Biometrika (1990), 77, 3, *pp.* 467–84
Printed in Great Britain

On inferences from Wei's biased coin design for clinical trials

BY COLIN B. BEGG
Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, New York 10021, U.S.A.

SUMMARY

Wei (1988) analyzed data from a clinical trial in which an urn-sampling model was used to allocate patients to treatments. The trial resulted in 11 patients being allocated to the experimental treatment, all successes, and with one patient allocated to the control treatment, a failure. Wei analyzed these data using a randomization algorithm and concluded that the results were almost significant at $p = 0{\cdot}051$. He asserted that an analysis which ignored the design and presumed complete randomization leads to a p-value of 0·001. In fact, if a more conventional analysis is used in which both margins are fixed, then this leads to $p = 0{\cdot}083$, Fisher's exact test, if complete randomization is assumed. If the urn-sampling allocation is taken into account much more conservative inferences are obtained. An analysis which conditions on both margins leads to $p = 0{\cdot}28$, while one which conditions on the observed sequence of responses, and the observed treatment totals leads to $p = 0{\cdot}62$. These serious discrepancies are discussed, in addition to the inappropriateness of biased coin design and small sample sizes in important medical trials.

Some key words: Clinical trial; Play-the-winner randomization; Randomization distribution.

1. INTRODUCTION

A recent clinical trial compared extracorporeal membrane oxygenation, ECMO, versus conventional therapy in the treatment of newborns with respiratory failure (Bartlett et al., 1985). Because of the seriousness of the condition and the anticipated large treatment effect a modified play-the-winner strategy was employed (Wei & Durham, 1978). Specifically, the first allocation was conceptually based on the selection of either an A ball, ECMO, or a B ball, control, from an urn containing one ball of each type. For subsequent allocations the mix of balls in the urn was progressively changed. If either a success on treatment A or a failure on B was recorded, an extra ball of type A was added to the urn, and vice versa. Unfortunately, the trial resulted in a rather peculiar sequence of outcomes. The first patient received treatment A, and the treatment was successful. The second patient failed on B. This was followed by ten consecutive successes on A at which point the trial was terminated. The statistical test used was based on the permutation distribution of the treatments conditional on the sequence of responses, taking into account the biased coin design. The test statistic was the number of successes on treatment A, which turned out to be the maximum possible given the sequence of responses, namely eleven. This led to a one-sided p-value of 0·051. For more complete details see Wei (1988). Wei also indicated that if the biased coin design is ignored in the analysis, and it is assumed that complete randomization was used, the corresponding p-value is 0·001.

The goal of this paper is to clarify the implications of Wei's methods, to demonstrate the radically different conclusions that are obtained when more conventional approaches are used, and to attempt to explain the reasons for the discrepancies.

2. METHODS

Wei's notation is used throughout with the exception that m denotes the sample size rather than n. That is, Y_j is 1 or 0 according to whether the jth patient is assigned to A or B respectively, $j = 1, \ldots, m$; X_j is 1 or 0 if the jth response is success or failure; and

$$S_i = \sum_{j=1}^{i} X_j Y_j, \quad N_i = \sum_{j=1}^{i} Y_j, \quad B_i = \sum_{j=1}^{i} X_j.$$

Following Wei's notational convention, we denote random variables in lower case in circumstances where we are conditioning on observed outcomes. For simplicity, formulae are specific to the urn model in which there is one ball of each type at the start, and one ball is added at each stage.

Wei uses S_m as the summary statistic. The probability distribution of S_m is calculated by first conditioning on the sequence of outcomes $\{X_i\}$, and then evaluating the probabilities of all possible sequences of treatment assignments. The distribution of S_m is evaluated by aggregating the probabilities of sequences that lead to the same value of S_m. The probability of any individual sequence of values of $\{Y_i\}$ is

$$\text{pr}\,(\{Y_i\}|\{X_i\}) = \prod_{i=1}^{n} p_i^{Y_i}(1-p_i)^{1-Y_i}, \tag{1}$$

where

$$p_i = (2S_{i-1} + i - N_{i-1} - B_{i-1})/(i+1). \tag{2}$$

When this is applied to the ECMO study, for which $S_{12} = 11$, the p-value is 0·051, since

$$\text{pr}\,(S_{12} \geqslant 11) = \text{pr}\,(S_{12} = 11) = 0{\cdot}051.$$

If one ignores the biased coin allocation probabilities, and assumes that complete randomization was used, then $p_i = \frac{1}{2}$, for all i, and pr $(S_{12} \geqslant 11) = 0{\cdot}00049$. More specifically, of the 2^{12} possible permutations of A's and B's, only two lead to $S_{12} = 11$, the one which actually occurred and the one in which all twelve patients are randomized to A.

There are several reasons why I believe Wei's test is inappropriate. First, the use of a test based on all possible permutations of the treatments is questionable. As indicated in the previous paragraph, the permutation in which all patients are randomized to A is included in the sample space, yet clearly such an outcome would provide no information about the treatment comparison. The same can be said for the permutation in which all patients are randomized to B. In fact, since only one patient actually received B in the trial, it is unclear that any permutation of treatments which has more or less than exactly one allocation to B should be included. Indeed, conditioning on both margins of the 2×2 table is the conventional approach in randomized experiments, and there are compelling theoretical reasons why this is so. In a fixed sample, completely randomized study, the number of responses on one of the treatments, given both margins, depends only on the treatment effect, i.e. the nuisance parameter is eliminated, and therefore is a suitable candidate as a test statistic (Cox, 1977, pp. 45–52). Indeed, the test of no treatment effect derived in this way is Fisher's exact test. For the data in the ECMO study the use of Fisher's exact test leads to a p-value of 0·083, or $\frac{1}{12}$.

The validity of Fisher's exact test is clearly violated when a biased coin randomization scheme is used since the various permutations of treatments and responses are not equiprobable. In fact, for the sequence of responses observed in the trial the permutation in which B is allocated to the first patient, a success, and A to the remaining eleven patients, one failure followed by ten successes, denoted P_1, has probability proportional to $\frac{1}{2}\times\frac{1}{3}\times\frac{1}{4}\times\frac{2}{5}\times\frac{3}{6}\times\cdots\times\frac{10}{13}$. This follows directly from (1). Likewise, the permutation in which B is allocated to the second patient, P_2, has probability proportional to $\frac{1}{2}\times\frac{1}{3}\times\frac{3}{4}\times\frac{4}{5}\times\cdots\times\frac{12}{13}$. Similarly, P_3, is proportional to $\frac{1}{2}\times\frac{2}{3}\times\frac{2}{4}\times\frac{2}{5}\times\frac{3}{6}\times\cdots\times\frac{10}{13}$, and $P_3=P_4=\ldots=P_{12}$. Consequently, if we restrict attention to permutations with eleven A's and one B, $P_1=0{\cdot}009$, $P_2=0{\cdot}617$, $P_3=0{\cdot}037$, etc. In other words, P_2, the permutation which was realized in the trial, is by far the most likely one. Therefore, if one conditions on the observed sequence of responses but restricts attention to permutations with $N_{12}=11$, the p-value is 0·62. By contrast, Wei's test which does not further condition on N_{12} leads to $p=0{\cdot}05$. However, the rationale for conditioning on the sequence of responses is itself unclear. Indeed if we construct a test in which N_{12} and B_{12} are both fixed, as in Fisher's exact test, there are 144 possible permutations of which 12 lead to $S_{12}=11$ and the rest lead to $S_{12}=10$. This leads to $p=0{\cdot}28$.

In interpreting these results, it is useful to return to the concept of sufficiency and the construction of conditional tests. Suppose that the logit of the probability of a response is $\gamma+\delta$ for treatment A and γ for treatment B. That is, δ is the treatment effect and γ is a nuisance parameter. The full likelihood is

$$L(\gamma,\delta)=\text{pr}\,(\{Y_i,X_i\})=\prod_{i=1}^{n}\text{pr}\,(X_i\,|\,Y_i)\,\text{pr}\,(Y_i\,|\,X_1,\ldots,X_{i-1},Y_1,\ldots,Y_{i-1})$$

$$=\prod_{i=1}^{n}\left(\frac{e^{\gamma+Y_i\delta}}{1+e^{\gamma+Y_i\delta}}\right)^{X_i}\left(\frac{1}{1+e^{\gamma+Y_i\delta}}\right)^{1-X_i}p_i^{Y_i}(1-p_i)^{1-Y_i}, \tag{3}$$

where p_i is as defined in (2) for the urn sampling design. Rearranging, the log likelihood has the form

$$\mathscr{L}(\gamma,\delta)=B_m\gamma+S_m\delta-(m-N_m)\log(1+e^{\gamma})-N_m\log(1+e^{\gamma+\delta})+\log\text{pr}\,[\{Y_i\}\,|\,\{X_i\}].$$

The three statistics B_m, S_m and N_m are jointly sufficient for γ and δ. The parameter γ is a nuisance parameter which represents the baseline probability of success. To obtain a test statistic which is independent of the nuisance parameter it is necessary to condition on both N_m and B_m. In this case the distribution of S_m is

$$\text{pr}\,(S_m=s\,|\,N_m=n,B_m=b)=\frac{e^{\delta s}\sum_C\text{pr}\,[\{Y_i\}\,|\,\{X_i\}]}{\sum_D e^{\delta q}\,\text{pr}\,[\{Y_i\}\,|\,\{X_i\}]}, \tag{4}$$

where q represents the different values of s in the set of permutations D,

$$C=[\{Y_i,X_i\}\colon S_m=s,B_m=b,N_m=n],\quad D=[\{Y_i,X_i\}\colon B_m=b,N_m=n].$$

The distribution is independent of the nuisance parameter. Also, as the test statistic increases, the probability associated with the statistic is increasingly enhanced under alternative hypotheses, i.e. the likelihood ratio is a monotonic function of S_m. This test leads to the p-value of 0·28 for the ECMO study, calculated as outlined in the previous paragraph.

By contrast, Wei's test statistic does not enjoy these properties. In this case there is no conditioning on N_m, and conditioning on B_m is further restricted to the observed sequence of responses $\{x_i\}$. The distribution of the test statistic is

$$\text{pr}[S_m = s \,|\, \{X_i\} = \{x_i\}] = \frac{e^{\delta s} \sum_F \{(1+e^{\gamma})/(1+e^{\gamma+\delta})\}^n \text{ pr}[\{Y_i\}|\{x_i\}]}{\sum_E e^{\delta q}\{(1+e^{\gamma})/(1+e^{\gamma+\delta})\}^n \text{ pr}[\{Y_i\}|\{x_i\}]}, \tag{5}$$

where

$$F = [\{Y_i, X_i\}: S_m = s,\ \{X_i\} = \{x_i\}], \quad E = [\{Y_i, X_i\}: \{X_i\} = \{x_i\}].$$

Under the null hypothesis, the nuisance parameter is eliminated. However, under the alternative, the distribution of S_m depends on both δ and γ in a complicated way. It is instructive to consider extreme special cases. For the case where the baseline response rate is small, we can consider the limiting distribution as $\gamma \to -\infty$. In this case

$$\lim_{\gamma \to -\infty} \text{pr}[S_m = s \,|\, \{X_i\} = \{x_i\}] = \frac{e^{\delta s} \sum_F \text{pr}[\{Y_i\}|\{x_i\}]}{\sum_E e^{\delta q} \text{ pr}[\{Y_i\}|\{x_i\}]}.$$

Under these circumstances, Wei's test is independent of γ and appears to be an appropriate test. However, at the other extreme a very different picture emerges. In this case

$$\lim_{\gamma \to \infty} \text{pr}[S_m = s \,|\, \{X_i\} = \{x_i\}] = \frac{\sum_F e^{\delta(s-n)} \text{ pr}[\{Y_i\}|\{x_i\}]}{\sum_E e^{\delta(s-n)} \text{ pr}[\{Y_i\}|\{x_i\}]}.$$

Now the likelihood ratio is a monotonic function of $S_m - N_m$, and this statistic appears to be more appropriate than S_m. It is this latter situation, i.e. a high value of γ, which is more relevant in the ECMO study.

To obtain a more intuitive picture of what is wrong with using S_m, unconditional on N_m, as the test statistic it is useful to examine different points in the sample space; Wei's statistic appears to rank them inappropriately for computing the p-value. For example, consider the possible outcome $(S_{12} = 10, N_{12} = 11)$, and compare this with $(S_{12} = 6, N_{12} = 6)$. Wei's test ranks the former as more consonant with the hypothesis that A is better than B than the latter, i.e. more significant. Yet the first outcome corresponds to a trial in which the observed response rate on A is lower than B, 91%, $\frac{10}{11}$, versus 100%, $\frac{1}{1}$, while the second outcome suggests A is superior, 100%, $\frac{6}{6}$, versus 83%, $\frac{5}{6}$. This is not unusual, Wei's sample space being of inconsistencies like this. By contrast, in the test which conditions on both B_m and N_m, no such inconsistencies are possible.

Despite this, the conditional test appears to be inefficient, since the conditioning removes the additional information about δ in N_m. That is, the conditional distribution of N_m, given S_m and B_m, is a function of δ. One could construct tests based on S_m and N_m in which the sample space is more appropriately organized than in Wei's test. For example, a referee suggested the use of $T = S_m/N_m - (B_m - S_m)/(m - N_m)$, the difference in the observed response rates, as the test statistic, permuted over all possible values of S_m and N_m. The operating characteristics of this test and the preceding three tests have been computed for a few values of γ and δ, using a sample size of 9, since the computation involved for larger sample sizes is too burdensome; see Table 1. For computational purposes the sample space for T was discretized into 40 equally-spaced intervals between -1 and 1, and for all tests the critical regions were randomized to ensure a size of 0·05 for the purposes of comparability. These calculations were also performed for smaller sample sizes and the results demonstrated similar trends, which provides us with reasonable confidence that the trends apply for larger sample sizes. The results indicate that

Table 1. *Power of various tests*

γ	δ	Test 1	Test 2	Test 3	Test 4
−4·0	1·0	0·06	0·05	0·06	0·06
	2·0	0·08	0·06	0·08	0·08
−2·0	1·0	0·09	0·06	0·09	0·11
	2·0	0·19	0·09	0·17	0·28
0·0	1·0	0·11	0·09	0·12	0·15
	2·0	0·16	0·15	0·20	0·24
2·0	1·0	0·06	0·07	0·07	0·07
	2·0	0·07	0·08	0·08	0·08

Test 1, Wei's test, (5); 2, test based on $S_m - N_m$, given $\{X_i\} = \{x_i\}$; 3, test conditional on N_m and B_m, (4); 4, test based on T, given B_m.

the conditional test based on (4), in column 3 of the table, compares favourably with Wei's test, in column 1, but is not uniformly more powerful. The test based on T, column 4, seems to be substantially more powerful in the middle ranges of γ, and is never worse than the other tests in the ranges studied. Incidentally this last test leads to a p-value of 0·038 in the ECMO study, assuming that permutations leading to degenerate estimates of the treatment difference are not included in the critical regions, i.e. permutations in which all patients receive the same treatment.

3. Discussion

Despite the superiority of T in terms of power it is not clear that it is appropriate to use any test which admits permutations involving values of N_m different from that which occurred in the trial. Along with B_m, N_m is a strong indicator of the information content of the trial regarding δ. Any test which does not condition on N_m will include hypothetical outcomes of vastly different information content. In the ECMO study the realized trial is relatively uninformative about δ, since only one patient received treatment B, yet the unconditional tests make use of hypothetical realizations which would have been more informative had they occurred. That is, a trial in which six patients each are randomized to A and B is much more informative about the treatment difference than one in which eleven receives A and one receives B, regardless of the outcomes. In considering the inference from an individual trial, the information content should be an important determinant of the conclusiveness of the result. By contrast the unconditional tests allow for supposedly conclusive inferences when the only thing that is unusual is the peculiar sequence of randomizations. This point of view is not new (Fisher, 1956, pp. 86–93). To be sure, the observed number of allocations to A in a play-the-winner strategy is influenced by δ and is informative about δ even after conditioning on S_m, and this presumably explains the increased power of the test based on T over the conditional test. Nonetheless, the realized value of N_m in this trial is abnormally large due to chance. Specifically, even under the most extreme alternative hypothesis possible, in which the true response rate on A is 1 and the true response rate on B is 0, the median value of N_m is approximately 9·8, while under the null it is, of course, 6. Therefore we have observed a demonstrably inflated value of N_m in this study, and so any test which does not condition on N_m is anti-conservative to the extent that this inflation contributes to a more extreme p-value. This affects Wei's test, but it is less clear that the influence on T is strong.

In any case, Fisher's reasoning applies in the context of play-the-winner randomization even though there may be some loss of information. If a trial has low information content about the parameter or hypothesis of interest it is inappropriate to make use of hypothetical trials which did not occur to create information where none exists. That is, the use of the likelihood function and sufficiency considerations as the basis for making inferences would seem much more reliable in the context of experiments such as this one. This line of reasoning also suggests that the common practice followed by Wei of conditioning on the observed sequence of responses in order to develop a permutation test that is solely induced by the randomization rather than conditioning on relevant sufficient statistics is unnecessarily restrictive, although it is fair to point out that conditioning on the observed sequence of responses dramatically reduces the computational burden.

There are a number of additional reasons why caution is necessary in interpreting important, small clinical trials of this nature. Medical trials are conducted on real patients, by nature heterogeneous with regard to the chances of successful treatment. They are not homogeneous experimental units, and so the applicability of methods based on strict randomization theory must be viewed with caution. To be sure, the appropriate permutation distribution under the null hypothesis will, on average, be unaffected by patient heterogeneity. However, in any individual small trial, serious imbalances with regard to prognostic factors are quite likely, and therefore are often capable of explaining unusual results. In fact, in the ECMO trial, the patient who failed on treatment *B* had the most extreme values on no fewer than four important covariates (Paneth & Wallenstein, 1985), and was clearly the sickest. In effect, the trial provides no information whatsoever regarding the treatment comparison.

As well as producing serious imbalances relative to covariates, randomized designs can produce serious overall imbalances, as happened in this trial since the design is tilted towards giving the better treatment with higher probability. Wei believes that such play-the-winner strategies are more ethical than designs which allocate to treatments with equal probability. I find this reasoning hard to understand. How do you explain to a patient that if the biased coin lands the wrong way, he or she has to submit to a treatment you believe to be inferior? In fact, when a clinical trial is in progress we are in a period of uncertainty regarding the relative merits of the treatments, even though the evidence is never entirely equivocal. If the evidence is sufficient that it is unethical to administer the inferior treatment, then it is time to stop the trial. If not, then allocation with equal probability is the fairest way. Randomized trials are almost never definitive, so the notion that subsequent treatment will be dictated by the trial results is false.

In summary, my conclusions from the analysis of the ECMO study are diametrically opposed to those of Wei. My view is that this study is an example of why not to use full randomized designs in small clinical trials. Randomization frequently leads to serious imbalances in treatment totals, and imbalances with regard to important covariates (Kalish & Begg, 1985). This can lead to loss of power, as well as serious loss of credibility in the analysis. Consequently, stratification and blocking are essential to ensure a reliable study.

ACKNOWLEDGEMENTS

I am grateful to Les Kalish, Butch Tsiatis, Marvin Zelen and two referees for useful advice, and to Karen Abbett and Terry Crespo for assistance with the manuscript. Part of the research was conducted at the Dana-Farber Cancer Institute, Boston.

REFERENCES

BARTLETT, R. H., ROLOFF, D. W., CORNELL, R. G., ANDREWS, A. F., DILLON, P. W., ZWISCHENBERGER, J. B. (1985). Extracorporeal circulation in neonatal respiratory failure: a prospective randomized study. *Pediatrics* **76**, 479–87.

COX, D. R. (1977). *The Analysis of Binary Data.* London: Chapman and Hall.

FISHER, R. A. (1956). *Statistical Methods and Scientific Inference.* New York: Hafner.

KALISH, L. A. & BEGG, C. B. (1985). Treatment allocation methods in clinical trials: a review. *Statist. Med.* **4**, 129–44.

PANETH, N. & WALLENSTEIN, S. (1985). Extracorporeal membrane oxygenation and the play the winner rule. *Pediatrics* **76**, 622–3.

WEI, L. J. (1988). Exact two-sample permutation tests based on the randomized play-the-winner rule. *Biometrika* **75**, 603–6.

WEI, L. J. & DURHAM, S. (1978). The randomized play-the-winner rule in medical trials. *J. Am. Statist. Assoc.* **73**, 830–43.

[*Received March* 1989. *Revised December* 1989]

Discussion of paper by C. B. Begg

BY RICHARD M. ROYALL

Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland 21205, U.S.A.

When the ECMO study of Bartlett et al. (1985) was published, it was criticized for its failure to provide a convincing test of the null hypothesis (Paneth & Wallenstein, 1985). In fact Ware & Epstein (1985) observed that '... the type I, or false positive, error rate for this design is 0·5!' and undertook a new trial designed to give a one-sided test at the 0·05 level (O'Rourke et al., 1989). In the meantime came a fascinating development: Wei (1988) showed that a null hypothesis significance test could be performed in the Bartlett study after all. Furthermore, his test produced a p-value that was within 0·001 of the 0·05 level; $p_W = 0{\cdot}051$. But now Dr Begg questions the validity of Wei's p-value. He argues for two larger ones, $p_{B1} = 0{\cdot}62$ and $p_{B2} = 0{\cdot}28$, but admits that a fourth test procedure suggested by a referee, which appears to be uniformly more powerful than the other three, yields $p_R = 0{\cdot}038$, even smaller than Wei's.

How should we judge Begg's claim that Wei's p-value is 'inappropriate'? On what basis are we to decide which, if any, of these four tests are 'appropriate' or 'valid'? A significance test requires choosing a test statistic T and determining the probability distribution of T under the null hypothesis. For an observed value of the test statistic, t_{obs}, the p-value is the probability of T's taking a value that large or larger, if the null hypothesis were true: $\text{pr}(T \geqslant t_{\text{obs}})$.

Now it is widely agreed that the probability model used to calculate the p-value should represent accurately the sampling procedure, including any use of deliberate randomization, that generated the observations. Thus Begg reasons that Fisher's exact test is clearly invalid in the ECMO study, because the various permutations of treatments and responses are not equally probable under the randomization scheme that was actually used. It is also widely agreed that T is an appropriate test statistic only if 'the larger the value of t, the stronger the evidence of departure from H_0 of the type it is desired to test' (Cox & Hinkley, 1974, p. 66).

An obvious problem with Wei's tests is that his choice of $T = S_{12}$ as the test statistic is inappropriate. As Dr Begg points out, the other outcome giving $S_{12} = 11$, in which all twelve patients are assigned to treatment A, has not only all the successes but also the failure occurring under A. It is clearly weaker evidence of departure from H_0 than the observed sequence, where the failure occurs under B, and its probability should not be included in the p-value. This correction reduces Wei's p-value from 0·051 to 0·038.

On the first criterion for test validity, accurate representation of the experiment, both Wei's and Begg's tests are deficient. The number of patients randomized in the ECMO study was not

fixed in advance, but was itself random. As it happened, randomization stopped after ten patients, not twelve. The abstract of the paper by Bartlett et al. (1985) is misleading on this point, but their discussion on page 484 is clear: 'The stopping rule, determined in advance, was to stop the randomization whenever ten balls of one type were added, and then to continue using only the treatment that gave the better results.' The subsequent publication of Cornell, Landenberger & Bartlett (1986) confirmed this, reporting that randomization did stop after the tenth patient.

Thus the probability of the twelve observations is not 0·038, as calculated by Wei and Begg, but $\frac{1}{2}\times\frac{1}{3}\times\frac{3}{4}\times\frac{4}{5}\times\ldots\times\frac{10}{11}\times u$, where u is the unknown probability of deciding, after randomization stopped, to include the two additional nonrandomized patients in the study. For example, if the rule was simply to continue with the urn scheme so long as patients were assigned to the winning treatment, the study would have stopped when it did because the thirteenth ball drawn was a B, an unacceptable treatment assignment. Then we would have $u=\frac{11}{12}\times\frac{12}{13}\times\frac{1}{14}$, and the probability of the observed treatment sequence would be 0·003. Or if the rule had simply required that after a winning treatment was declared, two more nonrandomized patients should be observed, then $u=1$ and the probability would be 0·045.

Although we know what rule was used to assign treatments to patients, we do not know what rule was used to decide when to stop accepting patients in the ECMO study. Thus we, i.e. Wei, Begg and I, cannot calculate even the probability of the observed data, and we cannot perform a valid test of significance. If we could justify ignoring the two nonrandomized observations, we could proceed, but would still encounter the problems described below.

For the sake of discussion, pretend from now on that we learn from those who conducted the trial that the second stopping rule described above was used; the number of nonrandomized patients to be included was fixed at two before the study began. In this case any reasonable one-sided test statistic shows that, for the observed sequence of successes and failures, the observed sequence of treatment assignments, Y_{obs}, is the strongest possible evidence of departure from H_0, so that the p-value is just $\text{pr}(Y_{\text{obs}})=0{\cdot}045$. Is this 'valid'? It is surely correct in the mathematical sense, i.e. it is probabilistically valid. But so are other candidates, such as $\text{pr}(Y_{\text{obs}}|E)$, where E is any set of outcomes containing Y_{obs}. In fact a version of the conditionality principle is often interpreted to mean that probabilistic validity is not enough, a p-value can be probabilistically valid and yet not be an appropriate measure of the evidence against H_0. To be inferentially valid as well, it must be 'properly' conditioned. When is a test properly conditioned? This question has never been answered in any generality. Apparently the answer is not to be found in considerations of size and power, since a 'properly' conditioned test can be less powerful than an unconditional one (Cox, 1958). In fact Dr Begg argues that the most powerful of the four tests he considers, i.e. the same test which, under a correct model for the random treatment assignments and the present assumptions about the stopping rule, yields the above p-value, 0·045, is inappropriate because it is not properly conditioned.

Dr Begg cites the attractive heuristic arguments for conditioning, and these suggest that the p-value should be conditioned on both (i) the number of patients randomized in the trial, $M=10$, and (ii) the number randomized to treatment A, $N=9$. Because the randomization test treats the sequence of successes and failures, $x=(1\ 0\ 1\ 1\ \ldots 1)$, as given, conditioning on M and N fixes the total number of successes as well as the number of patients randomized to each of the treatments.

Now under the randomized play-the-winner rule used, with randomization stopping when ten balls of either type have been added, the observed sequence of treatment assignments, Y_{obs}, and its complement $Y^{c}_{\text{obs}}=(0\ 1\ 0\ 0\ldots 0)$ are the only two possible results with $M=10$ patients randomized. If the first ten patients were all randomized to A, there would be only nine A balls added to the urn, so randomization would continue and M would be at least eleven; if one B occurred but not with the second patient, M would be at least 12, etc. The additional condition $N=9$ then implies that the only possible treatment sequence is the one observed: $\text{pr}(Y_{\text{obs}}|M=10, N=9)=1$.

Is this p-value 'properly' conditioned, and thus inferentially 'valid'? I think not. It implies, quite improperly, that the data are utterly worthless as evidence against the hypothesis of no treatment difference. It might be more reasonable to condition, not on M and N, but on M and

max $(N, M-N)$, to obtain a one-sided p-value of 0·50. Is this one valid? Maybe. But a two-sided test would double it, again denying that we have any evidence against the hypothesis of no treatment difference.

The arguments for conditioning are only heuristic. Those who do not find them persuasive might still maintain that the unconditional p-value, 0·045, is appropriate. Against this view we might observe that this value depends strongly and inappropriately on the order in which the observations were made. The patient who received treatment B and died was the second one randomized. If he had been the kth the p-value would have been $1/(11k)$. That is, if the same twelve patients had received the same treatments with the same results, the unconditional randomization test would indicate 'very strong evidence, $p \leqslant 0{\cdot}01$' or 'only weak evidence, $0{\cdot}05 < p \leqslant 0{\cdot}10$' depending on whether the patient who received B and died happened to be the last one randomized or the first.

Do we get 'more reasonable' p-values if we follow Begg in introducing a probability model for the successes and failures, x's, representing them as Bernoulli trials with success probabilities determined by which treatment is used? This model has the same form as (3), but with definitions of the p_i changed to reflect the design actually used. The jointly sufficient statistics are (S^*, N^*, M^*, B^*), where S^* is the number of successes and N^* the number of trials on treatment A, M^* is the total number of trials, and B^* is the total number of successes. If we agree that the only valid test statistics are functions of the sufficient statistics, as suggested by the sufficiency principle, then we find that any plausible candidate, conditioned on M^*, N^* and B^*, again brings us back to $\text{pr}\,(S^* = 11 \,|\, M^* = 12, N^* = 11, B^* = 11) = 1$.

Although the model (3) fails to generate a satisfactory p-value, it does provide a useful representation of the evidence in the ECMO trial, namely the likelihood function, $\theta_A^{11}(1-\theta_B)$. This function shows (Edwards, 1972, p. 30) that these data are strong evidence for large versus small values of the treatment A success probability; for instance, $\theta_A = 0{\cdot}8$ is better supported than $\theta_A = 0{\cdot}4$ by a factor of $(0{\cdot}8/0{\cdot}4)^{11} > 2000$. It also shows that the evidence concerning θ_B is weak. Here $\theta_B = 0{\cdot}4$ is better supported than $\theta_B = 0{\cdot}8$, but only by a factor of three.

This example seems to me to illustrate three important but widely ignored lessons that were contained in Birnbaum's landmark paper (1962) as follows.

(i) The intuitive concept of conditioning, when properly formulated in the conditionality principle, is not a rule for choosing an 'appropriate' conditional sample space to use in calculating p-values, biases, etc. It is essentially equivalent to the likelihood principle, which implies (ii).

(ii) Statistical methods that depend on how probabilities are distributed over sample space points that have not occurred, e.g. significance tests, are generally inappropriate for interpreting data as evidence. Thus we should not be surprised to find situations like the present one where our best efforts to produce and rationalize a satisfactory p-value have failed.

(iii) If you want to represent and interpret the data as evidence, look at likelihoods.

The fact that it is primarily the Bayesians who have been trying to teach these lessons, e.g. Cornfield (1966) and Berger & Wolpert (1988), does not excuse the rest of us for not having learned them.

Additional references

Berger, J. O. & Wolpert, R. L. (1988). *The Likelihood Principle*, 2nd ed. Hayward, California: Institute of Mathematical Statistics.

Birnbaum, A. (1962). On the foundations of statistical inference. *J. Am. Statist. Assoc.* **57**, 269–306.

Cornell, R. G., Landenberger, B. D. & Bartlett, R. H. (1986). Randomized play the winner clinical trials. *Comm. Statist.* A **15**, 159–78.

Cornfield, J. (1966). Sequential trials, sequential analysis and the likelihood principle. *Am. Statistician* **20**, 18–23.

Cox, D. R. (1958). Some problems connected with statistical inference. *Ann. Math. Statist.* **29**, 357–72.

Cox, D. R. & Hinkley, D. (1974). *Theoretical Statistics.* London: Chapman and Hall.

Edwards, A. W. F. (1972). *Likelihood.* Cambridge University Press.

O'ROURKE, P. P., CRONE, R. K., VACANTI, J. P., WARE, J. H., LILLEHEI, C. W., PARAD, R. B. & EPSTEIN, M. F. (1989). Extracorporeal membrane oxygenation and conventional medical therapy in neonates with persistent pulmonary hypertension of the newborn; a prospective randomized study. *Pediatrics* **84**, 957–63.
WARE, J. H. & EPSTEIN, M. F. (1985). Extracorporeal circulation in neonatal respiratory failure: a prospective randomized study. *Pediatrics* **76**, 849–51.

[*Received August* 1989. *Revised February* 1990]

Discussion of paper by C. B. Begg

BY L. J. WEI

Department of Statistics, University of Wisconsin, Madison, Wisconsin 53706, U.S.A.

Dr Begg raises two important and intriguing questions: 'Should we use a data-dependent treatment allocation rule in a clinical trial?' and 'How should we analyse the results from a trial which was done with an adaptive design?'. Dr Begg's answer to the first question is no. He feels that it is hard for a medical investigator to explain to a patient that if the biased coin lands the 'wrong' way, he or she has to submit to a treatment the investigator thinks to be inferior. Now, suppose that we use a stratified 50/50 allocation with some kind of sequential monitoring scheme for an ECMO trial. Furthermore, assume that we do not mask the treatment assignment rule or the results from the interim analyses performed during the trial to the incoming patients. At some point in the study, suppose that the accruing data show an advantage for the ECMO though not a significant enough of an advantage to terminate the trial early. Then, would it be very difficult for the investigator to tell an eligible patient for the study that he or she has to be assigned to the control group in order to avoid imbalances in treatment totals and imbalances with regard to covariates? Masked 50/50 allocation rule and sequential stopping protect trials from ethical criticisms. Adaptation on outcome can also be protected by masking (Louis, 1985).

As well put by Bather (1985), for a clinical study on human beings, there is a conflict between the interests of an individual study patient and the advancement of science. For the ECMO study conducted in Michigan, this conflict was rather prominent. A nondeterministic adaptive design, such as the randomized play-the-winner rule, seemed to be a good compromise for the conflict in this situation. However, in retrospect it might have been better to have run the Michigan study with more than one ball of each type for the first allocation. General issues on adaptive designs have been extensively discussed by Simon (1977), Bather (1985), Armitage (1985), and recently by Ware (1989). In my opinion, the implementation of a data-dependent design may be more feasible in single centre short-term trials than in multi-centre trials controlled by a group of investigators.

As to the second question, the test proposed by Wei (1988) was constructed under the so-called randomization model. That is, those twelve study patients in Ann Arbor, Michigan, might not have been drawn from a well-defined population according to some random sampling scheme. Now, let us suppose that those patients were obtained from a simple random sampling scheme. Then, as indicated by Dr Begg, Wei's test is unnecessarily restricted by conditioning on the observed sequence of responses, and consequently is less powerful than his unconditional test T. With an adaptive design under the above population model assumption, the likelihood function is proportional to that for the usual two-sample problem of proportions. Therefore, a Bayesian statistician or other believer in the likelihood principle would not care about the allocation rule used in the trial. It is, however, less clear how to analyse the ECMO data from a frequentist point of view. By conditioning on the two marginal sums in the 2×2 table, one can get rid of the nuisance parameter. However, as Dr Begg pointed out, with an adaptive design the total number of patients treated by ECMO carries strong evidence about the relative merits of the two treatments.

Inferences about the odds ratio based on his conditional procedure (4) may not be very efficient. Indeed, the unconditional test T has been shown to be much more powerful than its conditional counterpart in an extensive numerical study (Wei et al., 1990). For instance, under the alternative $p_A = 0{\cdot}9$ and $p_B = 0{\cdot}1$, with a Type I error probability 0·05 and $m = 10$ and with the design used in the Michigan ECMO trial, the exact powers for the unconditional and conditional tests are 0·72 and 0·22, respectively. For a less drastic alternative, say, $p_A = 0{\cdot}8$ and $p_B = 0{\cdot}3$, the corresponding powers are 0·34 and 0·09. Another advantage of using unconditional procedures is that exact confidence intervals for various parameters, e.g. the difference between the two proportions, can be constructed. It seems rather difficult to do so through the conditional method based on (4) except for the odds ratio parameter.

Since there was only one patient assigned to the control group in the Michigan trial, Dr Begg claimed that the 'information content' from this experiment was low, and one simply could not say much about the difference between the ECMO and the control. The unconditional test T was, therefore, guilty of producing a significant result. Dr Begg suggested use of a more conservative test such as a conditional test for this case. However, it was not clear why the test based on (4) was preferred over other conditional tests, e.g. Wei's test with conditioning further on N_m. Furthermore consider a hypothetical trial done with the play the winner rule for twelve patients. Suppose that it ended up with 8 A's and 4 B's. Does this contain enough 'information content' about the treatment difference? Is it now all right to use a more powerful test procedure such as an unconditional one? Is there an objective way to quantify the so-called 'information content' from the data so that a 'correct' inference procedure can be chosen? The unconditional procedure, such as T, is based on the likelihood function and sufficient statistics. An efficient algorithm is available to compute the distribution of T for any combination of p_A and p_B with adaptive designs (Wei et al., 1990). I am not convinced that, in general, one should give up a powerful procedure such as the unconditional test T in analyzing the data from a trial done with an adaptive design. However, I do appreciate Dr Begg's interesting comments.

Additional references

ARMITAGE, P. (1985). The search for optimality in clinical trials. *Int. Statist. Rev.* **53**, 15–24.

BATHER, J. A. (1985). On the allocation of treatments in sequential medical trials. *Int. Statist. Rev.* **53**, 1–13.

LOUIS, T. A. (1985). Discussion of the papers by Bather and Armitage. *Int. Statist. Rev.* **53**, 28–31.

SIMON, R. (1977). Adaptive treatment assignment methods and clinical trials. *Biometrics* **33**, 743–9.

WARE, J. H. (1989). Investigating therapies of potentially great benefit: ECMO (with comments). *Statist. Sci.* **4**, 298–340.

WEI, L. J., SMYTHE, R. T., LIN, D. Y. & PARK, T. S. (1990). Statistical inferences with data-dependent treatment allocation rules. *J. Am. Statist. Assoc.* **85**, 156–62.

[*Received January* 1990]

Discussion of paper by C. B. Begg

BY THOMAS R. FLEMING

Department of Biostatistics, University of Washington, Seattle, Washington 98195, U.S.A.

Dr Begg provides well stated arguments against the use of the randomized biased coin rule, and arguments for using considerable caution when interpreting results from small clinical trials. His paper also raises important criticisms of Wei (1988) who used a biased coin design in which $S_m = X_1 Y_1 + \ldots + X_m Y_m$, the number of successes on treatment A, was used as the summary test statistic for the hypothesis of equal efficacy of treatments A and B, with the distribution of S_m based on the permutation distribution of treatment assignment conditional on the sequence of responses $\{X_j\}$ and taking into account the biased coin design. This approach does lead to some undesirable orderings of the sample space.

Dr Begg suggests a valid although somewhat inefficient approach by proposing to use the distribution of S_m, accounting for treatment assignment by the biased coin design after conditioning both on the number assigned to treatment A, $N_m = \Sigma_j Y_j$, and on the total number of successes, $B_m = \Sigma_j X_j$. He established (B_m, S_m, N_m) to be sufficient for (γ, δ) in a logistic model where δ represents treatment effect and γ is a nuisance parameter. It is clear that the information in B_m essentially relates to γ. However, due to the biased coin design treatment assignment, N_m as well as S_m carries considerable information about δ. Agreeing with the author's statement concerning 'sufficiency considerations as the basis for making inference', I have preferred an alternative approach based on $T = S_m/N_m - (B_m - S_m)/(m - N_m)$ as a test statistic, conditioning on B_m and permuting over all possible values of S_m and N_m while taking into account the biased coin design. The author's nice computations summarized in Table 1 validate the increase in power provided by T through its use of the information in N_m.

Dr Begg's reference to Fisher (1956, pp. 86–93) to justify the less efficient approach defined by conditioning on both N_m and B_m could be debated. Unlike in classical randomized designs where N_m is not informative and often is deterministic, one could use N_m alone in a biased coin design to formulate an inefficient but valid test procedure for $H: \delta = 0$ which actually would be consistent against $H: \delta > 0$. It is not clear Fisher intended that his arguments for conditioning on margins should be applied to this setting.

In clinical trials, type II errors as well as type I errors have considerable detrimental effects on our ability to identify and establish more effective treatment approaches. For this reason, we need valid and efficient statistical procedures which are thoughtfully applied. In the ECMO trial data, the test based on T yielded a p-value of 0·038 while the test conditioning on both N_m and B_m yielded a much higher p-value of 0·28. The fact that the single patient receiving treatment B also had the poorest levels of several prognostic factors does not support an argument favouring the less efficient test. Rather, it supports the need for careful thoughtful interpretation of clinical results and it argues strongly against implementing designs which allow very small sample sizes. The author's concerns about the interpretation of results from small trials due to problems induced by patient heterogeneity are quite relevant.

The discussion concerning use of T rather than an approach involving conditioning on both N_m and B_m probably is moot in the setting of medical trials. I strongly concur with the author's position challenging the ethics and practicality of the biased coin design. In addition to allowing absurdly small sample sizes, the biased coin design requires physicians to present a randomized treatment assignment to patients who might have been informed that the therapeutic index of one of the therapies appears to be better. The principle of equipoise, which justifies the use of randomization, requires the clinical investigator acknowledge a state of genuine uncertainty among the medical community regarding the comparative therapeutic merits of each arm in a trial. When equipoise no longer exists, the trial should be stopped. The biased coin design also allows increased risk of bias in the randomization process. Consciously or not, physicians for example might randomize patients with better prognosis at a time when one regimen has had a string of successes, leading to systematic prognostic factor imbalances between the treatment groups. Indeed, this could explain why the final 10 patients on the ECMO trial had better prognoses.

The biased coin rule does not provide a desirable approach to the allocation of patients to treatments in medical trials.

[*Received February* 1990]