}/(1 - P_{\mathrm{DK}}), \quad \text{for } P_{\mathrm{DK}} \neq 1,$$

and the weighted proportion of expected guesses is

$$p_{\mathrm{De}} = \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}p_{i.}(p_{.j} - p_{0j})/(1 - P_{\mathrm{DK}})^2$$

where the w_{ij} are the weights for the specific responses and the p's are the expected relative frequencies of the $(k+1)\times k$ table. That is, if we denote the estimated probabilities by $\hat{p}$, then $\hat{p}_{ij} = n_{ij}/N, \hat{p}_{0j} = n_{0j}/N, \hat{p}_{.j} = n_{.j}/N, \hat{p}_{i.} = n_{i.}/N$, and $\hat{P}_{\mathrm{DK}} = n_{0.}/N$, $n_{0.}$ being the total number of 'don't know' responses and N the total sample size. In this index, κ_{D}, for the upper $k \times k$ ('guess') portion of the matrix, is similar to κ_{A} and is scaled on an interval of -1 to $+1$. $(1 - P_{\mathrm{DK}})$ and P_{DK} apportion the 'guesses' and the 'don't knows'. Adding 1 and dividing the quantity in brackets by 2 scales BI to 0–1, which is, in general, a desirable index interval. In our application, $k = 3$ and we assigned the weights as follows: $w_{11} = w_{22} = w_{33} = 0{\cdot}00$ (correct guess), $w_{12} = w_{21} = 0{\cdot}5$ (correctly guessed the medication, but wrong dose), $w_{13} = w_{31} = w_{23} = w_{32} = 0{\cdot}75$ (guessed wrong medication), and $w_{01} = w_{02} = w_{03} = 1{\cdot}00$ (responded 'don't know').

If all respondents report 'don't know', then the value of the index is 1 indicating success of blinding procedures, while if all respondents report correct answers, the value of the index is 0. Table II shows some hypothetical cases between these extremes. As the responses shift from being correct to 'don't know', the estimate of BI increases from 0 to 1. In the null case in which there are no 'don't knows' and the answers are reported randomly (the cell probabilities are products of the corresponding marginal probabilties), the index is 0·5.

The estimate of of BI, $\widehat{BI}$, is

$$[1 + \hat{P}_{\mathrm{DK}} + (1 - \hat{P}_{\mathrm{DK}})\hat{k}_{\mathrm{D}}]/2$$

and its asymptotic variance (see Appendix) is

$$\operatorname{var}(\hat{B}I) = \left\{ \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}(1 - p_{\mathrm{DK}})^2 \left[(1 - p_{\mathrm{DK}})w_{ij} - (1 + \kappa_{\mathrm{D}}) \sum_{r=1}^{k} \{p_{r.}w_{rj} + (p_{.r} - p_{0r})w_{ir}\}\right]^2}{4\left[\sum_{i=1}^{k}\sum_{j=1}^{k} p_{i.}(p_{.j} - p_{0j})w_{ij}\right]^2} \right.$$

$$\left. + p_{\mathrm{DK}}(1 - p_{\mathrm{DK}}) - (1 - p_{\mathrm{DK}})(1 + \kappa_{\mathrm{D}}) \left[p_{\mathrm{DK}} + \frac{(1 - p_{\mathrm{DK}})(1 + \kappa_{\mathrm{D}})}{4}\right] \right\} \Bigg/ N.$$

We also used the jack-knife procedure[12] (see Appendix) to compute the variance of $\widehat{BI}$.

Assessment of the Blind Across Participating Hospitals

The estimate of the index, $\widehat{BI}$, was proposed to measure how well a programme therapist or study co-ordinator guessed the correct treatment group, and in turn, how well the treatment bind was maintained for each hospital.

RESULTS

Application to Cooperative Study No. 107

Study co-ordinators

Table III shows the degree of association between the guesses and actual treatments for the study co-ordinators pooled across the nine participating hospitals. This configuration yielded a $\hat{B}I$ of 0·56 with 95 per cent confidence limits of 0·52 to 0·59, indicating a response pattern close to that expected by random guessing, that is, partial but not complete blindness with no 'don't knows'. For the nine hospitals overall, 177 (33·5 per cent) of the 529 responses were correct. There were 147 patients (28 per cent) for whom the co-ordinators indicated that they did not know which drug the patient was taking. Table IV shows the coefficients indicating the success of blinding in the nine participating hospitals. Note that the hypothesis of non-blinding (BI = 0) is rejected at the two-tailed 5 per cent level in all of the nine hospitals. The success of blinding was significantly above random guessing (BI = 0·50) in four hospitals (A, E, F and G), the magnitude of $\widehat{BI}$ being 0·65, about one-third of the way between random guessing and perfect blindness and two-thirds of the way between non-blindness and perfect blindness. The success of blindness was not significantly above random guessing for the other five hospitals.

Programme Therapists

The programme therapists were aware that the patients received disulfiram or riboflavin, but were not aware that disulfiram consisted of two possible doses. Thus there were two possible

Table III. Study co-ordinators' responses pooled across hospitals

Guessed treatment	Actual treatment			
	Disulfiram 1 mg	Disulfiram 250 mg	Riboflavin	Total
	Number Weight Proportion	Number Weight Proportion	Number Weight Proportion	Number Proportion
Disulfiram 1 mg	41 0·00 0·08	27 0·50 0·05	22 0·75 0·04	90 0·17
Disulfiram 250 mg	66 0·50 0·13	72 0·00 0·13	36 0·75 0·07	174 0·33
Riboflavin	30 0·75 0·05	24 0·75 0·05	64 0·00 0·12	118 0·22
Don't know	44 1·00 0·08	51 1·00 0·10	52 1·00 0·10	147 0·28
Total sample	181	174	174	529
Proportion of sample	0·34	0·33	0·33	1·00
$\widehat{BI}$	95% jack-knife confidence limits		95% asymptotic confidence limits	
0·556	0·520 to 0·592		0·521 to 0·592	

Table IV. Study Co-ordinator blindness indices, $\widehat{BI}$'s, by hospital

Hospitals	$\widehat{BI}$	95% jack-knife confidence limits	95% asymptotic confidence limits
A	0·678	0·589 to 0·765	0·592 to 0·764
B	0·500	0·331 to 0·662	0·342 to 0·658
C	0·472	0·372 to 0·570	0·375 to 0·569
D	0·418	0·234 to 0·588	0·251 to 0·585
E	0·645	0·548 to 0·738	0·552 to 0·738
F	0·648	0·540 to 0·750	0·545 to 0·751
G	0·615	0·503 to 0·726	0·506 to 0·724
H	0·379	0·285 to 0·469	0·289 to 0·469
I	0·501	0·409 to 0·591	0·412 to 0·591

responses and the probability of responding correctly was 0·5. Table V shows a representation of the therapists' responses with a 3 × 2 frequency table. Overall, 204 (48 per cent) of the 423 responses were correct. The programme therapists indicated that they did not know which treatment the patient received in 114 cases (27 per cent). The estimate of BI was 0·54, with 95 per cent confidence limits of 0·49 to 0·58, indicating a level of blinding expected by random guessing

Table V. Programme therapists' responses pooled across hospitals

Guessed treatment	Actual treatment		
	Disulfiram	Riboflavin	Total
	Number	Number	Number
	Weight	Weight	
	Proportion	Proportion	Proportion
Disulfiram	145	34	179
	0·00	0·75	
	0·34	0·08	0·42
Riboflavin	71	59	130
	0·75	0·00	
	0·17	0·14	0·31
Don't know	76	38	114
	1·00	1·00	
	0·18	0·09	0·27
Total sample	292	131	423
Proportion of sample	0·69	0·31	1·00
$\widehat{BI}$	95% jack-knife confidence limits		95% asymptotic confidence limits
0·535	0·487 to 0·582		0·487 to 0·582

Table VI. Programme therapists' blindness indices, $\widehat{BI}$'s, by hospital

Hospitals	$\widehat{BI}$	95% jack-knife CI	95% asymptotic CI
A	0·543	0·406 to 0·682	0·413 to 0·681
B	0·544	0·341 to 0·745	0·356 to 0·732
C	0·405	0·259 to 0·546	0·267 to 0·544
D	0·480	0·234 to 0·704	0·263 to 0·700
E	0·793	0·700 to 0·888	0·701 to 0·884
F	0·857	0·679 to 0·998	0·736 to 0·978
G	0·170	0·050 to 0·286	0·054 to 0·286
H	0·459	0·339 to 0·577	0·342 to 0·576
I	0·508	0·394 to 0·621	0·397 to 0·619

with no 'don't knows'. Table VI shows the coefficients indicating success of blinding in the nine participating hospitals. The estimated index was significantly above the chance level for two hospitals (E and F) with indices of 0·79 and 0·85 and 95 per cent confidence limits on the order of 0·70 to 0·98.

To compare the blinding success of the study co-ordinators to the programme therapists, the 4 × 3 frequency tables for the co-ordinators were collapsed into 3 × 2 tables that corresponded to the programme therapists' responses (data not shown). Except for three hospitals, the blinding success was very similar in both groups of respondents. The study co-ordinators' blindness indices for hospitals E and F were approximately 0·60 compared to indices of about 0·80 for the

Table VII. Effect of varying weights on blindness indices, $\widehat{BI}$'s, study co-ordinators pooled across hospitals

Weight 1*	Weight 2	$\widehat{BI}$	95% jack-knife confidence limits
0·20	0·40	0·550	0·513 to 0·587
0·20	0·75	0·539	0·500 to 0·578
0·20	0·90	0·537	0·497 to 0·577
0·50	0·40	0·573	0·538 to 0·610
0·50	0·75	0·556	0·520 to 0·592
0·50	0·90	0·552	0·516 to 0·589
0·80	0·40	0·588	0·551 to 0·624
0·80	0·75	0·569	0·534 to 0·604
0·80	0·90	0·564	0·528 to 0·599

*Varying weights
Weight 1: Correct treatment, incorrect dose
Weight 2: Incorrect treatment
Fixed weights
0·00: Correct identification
1·00: Don't know

programme therapists. In hospital G the co-ordinators' index was 0·62 compared to the therapists' index of 0·17, which indicated almost total lack of blinding.

Robustness of the Procedure for Varying Weights

Table VII shows the coefficients indicating success of blinding pooled across the nine participating hospitals for varying weights. With the first weight (correct treatment, incorrect dose) fixed, $\widehat{BI}$ decreases as the second weight (incorrect treatment) increases. With the second weight fixed, $\widehat{BI}$ increases as the first weight increases. The $\widehat{BI}$'s and their 95 per cent confidence limits remained in the range of 0·50 to 0·60, illustrating that varying weights over a wide range of values had little effect on the test statistic. Usually with ordered response categories, the particular choice of weights is not too influential if they reflect the response order.[13]

DISCUSSION

Argument for Testing Blinding

In most clinical trials steps are taken to test the success of treatment randomization. Usually this consists of examining the treatment regimens for differences in patient characteristics. When differences are found post-stratification, covariance analysis is often used to adjust for them. Less frequently are measures taken to check for the blindness of the treatments administered, although this certainly represents an important consideration and a basic assumption in the conduct of many trials.[14] Blindness is particularly important in trials where the endpoint is open to subjective evaluation, as was the case in Cooperative Study Number 107 where abstinence from alcohol was among the primary endpoints. It is less of an issue in trials where more objective endpoints such as mortality are used.

The index described in this paper provides a test of blinding. It is unavoidably a mix of the effects of actual unblinding because of 'slips of the tongue' and cheating as well as physiologic effects due to extreme efficacy and dramatic side-effects. How lack of blinding will affect the results of the trial or what steps one should take in the analysis to adjust for unblinding is an important consideration. One can assess the effect of unblinding in a multi-centre trial by comparing the endpoints in the hospitals that maintained a moderate to high degree of blindness (say $\widehat{BI} > 0{\cdot}60$) to those in the hospitals where there appeared to be some degree of unblinding (say $\widehat{BI} < 0{\cdot}50$). If the results differ between these hospital groups, additional analyses may be necessary to determine the reasons for such a difference, although the nature of these analyses may be open to question. Oxtoby *et al.*[14] argue for reanalyses excluding those patients whose treatments were identified because of side-effects, appearance, taste or smell. On the other hand, Newcombe[15] asserts that all patients should be included in the analysis, primarily because of the intention-to-treat principle, but for other reasons as well.

For Cooperative Study Number 107, abstinence was significantly greater for disulfiram (1 mg disulfirm, $P = 0{\cdot}025$; 250 mg disulfiram, $p = 0{\cdot}018$) in the hospitals with blinding indices $> 0{\cdot}60$ (A, E, F, G) than in the hospitals with indices $\leqslant 0{\cdot}50$ (B, C, D, H, I), when abstinence was based solely on interviews with the patient and his relative/friend. We hypothesize that beliefs by the study co-ordinators that patients assigned to disulfiram should have been more abstinent may have unconsciously biased the manner in which they elicited the abstinence information. Adding more definitive information from the urine and blood laboratory test and other medical records reduced the overall abstinence rate for each treatment and decreased the significance of the difference in abstinence between the blinded and unblinded centres (1 mg disulfiram, $P = 0{\cdot}08$; 250 mg disulfiram, $P = 0{\cdot}14$). Unblinding appeared to affect abstinence rates in disulfiram assigned patients, but not in riboflavin assigned patients ($P = 0{\cdot}23$).

Asking about Blindness

The blindness index proposed in this paper assumes that the most desirable response is 'don't know'. If the respondents detect that this is the most socially acceptable response, they may provide this answer even though they really think that the patient received a specific study treatment. Therefore, the key to blindness assessment is to elicit the information so that the respondent provides an honest accounting of which treatment s/he thinks the patient received. One should encourage respondents to answer 'don't know' only if they truly do not know or cannot make an educated guess. In Cooperative Study Number 107, blindness information was elicited by asking an open ended question about the study medication the patient received. The study co-ordinator could specify the medication and dosage, say that s/he did not know, or provide another answer. The programme therapist had the same response options, except that they were unaware that there were two doses of disulfiram. The responses were subsequently coded by the Chairman's Office into three categories: disulfiram (with the dose); riboflavin, or 'don't know'. Medication assignments were elicited in this manner to provide the study co-ordinator or programme therapist with maximum leeway in answering the questions and lessen the motivation for giving a socially acceptable answer. In retrospect, it is possible that we should have given more encouragement to specifying the medication and dose if the study co-ordinator or programme therapist was fairly certain of what they were.

Variability of Blindness Across Hospitals

As discussed previously, blinding success varied across hospitals. The study co-ordinator's blindness indices for hospitals A, E, F and G were significantly above the chance level. The proportion of correct responses for these hospitals ranged from 24 per cent to 29 per cent, with the proportion of 'don't knows' ranging from 25 per cent in hospital G to more than 40 per cent in the other three hospitals, the highest being 51 per cent in hospital F. The proportion of correct responses for the remaining five hospitals, which had $\widehat{BI}$'s less than 0·5, ranged from 39 per cent to 47 per cent, and the proportion of 'don't knows' ranged from 3 per cent to 19 per cent. The co-ordinators in these hospitals appeared to be at least partially unblinded, which may have resulted from the stability of the co-ordinators over the course of the study, thereby increasing the chance for exposure to the treatment identities. Co-ordinators hired later, particularly after the completion of patient recruitment and randomization, had less opportunity to speak with the patients about their treatment assignments.

There was greater variability in the blinding success for the programme therapists with $\widehat{BI}$'s ranging from 0·16 in hospital G to 0·85 in hospital F. In hospital G, 84 per cent of the responses were correct and 9 per cent of the responses were 'don't know', whereas in hospital F only 15 per cent of the answers were correct and 82 per cent were 'don't knows'. Most of the $\widehat{BI}$'s were in the 0·45–0·55 range, but the confidence intervals on these indices were wider than confidence intervals for the study co-ordinators because they were based on fewer responses. The programme therapists may have been become unblinded because some hospitals encouraged open and free communication between the programme therapists and their patients. While patients were asked not to discuss their treatment with clinic personnel, a 'slip of the tongue' is more likely to have occurred in such an environment. As with the study co-ordinators, one can speculate that those treatment programme with stable personnel provided the opportunity for greater interaction between patients and therapists and this could have resulted in more chance for unblinding. Also, one might expect that patients who exhibited greater compliance in keeping their scheduled visits might have been more likely to divulge their treatment assignment.

Assignment of Weights

In the construction of the blinding index we assigned weights of 0·0 and 1·0, respectively, to the correct and 'don't know' responses to reflect their desirability. The assignment of intermediate weights is more open to judgement, and, again, depends upon the desirability of the response. For Cooperative Study Number 107, we assigned the weight of 0·50 to 'correct treatment, wrong dose' reflecting sufficient information to guess the right treatment but not enough knowledge to guess the 'right dose'. We assigned the weight of 0·75 to 'wrong treatment' (which was more desirable from a blindness standpoint than 'right treatment, wrong dose') as an indication that the respondent felt sure enough to venture a guess, but guessed the wrong treatment. Thus the weights progressed from less to more blindness. In general, when a study involves multiple treatments, one can feel safe in assigning weights that reflect the desirability of the responses. We found, however, that with anchor points of 0·0 and 1·0, one can vary the intermediate weights considerably without substantially altering the blindness index.

CONCLUSIONS

The blindness index, BI, appears to have several desirable qualities for assessing blinding success in a clinical trial: it increases as the success of blinding increases; it accounts for the 'don't know' responses, which, in a truly blinded study, is the most desirable response; it is scaled to a meaningful interval of 0·0 to 1·0; and it is robust under changing intermediate weights. We used jack-knife and asymptotic procedures to approximate the index variance and 95 per cent confidence limits; both methods are straightforward and yield comparable estimates for large sample size.

APPENDIX

Asymptotic Variance of the Blinding Index Statistic

Let

$$T = 2\widehat{BI} - 1 = \hat{p}_{\mathrm{DK}} + (1 - \hat{p}_{\mathrm{DK}})\hat{\kappa}_{\mathrm{D}}$$

where

$$\hat{p}_{\mathrm{DK}} = n_{0.}/N, \qquad \hat{\kappa}_{\mathrm{D}} = (p_{\mathrm{o}} - p_{\mathrm{e}})/p_{\mathrm{e}},$$

$$p_{\mathrm{o}} = \sum_{i=1}^{k} \sum_{j=1}^{k} n_{ij} w_{ij}/(N - n_{0.})$$

and

$$p_{\mathrm{e}} = \sum_{i=1}^{k} \sum_{j=1}^{k} n_{i.}(n_{.j} - n_{0j}) w_{ij}/(N - n_{0.}).$$

Here, w_{ij} is the weight assigned to the (i, j)th cell of the $(k + 1) \times k$ frequency table, k equals the number of treatments, and the frequencies are as defined in the text; see Table I for the case where $k = 3$. The subscript i equals $k + 1$ for the row of 'don't know' responses.

We use the following result due to Fisher[16] to derive the variance of T to order $O(1/N)$. Let $T(m_1, m_2, \ldots, m_l)$ be any function of the observed frequencies $m_1, m_2, \ldots, m_l$ of a sample of size M from an l-nominal distribution with probabilities $p_1, p_2, \ldots, p_l$ ($\Sigma m_h = M$, $\Sigma p_h = 1$). Then

$$\frac{1}{M}\operatorname{var}(T) = \sum_{h=1}^{l} p_h \left(\frac{\partial T}{\partial m_h}\right)^2 - \left(\frac{\partial T}{\partial M}\right)^2$$

asymptotically, the derivatives being taken at the values $m_h = Mp_h$.

In our application, $M = N$ and we need only consider $l = k^2 + 1$ cells, where the m_h are n_{ij} for $i = 1, \ldots, k$ and $j = 1, \ldots, k$ and the last cell frequency equals $n_{0.}$. Then, evaluating the derivatives at $n_{ij} = Np_{ij}$ and $n_{0.} = Np_{\mathrm{DK}}$ leads to

$$\frac{\partial T}{\partial n_{ij}} = \frac{(1 - p_{\mathrm{DK}})\left[(1 - p_{\mathrm{DK}}) w_{ij} - (1 + \kappa_{\mathrm{D}}) \sum_{r=1}^{k} \{p_{r.} w_{rj} + (p_{.r} - p_{0r}) w_{ir}\}\right]}{N \sum_{i=1}^{k} \sum_{j=1}^{k} p_{i.}(p_{.j} - p_{0j}) w_{ij}}$$

for $i = 1, \ldots, k$ and $j = 1, \ldots, k$,

$$\frac{\partial T}{\partial n_{0.}} = \frac{-2\kappa_{\mathrm{D}}}{N}$$

and

$$\frac{\partial T}{\partial N} = \frac{(1 - p_{\mathrm{DK}}) + \kappa_{\mathrm{D}}(1 + p_{\mathrm{DK}})}{N}$$

Hence to order $O(1/N)$

$$\operatorname{var}(\widehat{BI}) = \left\{ \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} p_{ij}(1-p_{\mathrm{DK}})^2 \left[(1-p_{\mathrm{DK}})w_{ij} - (1+\kappa_{\mathrm{D}}) \sum_{r=1}^{k} \{ p_{r.}w_{rj} + (p_{.r} - p_{0r})w_{ir} \} \right]^2}{4\left[\sum_{i=1}^{k} \sum_{j=1}^{k} p_{i.}(p_{.j} - p_{0j})w_{ij} \right]^2} \right.$$

$$\left. + p_{\mathrm{DK}}(1-p_{\mathrm{DK}}) - (1-p_{\mathrm{DK}})(1+\kappa_{\mathrm{D}}) \left[p_{\mathrm{DK}} + \frac{(1-p_{\mathrm{DK}})(1+\kappa_{\mathrm{D}})}{4} \right] \right\} \Big/ N.$$

Jack-knife Procedure for Constructing 95 per cent Confidence Limits

For a defined test statistic, $T_0(x)$, consisting of observations $x_1, x_2, \ldots, x_n$, the jack-knife procedure consists of deleting one observation at a time and computing n pseudo values of the form

$$J_i = nT_0(x) - (n-1)T_i,$$

where T_i is the test statistic $T(x)$ computed on the remaining $n-1$ observations.

The jack-knife mean is

$$J(\theta) = \sum_{i=1}^{n} J_i/n$$

and the jack-knife variance is

$$s_J^2 = \sum_{i=1}^{n} [J_i - J(\theta)]^2/(n-1).$$

$\mathrm{SE}_J = s_J/\sqrt{n}$, from which 95 per cent confidence limits on θ are constructed by

$$J(\theta) \pm 1{\cdot}96\,\mathrm{SE}_J \qquad (1)$$

With the test statistic $T_0(x)$ defined as the estimate of the blinding index $\widehat{BI}$, deletion of an observation in each of the $(k+1) \times k$ cells (4×3 for study co-ordinators, 3×2 for programme therapists) yields $(k+1) \times k$ pseudo values and the jack-knife mean becomes

$$J(\widehat{BI}) = \sum_{i=1}^{k+1} \sum_{j=1}^{k} n_{ij} J_{ij}/N$$

where J_{ij} is the pseudo value obtained by reducing the ijth cell ($n_{ij} > 0$) by 1 and computing $J_{ij} = \widehat{BI}$ on the remaining $N-1$ observations. Likewise, the jack-knife variance becomes

$$s_J^2 = \sum_{i=1}^{k+1} \sum_{j=1}^{k} n_{ij}[J_{ij} - J(BI)]^2/(N-1)$$

and the 95 per cent confidence limits on $\hat{B}I$ follow from equation (1). Copies of FORTRAN programs for the estimation and variances procedures can be obtained from Dr. James.

Cooperative Study Number 107 Personnel

Reference 9 provides a complete listing of Cooperative Study Number 107 personnel.

REFERENCES

1. Byington, R. P., Curb, J. D., Mattson, M. E., for the Beta-Blocker heart Attack Trial Research Group. 'Assessments of double-blindness at the conclusion of the Beta-Blocker Heart Attack Trial', *Journal of the American Medical Association*, **253**, 1733–1736 (1985).
2. Moscucci, M., Byrne, L., Weintraub, M. and Cox, C. 'Blinding, unblinding, and the placebo effect: An analysis of patients' guesses of treatment assignment in a double-blind clinical trial', *Clinical Pharmacology and Therapeutics*, **41**, 259–265 (1987).
3. Howard, J., Whittemore, A. S., Hoover, J. J., Panos, M. and the Aspirin Myocardial Infarction Study Research Group. 'Commentary: How blind was the patient blind in AMIS?', *Clinical Pharmacology and Therapeutics*, **32**, 543–553 (1982).
4. Hughes, J. R. and Krahn, D. 'Blindness and variability of the double-blind procedure', *Journal of Clinical Psychopharmacology*, **5**, 138–142 (1985).
5. Bownell, K. D. and Stunkard, A. J. 'The double-blind in danger: Untoward consequences of informed consent', *American Journal of Psychiatry*, **139**, 1487–1489 (1982).
6. Deyo, R. A., Walsh, N. E., Schoenfield, L. S. and Ramamurthy, S. 'Can trials of physical treatment be blinded', *American Journal of Physical Medicine and Rehabilitation*, **69**, 6–10 (1990).
7. Jesperson, C. M. and the Danish Study Group on Verapamil in Myocardial Infarction. 'Assessment of blindness in the Danish Verapamil Infraction Trial II (DAVIT II)', *European Journal of Clinical Pharmacology*, **39**, 75–76 (1990).
8. Marini, J. L., Sheard, M. H., Bridges, C. I. and Wagner, E. 'An evaluation of the double-blind design in a study comparing lithium carbonate with placebo', *Acta Psychiat. Scand.*, **53**, 343–354 (1976).
9. Fuller, R. K., Branchy L., Brightwell, D., Derman, R. M., Emrick, C. D. *et al.* 'Disulfiram treatment of alcoholism: A Veterans Administration cooperative study', *Journal of the American Medical Association*, **256**, 1449–1455 (1986).
10. Fuller, R. K., Williford, W. O., William, O., Lee, K. K. and Derman, R. 'Veterans Administration cooperative study of disulfiram for the treatment of alcoholism: Study design and methodological considerations', *Controlled Clinical Trials*, **5**, 263–273 (1984).
11. Fleiss, J. L. *Statistical Methods for Rates and Proportions*, 2nd edn, Wiley, New York, 1977, p. 219.
12. Efron, B. *The Jackknife, Bootstrap, and Other Resampling Plans*, Society for Industrial and Applied Mathematics, Philadelphia, 1982.
13. Moses, L. E. *Think and Explain with Statistics*, Addison-Wesley, Menlo Park, 1986, p. 433.
14. Oxtoby, A., Jones, A. and Robinson, M. 'Is your double-blind design truly double-blind?', *British Journal of Psychiatry*, **155**, 700–701 (1989).
15. Newcombe, R. G., 'Letter to the editor', *British Journal of Psychiatry*, **156**, 282 (1990).
16. Fisher, R. A. *Statistical Methods for Research Workers*, 13th edn, Hafner, New York, 1958, p. 309.