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.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *Statistical Science*. www.jstor.org