Comment

C. Jennison

I would like to congratulate Professor Breslow on his thought-provoking paper. The many and varied biostatistical problems discussed serve to illustrate the advantages, and sometimes the limitations, of the Bayesian approach. The paper is typical of recent trends in that it promotes Bayesian methods by demonstrating their application to real and often complex problems, a refreshing change from abstract justifications from first principles of the Bayesian paradigm. While I do not subscribe to an all-embracing Bayesian philosophy, I am happy to recognize the advantages of Bayesian methods in many specific instances. But, for readers like myself it is essential that the debate addresses real examples; we need to be able to compare other methods of analysis and to assess the suitability of assumed prior distributions. This latter point is of prime importance since doubt about the validity of an informative prior weakens the credibility of any analysis that depends strongly on its prior.

PROBLEMS WITH MANY PARAMETERS

The examples in Section 3 of the paper demonstrate that the Bayesian approach has much to offer in problems with many parameters, be they parameters of interest or nuisance parameters. Introducing a joint distribution for parameter values is a useful way of incorporating prior information or expectations, and the examples of Section 3 provide convincing illustrations of this technique. Nonetheless, I am pleased to see in the closing paragraph some healthy skepticism about the use of arbitrary priors and warnings about their possible ill-effects.

In some problems, similar models have been developed outside the Bayesian school, for example random effects models are commonly used in classical analysis of variance. However, the interpretations of such models that different people are willing to recognize can vary substantially. I am reminded of my own experience in connection with work on the estimation of treatment effects in agricultural field trials in the presence of fertility trends (Green, Jennison and Seheult, 1985). We proposed treatment estimates which could be derived either as least-squares estimates in the presence of a fixed but unknown smooth trend or as general least-squares estimates in the presence of a random trend; the random trend model has the advantage that it facilitates calculation of standard errors for estimates of treatment differences. My own view is that there is no real difference between an unknown fixed trend and an appropriately defined random trend, but these appear to be quite separate concepts to many agricultural statisticians. Unfortunately, the distinction is important, and the difficulty of calculating standard errors has been a major factor in the failure of "neighbour adjustment" methods to challenge classical randomization methods for the analysis of field trial data.

It is interesting to note the willingness of researchers working in problems with very many parameters to embrace the Bayesian approach. Prior distributions for regression functions have been proposed in the context of nonparametric regression (see, for example, Silverman, 1985); Bayesian models have been widely adopted in statistical image reconstruction (see Geman and Geman, 1984; Besag, 1986). In both these areas, the Bayesian paradigm appears to be alone in offering hope of progress toward formal inferences in the presence of so many related parameters.

BIOEQUIVALENCE

The issue of bioequivalence testing is clearly an important one with major financial implications for pharmaceutical companies. While not disputing the Bayesian approach to this problem (apart from possible doubts regarding the origins of prior distributions), I would challenge the suggestion that suitable "frequentist" procedures cannot be found. I believe the difficulties alluded to arise from unfortunate and rather confused early formulations of the problem.

The repeated-sampling requirements of a frequentist testing procedure are clear. It should accept bioequivalence with high probability when two compounds have the same response distribution, and it should reject bioequivalence with high probability when response distributions differ by some specified amount. For many types of response, standard methods are available to construct a test meeting such requirements. A problem that has arisen is a semantic one, concerning the naming of hypotheses. To protect the consumer, one must guard against wrongly accepting bioequivalence, and it is the probability of such an error that will be the main concern of a regulatory body. Under the usual convention, the "more serious" form of error is called *type I* and this

C. Jennison is a Senior Lecturer in the School of Mathematical Sciences at the University of Bath, Bath BA2 7AY, England.

should correspond to wrongly rejecting the null hypothesis; in order to achieve this in bioequivalence testing, the null hypothesis must be set at a point of nonequivalence, i.e., a "non-null" null hypothesis. If this notion is unacceptable, the simple solution is to retain a null hypothesis of exact equivalence and express the bound on the probability of wrongly accepting bioequivalence as a power requirement. From an operational point of view, the naming of hypotheses is immaterial and the properties of a test can be represented by its operating characteristic, a graph of the probability of accepting bioequivalence against treatment difference. Such a graph should suffice to demonstrate the acceptability of a good frequentist method and, in view of Professor Breslow's experience, I would suggest its inclusion as standard practice in protocols for bioequivalence testing.

SEQUENTIAL CLINICAL TRIALS

Most of my reservations about Professor Breslow's paper concern the discussion of sequential clinical trials. The initial example of a trial in which the study committee changed the end point in the middle of the study is interesting, but it should usually be possible to avoid such problems through detailed discussion at the planning stage. A related problem is the shifting during the course of a study of the point at which two treatments are deemed to be equivalent with respect to the primary response, for instance as a result of observed outcomes of a secondary response; the repeated confidence interval approach of Jennison and Turnbull (1989) is designed to allow frequentists sufficient flexibility to cope with such mid-study changes. Professor Breslow's remarks about the difficulty of evaluating operating characteristics of sequential tests are outdated since direct calculation by repeated numerical integration is straightforward and fast, both for group sequential designs and for quite fine discretizations of continuous time procedures; however, given the large number of papers and books on analytic approximations to these quantities, one might be forgiven for assuming that no simple direct solutions exist! I am surprised that the suggestion to discard observations in order to comply with a group sequential design is given such prominence. Pocock (1977) noted the robustness of his group sequential designs to small fluctuations in the group sizes and, for larger fluctuations, the Lan and DeMets (1983) "error spending function" provides an exact adjustment; also, Jennison (1987) has adapted the error spending function approach to one-sided tests.

Before addressing the troublesome question of whether inferences on termination should depend on the stopping rule used to collect data, I think it is important to consider the choice of the stopping rule itself. The motivation behind any sequential design is the need to weigh the cost of collecting further data against the reduction in the probability of an incorrect decision gained thereby. If costs of sampling and loss functions for incorrect decisions are specified, a formal decision problem results but, even if this is not done explicitly, the notional existence of a common underlying framework suggests that good frequentist and Bayes stopping rules should not be too dissimilar. This idea is confirmed by the complete class theorems for sequential tests of Brown, Cohen and Strawderman (1980) which state that all "admissible" frequentist rules are Bayes for some combination of prior and loss function. This equivalence is exploited by Eales and Jennison (1990), who use dynamic programming solutions of Bayes decision problems to compute optimal frequentist tests. In practice, formal decision theory is rarely, if ever, used to determine stopping rules; to specify a loss function as well as a prior appears to be too daunting a task. Perhaps some progress could be made by approaching the problem from the opposite direction: it would certainly be an interesting exercise to determine a decision problem for which a proposed sequential test is optimal or near-optimal and ask the study committee to check that the loss functions and prior involved are at least reasonable.

A major difference in the description of frequentist and Bayes tests is that the error probabilities of a frequentist test are quoted at a small number of parameter values while the error probability of the Bayes test is quoted, if at all, as an integral over the parameter space with respect to the prior distribution. The fact that these probabilities are sometimes numerically equal can be the source of some confusion since they are really very different items. Suppose, for example, normal observations $Y_1, Y_2, \cdots$ with mean θ and known variance σ^2 are available and it is desired to test between $\theta \leq 0$ and $\theta > 0$; suppose also that a flat generalized prior for θ is assumed in the Bayes analysis. After collecting n observations with mean $\bar{Y}$, the posterior distribution of θ is normal with mean $\bar{Y}$ and variance σ^2/n. One possible Bayes test is to reject $\theta \leq 0$ if

$$(1) \qquad P(\theta \leq 0 \mid \bar{Y}) \leq 0.05,$$

i.e., if $\bar{Y} \geq 1.645\sqrt{(\sigma^2/n)}$. As it happens, this test also satisfies the frequentist property

$$(2) \qquad P(\text{reject } \theta \leq 0 \mid \theta) \leq 0.05 \quad \text{for all } \theta \leq 0,$$

with equality when $\theta = 0$. Note, however, that the probability in (1) is with respect to random θ for fixed $\bar{Y}$, while the probability in (2) is with respect to random $\bar{Y}$ for fixed θ. The apparent similarity between the frequentist and Bayesian properties disappears in a sequential setting. If observations are collected in

up to K groups of size n, a sequential version of the Bayes procedure stops and rejects $\theta \leq 0$ after k groups of observations if

$$(3) \qquad P(\theta \leq 0 \mid \bar{Y}_k) \leq 0.05, \quad k = 1, \cdots, K,$$

i.e., if $\bar{Y}_k \geq 1.645\sqrt{(\sigma^2/nk)}$, where $\bar{Y}_k$ is the mean of the first kn observations; if $\bar{Y}_K \leq 1.645\sqrt{(\sigma^2/nK)}$, $\theta \leq 0$ is not rejected. Rules similar to this have been proposed by Mehta and Cain (1984), Berry (1985) and Freedman and Spiegelhalter (1989). On termination at analysis k, the posterior distribution for θ is normal with mean $\bar{Y}_k$ and variance σ^2/kn, even though a data-dependent stopping rule is used and, thus, (3) implies that in the Bayesian analysis 95% confidence of not having wrongly rejected $\theta \leq 0$ is preserved. However, if the repeated-sampling properties of this test are evaluated at specific values of θ, we find that (2) is not preserved, for example $P(\text{reject } \theta \leq 0 \mid \theta = 0) = 0.17$ for $K = 10$; qualitatively similar results are found if the flat prior is replaced by a proper normal prior with mean 0 and moderately high variance while a positive prior mean leads to even higher rejection probabilities. In contrast, frequentist group sequential tests, for example those of Pocock (1977) and O'Brien and Fleming (1979), maintain a specified type I error, (2), by raising the boundary for rejection of $\theta \leq 0$ at each analysis.

It is tempting to ask which is the "correct" adaptation of the fixed sample test to the sequential problem. However, I do not think there is enough information in the question to allow a clear answer. There is an element of arbitrariness about both the frequentist and Bayes fixed sample tests and the sequential versions are produced simply by preserving a particular familiar property in either the frequentist or Bayes setting. One route to a definitive answer would be to note the Bayes decision problem implicit in the fixed sample test, find the optimal sequential procedure for this problem and compare this with the frequentist and Bayes sequential tests. However, it turns out that the same fixed sample test results from a multitude of different decision problems, combining different priors and loss functions; moreover, there is further freedom in choosing the cost of sampling function in the sequential problem. Depending on the choice of prior, loss function and cost of sampling function, one of a whole range of different optimal sequential tests can result and, so, the original question remains unresolved. My conclusion is that, without further information, it is not possible to make an absolute judgment between the two types of test in question and it would be foolish to try to do so. Comparison must, therefore, be based on properties of the tests such as their operating characteristics and expected sample size functions, measured either at individual values of θ or integrated with respect to a prior.

I still have strong reservations about specifying the properties of a sequential test only after integrating out the prior distribution for θ. If members of an external audience with different prior beliefs are to be convinced by a Bayes test's outcome, some conservatism must be built into the test; with this in mind, the results of Rosenbaum and Rubin (1984) that properties of Bayes sequential tests are very sensitive to the chosen prior are a serious cause for concern. In many medical applications trial organizers have vested interests, either as physicians hoping to demonstrate the efficacy of a treatment they have pioneered or drug manufacturers seeking approval for a new product. The traditional role of phase III trials is to test the efficacy of promising new treatments in as controlled and impartial a way as possible, but it is difficult to imagine how anything like an impartial prior could be agreed upon for many such studies. Finally, if Bayes tests *are* to be employed, I would stress the importance of examining their repeated-sampling properties at individual parameter values; with reference to my previous example, even if I believed the relevant prior for θ, I would find a test with a "false positive" error probability of 0.17, in the usual frequentist sense, very worrying. Perhaps regulatory bodies and skeptical observers in general concentrate too much on worst case situations, but in drug testing, and especially in drug safety, such caution seems only prudent.

Having identified myself as an advocate of frequentist sequential tests specified in terms of repeated-sampling properties, I must now defend some of their peculiarities. It is undeniable that adherence to strict limits on repeated-sampling error probabilities can lead to some awkward problems. Professor Breslow notes the problem that arises in Lan and DeMets' (1983) method and which also occurs in Jennison and Turnbull's (1989) method of computing repeated confidence intervals, that no further data can be used once all the type I error has been "spent." The difficulty can be largely alleviated by careful experimental design and by taking full advantage of the flexibility afforded by the Lan and DeMets approach. However, the underlying problem *must* still remain for any test with limited type I error. Although, at first sight, Bayes methods appear to provide a convenient solution to this sort of problem, that they do so is also further evidence that they do not guarantee repeated-sampling type I error rates and, if the repeated-sampling properties of the Bayes test described above are representative, this may not be so desirable after all!

I shall now turn to the thorny issue of making inferences on termination of a sequential test. I would first note that it is essential to take account of the stopping rule when computing frequentist confidence intervals or p-values, if they are to satisfy their usual

repeated-sampling definitions. It may be tempting to try to avoid complications by proceeding as if the final data had arisen from a fixed sample size study, but this can only lead to greater problems; examples can be constructed where this method leads to guaranteed arbitrarily small p-values under the null hypothesis and where confidence intervals fail to contain the true parameter value with probability one.

In defending frequentist analyses which take account of the stopping rule, I start from the assumption that the sequential test is closely linked with a decision problem. As an example, consider a study comparing a new drug against the current standard treatment and let θ denote a measure of the improvement offered by the new drug. Suppose a sequential test is conducted to test H_0: $\theta = 0$ against the one-sided alternative $\theta > 0$ with type I error probability α and rejection of H_0 will be used to justify the efficacy of the new treatment to a regulatory body. In such a situation, the role of a p-value against H_0 or a confidence interval for θ upon termination is to provide additional information over and above the result of the formal test. A basic requirement, then, is agreement with the outcome of the test: the p-value should be less than α and a $1 - 2\alpha$ equal-tailed confidence interval for θ completely above 0 if and only if the test rejects H_0. Such agreement will be obtained if one uses standard frequentist methods and any sensible ordering of the sample space when computing the p-value or the confidence interval (see, for example, Fairbanks and Madsen, 1982; Jennison and Turnbull, 1983; Tsiatis, Rosner and Mehta, 1984). Given the close interconnections with the underlying decision problem, it might be argued that p-values and confidence intervals on termination should be termed "decisions" rather than "inferences," although I am not convinced that this is a very useful distinction, anyway.

The major argument against frequentist inferences upon termination seems to revolve around the possibility that two statisticians faced with the same set of data may report different p-values. There is no doubt that the apparent seriousness of this problem is magnified by the undue emphasis frequentist statisticians have placed on the p-value as *the* single summary of a set of data. The p-value is defined with respect to a sample space, and there is no reason that different sampling rules should not lead to different p-values for the same final set of data; this does not necessarily imply that different decisions should ensue although this possibility cannot be excluded. Suppose in the above example of a comparison between a new drug and a standard, a repeated significance test is used and this fails to reject H_0 but a "fixed sample" p-value on termination is less than α. The manufacturers of the new drug will doubtless feel disappointed that they had not chosen a fixed sample study with the same sample size as their completed sequential design; however, they would then have risked the complementary outcome of a final p-value greater than α but interim results that crossed the repeated significance testing boundary. If a regulatory body lays down limits on the acceptable type I error, these possibilities have to be balanced one against the other, just as increased power has to be weighed against the cost of greater expected sample size. It is an inescapable consequence of specifying frequentist error rates that such things can happen; however, once a frequentist accepts that a p-value is only one convenient summary of a set of data, the apparent paradox is not too hard to live with. As I have argued earlier, it is eminently reasonable that statisticians should choose to steer clear of introducing priors into what are intended to be impartial and definitive studies; the price to be paid, namely, the problem of dealing with a rather peculiar sample space and the need to examine critically the interpretation of familiar methods of inference, is far from prohibitive.

CONCLUSION

Bayesian methods have much to offer in the many biostatistical applications where substantial prior information is available. I am particularly interested in the prospect of Bayesian strategies for drug development, where sequences of experiments are carried out on many, often related, compounds. However, I retain my reservations about the practicality of obtaining an impartial prior for a final phase III trial, the results of which must convince the wider external audience.

Bayesians do not have a monopoly on good ideas, and many of the benefits of Bayesian methods have also been obtained by frequentists working with extended, sometimes hierarchical, models. The intent of this remark is not to quibble over who should take the credit for methodological advances but rather to note the similarities in current directions of frequentist and Bayesian research. Lively and constructive debate about real applied problems is an excellent stimulus to the flow of ideas between the various statistical schools, and I repeat my congratulations to Professor Breslow for his contribution to this discourse.