Comment

David J. Spiegelhalter and Laurence S. Freedman

We are grateful for the opportunity to contribute to the discussion of this important paper, which is all the more impressive for being written by someone who is not already identified with "the faith." Professor Breslow describes a number of exciting developments that are at last bringing Bayesian techniques into mainstream biostatistics, and we would like to comment on clinical trials. Here the increasing interest in Bayesian methods is a reflection of dissatisfaction with the established Neyman–Pearson methodology, and where the Bayesian analysis appears to provide both insight into the interpretation of frequentist procedures and a qualitative improvement in the communication of issues in the design and monitoring of trials.

The author describes in Sections 5 and 6 the use of Bayesian techniques in bioequivalence studies and sequential clinical trials. The inferential problem underlying each of these applications can be characterized by the superimposition of a distribution for the parameter of interest on a domain in which regions of different clinical implications have been displayed. Figure 1 is taken from Freedman and Spiegelhalter (1989), and shows how the distribution can be summarized by the areas lying in each of three regions.

In bioequivalence testing, the range of equivalence is generally taken to be ±20%, and, as described in Section 5, "equivalence" is declared if $p_C + p_E < .05$. When a uniform prior is assumed for δ, this procedure is equivalent to the symmetric confidence interval procedure of Westlake (1976) (although the Bayesian interpretation is much simpler). O'Quigley and Baudoin (1988) show how other frequentist proposals for bioequivalence testing are interpretable as analysis of posterior distributions, which reveals, for example, the rather unintuitive nature of the Hauck and Anderson (1986) method.

In general clinical trials there is a great advantage in using pictures such as Figure 1 to explain to clinicians the essential difference between a treatment difference that they would like to have in order to recommend routine use of a new treatment (i.e., $\delta > \Delta_E$) and a difference they think it is reasonable to expect. In standard works on the design of clinical trials, there is often great confusion between differences that are *desired* and those that are *expected*, with both being cited as a basis for deciding an alternative hypothesis. A simple picture clarifies the issues and can be used both before trial is started, and while sequentially monitoring data.

Before a trial, it seems reasonable to obtain the best possible assessment of the likely treatment difference, either from past trials or from careful questioning of trial participants, and to summarize that opinion as a prior distribution superimposed on an assessment of the range of clinical equivalence, where this range takes into account the possible side-effects and other secondary disadvantages of the treatments. The juxtaposition of belief upon demands can be used for two types of reassurance. First, neither p_C nor p_E should be very large, otherwise it would appear unethical to randomize patients when there was already substantial belief in the clinical superiority of one or other treatments. Second, the prior distribution can be used to assess the predictive power of obtaining a convincing result, which is essentially obtained by averaging the standard power curve with respect to the prior plausibility of each value of δ. Applications of such analyses are reported in Spiegelhalter and Freedman (1986, 1988), who found clinicians quite willing to express their judgments and actually surprised that they had not previously been asked to do so.

Once a trial is underway, the updated posterior should be monitored for the ethical basis of randomization; the likelihood could also be monitored since this will presumably be transmitted to regulatory authorities and journal editors. Monitoring the tail areas p_C and p_E appears appropriate, and it is possible for the range of equivalence to be adapted during the trial provided it is done by an adverse event committee ignorant of the current results on the primary outcome measure. Professor Breslow illustrates an alternative input to the decision of whether to stop: the predictive probability of achieving a firm conclusion were the trial to continue. In fact, it is remarkable that this measure is so rarely used (one example being Frei, Cottier, Wunderlich and Ludin, 1987), as Armitage (1988, 1989) has shown that a Bayesian derivation is unnecessary. If at an interim stage δ is estimated by d_0, say, then the final estimate d_N will generally depend on d_0 and the unobserved estimate d_u to be based on the future observations. Then $d_u - d_0$ will be at least approximately independent of δ, and this will

David J. Spiegelhalter works for the Biostatistics Unit of the Medical Research Council, 5 Shaftesbury Road, Cambridge CB2 2BW, United Kingdom. Laurence S. Freedman works in the Biometry Branch of the National Cancer Institute, Bethesda, Maryland 20892.

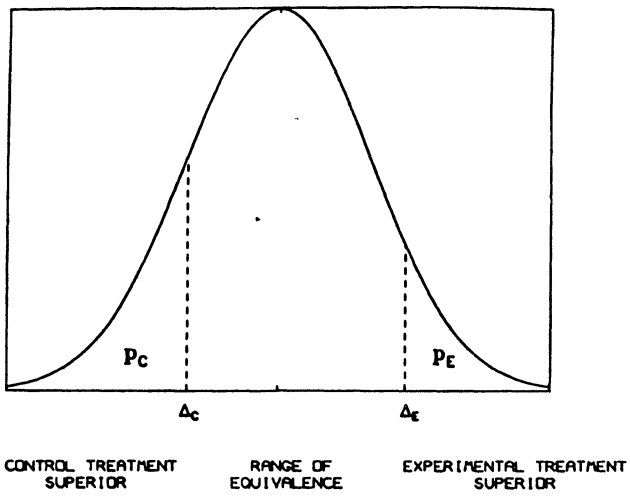

FIG. 1. *Distribution of parameter δ summarizing treatment difference, superimposed on regions of differing clinical implication.*

allow us to calculate the unconditional chance of d_N achieving "significance." When applied to the example in Section 6, Armitage (1988) obtained a predictive probability of .948 in contrast to our estimate of .946 using much more complex techniques. Armitage (1989) does, however, express reservations about this predictive approach, arguing that decisions to stop should be based on current opinion, and that concentration on attaining future "significance" understates the value of a "nonsignificant" result.

The suggestion of monitoring a trial using posterior tail areas will inevitably lead to accusations that we are essentially advocating the use of "naive" significance tests that do not take into account the repeated looks at the data, and so can "sample to a foregone conclusion." Breslow cites Cornfield's observation that if one is concerned with this behavior, then it must reflect some belief in the correctness of the null hypothesis. Cornfield's approach was thus to put a lump of probability on the null and smear the rest over the entire parameter range; such a mixture prior should be a reasonable expression of a belief that either the new treatment will have some effect, or there is some mechanism that will make it irrelevant. Alternatively, reasonable prior skepticism about the existence of large treatment differences can be expressed by a smooth distribution centered on the null hypothesis. Such a prior could be considered equivalent to an implicit sample of patients whose observed treatment difference is zero, and hence forms a "handicap" that the data must overcome in order to convince us that extreme differences observed early in the trial are not just lucky events. Stopping can be based on the posterior tail areas, with the prior opinion providing a degree of shrinkage toward the null hypothesis, which will impose the type of conservatism sought in frequentist sequential analysis.

Figure 2 is taken from Freedman and Spiegelhalter (1989) and shows the stopping boundaries obtained in a trial of 200 patients with 5 interim analyses, in which the classical boundaries of Pocock and O'Brien and Fleming are contrasted with those derived from a Bayesian stopping rule based on a prior equivalent to 22 pairs of patients who on average showed no treatment difference. The Bayesian stopping rule assumes a zero range of equivalence and recommends stopping if either of the posterior tail areas p_C or p_E are $< .025$. It can be seen that the Bayesian stopping boundaries, expressed in terms of the naive fixed-sample Z-statistic, are fairly conservative at the start of the trial and then approach the nominal value of 1.96 as the data overwhelm the prior. Thus the qualitative behavior of frequentist sequential stopping rules can be imitated by a Bayesian procedure based on an explicit declaration of prior skepticism. (The prior can even be juggled until the Bayes procedure has exactly the correct size of 5%, but that appears rather sacreligious; alternatively, specific sampling characteristics might be achieved by adjusting the number of looks at the data.)

The domain of clinical trials appears to be distinguished from others described by Professor Breslow by the fact that the Bayesian approach is not intended to provide a more sophisticated statistical analysis. Rather, explicit representation of beliefs and demands provides a qualitative improvement in the tools available in the complex collaborative process of designing and carrying out a trial. We and others have found it extremely valuable to elicit beliefs and demands before

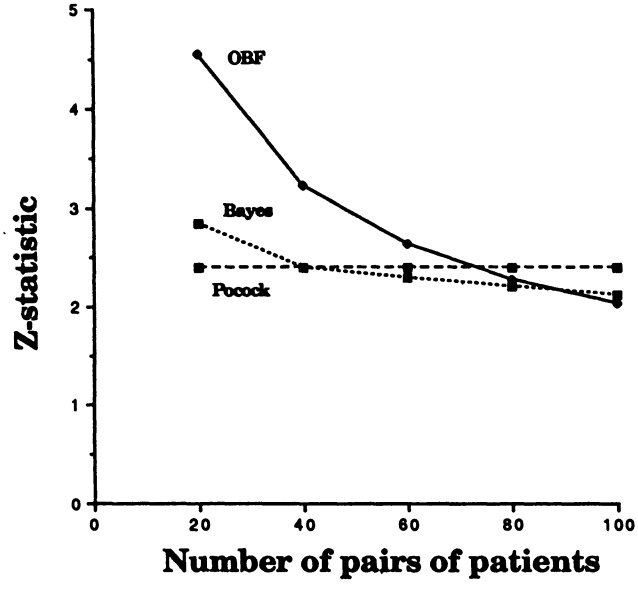

FIG. 2. *Stopping rules for two classical and one Bayesian procedure, where Z is the standard fixed-sample classical test-statistic for zero treatment difference.*

the start of a trial, even if they are only to be used informally. Contrasting prior beliefs with accumulating data can provide a means of identifying over-optimistic expectations, but this can only take place if those expectations have been explicitly recorded. Sometimes prior expectations can be dead on: the Beta-blocker heart attack trial (BHAT) was designed around an expected 28% drop in mortality derived from previous studies (BHAT Research Group, 1984); after 3837 patients had been randomized the observed improvement was exactly 28%! (See BHAT Research Group, 1987.) It would be rather optimistic to think that all prior judgments will be so accurate, especially when similar studies have not been carried out and one is reliant on purely subjective opinion.

An encouraging sign is the willingness of established researchers in clinical trials to take the Bayesian approach seriously: Armitage (1988, 1989) illustrates many of the points made in this discussion, while Pocock and Hughes (1990) provide details of Bayesian estimation following early termination of a trial in order to overcome the excessive bias of the standard estimate. We feel confident that slow but steady progress toward Bayesian design and monitoring will continue to be made in the future, and feel sure that Professor Breslow's paper will help in this regard.

Comment

M. Zelen

Bayesian methods have influenced our thinking about the foundations of statistical inference but have not enjoyed widespread popularity in applications. Professor Breslow's paper is a fine summary of some of the settings in which Bayesian methods have been applied with success to real data problems. The paper serves as a reminder that Bayesian methods are beginning to be utilized in the analysis of data arising in the health sciences. I would expect this trend to increase as Bayesian software becomes more available. However, even with access to appropriate software, the increased use of Bayesian methods will be dampened by the sensitivity of these methods to model specificity. A widely prevailing view is that inferences should rely on reasonably robust procedures. As a result, one is likely to see Bayesian methods applied to situations which have insufficient data to make frequency-based inferences or to situations which directly arise from Bayesian considerations. It is this latter remark on which I will comment further.

The Bayesian philosophy seems particularly appealing and appropriate in case–control settings. This methodology is aimed at inferring whether exposure to a potential causal factor is associated with the incidence of a particular disease. Starting from a collection of cases and controls, one must infer if the case exposure to a causal factor is "unusual" when compared to controls. One can use the information from a control group to calculate the posterior distribution of exposure. In many instances, there may be so much prior information available about exposure of a population (e.g., lifestyle habits of smoking and drinking, etc.) that the limited information available from a sample of controls may generate a posterior distribution of exposure which is nearly the same as the prior distribution. In such situations, one can carry out an analysis of the cases and their exposure without even generating data on a control group. The frequentist view of case–control studies does not permit studies without controls. This represents a serious shortcoming of the frequentist methodology for case–control studies. To cite an extreme example, suppose one has a potential causal factor which is rare in the population, yet the available cases all have been exposed to the causal factor. It would be ludicrous to carry out a case–control study. Yet this is what the frequency point-of-view dictates.

The frequentist model for case–control studies is that random samples are drawn from a population of cases and controls. In practice, this assumption is unrealistic and is rarely met in practice. Data on cases are usually drawn from hospitals, registries or whatever data collection mechanism would yield a convenient set of cases. Controls may be gathered in a variety of ways, but often it is not at all clear if the controls are from the same population as the cases. Various matching techniques are used to attempt to make cases and controls comparable, but there is no way to account for unknown factors which can influence

M. Zelen is Professor of Statistics and Chairman of the Department of Biostatistics at Harvard University. His mailing address is Department of Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston, Massachusetts 02115.