STATISTICS IN MEDICINE, VOL. 16, 1413–1430 (1997)

TUTORIAL IN BIOSTATISTICS
BAYESIAN DATA MONITORING IN CLINICAL TRIALS

PETER M. FAYERS,[1]*† DEBORAH ASHBY[2] AND MAHESH K. B. PARMAR[1]

[1] *MRC Cancer Trials Office, 5 Shaftesbury Road, Cambridge CB2 2BW, U.K.*
[2] *Department of Mathematical Sciences, University of Liverpool, Liverpool L69 3BX, U.K.*

SUMMARY

Many clinical trials organizations use regular interim analyses to monitor the accruing results in large clinical trials. In disease areas such as cancer, where survival is usually a major outcome variable, ethical considerations may lead to a stipulated requirement for data monitoring of mortality. This monitoring has frequently taken the form of limiting interim analyses to be few in number, and specifying an extreme p-value of, for example, $p < 0{\cdot}001$ or $p < 0{\cdot}01$ as grounds for early termination of the trial. Group-sequential methods are also used. However, none of these approaches formally assesses the impact that the results of a clinical trial may have upon clinical practice. Thus a trial might be terminated early because of apparent treatment benefits, but might fail to influence sceptical clinicians to modify their future treatment policy. We discuss the application of Bayesian methods, including the use of uninformative, sceptical and enthusiastic priors, and demonstrate that the necessary calculations are both straightforward to perform and easy to interpret statistically and clinically. Methods are illustrated with interim analyses of a clinical trial in oesophageal cancer. © 1997 by John Wiley & Sons, Ltd.

Statist. Med., **16**, 1413–1430 (1997)
No. of Figures: 0 No. of Tables: 1 No. of References: 20

1. INTRODUCTION

Interim analysis of accruing information in clinical trials is necessary in order to monitor for unexpectedly large treatment effects and for excess toxicity. In many clinical trials survival may be one of the main outcome measures, and it would clearly be unethical and unacceptable to continue recruiting patients to the trial if early results provide conclusive evidence of a convincing superiority of one or other treatment policies. These considerations have led many clinical trials organizations to institute formal procedures for regular monitoring and interim analyses of their trials, especially those trials which are large, have lengthy recruitment periods, and involve patient survival. In many cases the results of such monitoring are reviewed by specially convened Data Monitoring Committees.

* Correspondence to: P. M. Fayers, Unit for Epidemiology and Clinical Research, Faculty of Medicine, Medisinsk Teknisk Senter, N 7005 Trondheim, Norway

† Current address: Unit for Epidemiology and Clinical Research, Faculty of Medicine, Medisinsk Teknisk Senter, N 7005 Trondheim, Norway

CCC 0277–6715/97/121413–18$17.50

Bayesian methods which may be used for data monitoring in clinical trials are illustrated, and a simple exposition provided showing how to apply these techniques. Previously published papers on this subject have tended either to be theoretical or to be wide ranging;[1–6] in all cases, details of the methods have been presented in a manner which may not be readily accessible to those who simply seek to apply Bayesian methods to their own trials but find the mathematics unfamiliar. In this paper the description of the methods should be sufficiently simple for non-mathematical readers to appreciate the objectives, to be able to perform the calculations, and to understand the interpretation of the results; however, mathematical details intended for applied statisticians are also included. Thus this tutorial adopts a position mid-way between exposition and 'cook-book', and should be suitable for both clinicians and statisticians involved in clinical trials.

Three worked examples are given, based upon data monitoring of the Medical Research Council (MRC) OE02 trial which is evaluating the role of surgery with or without adjuvant chemotherapy for treatment of patients with oesophageal carcinoma. Confidentiality precludes the publication of interim results, and so hypothetical scenarios are presented illustrating monitoring of: (i) a trial with positive results, but at an early stage of patient recruitment; (ii) the same trial at a later stage, when early termination would be recommended, and (iii) a trial which could be terminated because early results suggest there is unlikely to be any treatment difference. These examples relate to survival comparisons, which are very pertinent to trials in many disease areas, but adaptation to binary or continuous endpoints is relatively simple.

2. MONITORING OF CLINICAL TRIALS

When a trial accrues patients over several years, results on earlier patients become available before the later ones are randomized. If the results look promising in favour of one treatment, the question can arise as to whether the trial should be terminated early. In particular, if the early results provide reasonably conclusive evidence of an advantage in favour of one of the treatments, it may be considered unethical to continue recruiting patients. This especially applies to clinical trials in potentially fatal diseases, in which patients receiving an inferior therapy may be at higher risk of death. It is crucial that such studies should be closely monitored so that if the new treatment has an effect that is larger than expected the trial may be terminated early; it is equally crucial that if the new treatment is unexpectedly found to be inferior, the trial should also terminate early. However, even where survival is not the primary outcome, it may still be unethical to continue exposing patients to an inferior treatment. Furthermore, one can also argue that it is an abuse of research funds to continue even a harmless clinical trial beyond the point at which there is sufficient evidence as to which therapy more is effective. Thus many clinical trial protocols contain explicit statements about the frequency and timing of interim analyses, and many trials have a formal, independent Data Monitoring Committee which is assigned the task of overseeing the monitoring of the trial.[7]

Superficially, therefore, it might appear that clinical trials should be terminated as soon as there is a convincing and statistically significant difference between the treatments. Thus one might envisage a sequential or group sequential trial,[7,8] which would specify a stopping rule based upon formal statistical tests. This approach is being used in a few MRC trials, albeit with some reservations.[9] It is worth remembering that if a trial was important enough to start in the light of the knowledge available then caution should be exercised before too lightly concluding that there is sufficient evidence to terminate the trial; hence p-values alone are unlikely to suffice for decision making about the future of a clinical trial, although they may be an important consideration.

Nevertheless, there is an opposing school of thought which argues persuasively that the role of a clinical trial is to influence clinical opinion and clinical practice. Thus if a clinical trial detects a large treatment effect after half the patients have been entered, and as a consequence is terminated early, that trial may be received with considerable scepticism by clinicians; despite any significant *p*-values that are cited, many clinicians may still remain unconvinced by the weight of evidence that has been produced. These clinicians are likely to continue treating new patients in the same way that they have done in the past. The clinical trial, therefore, will have failed in its primary objective. Through early termination, it has failed to obtain sufficient evidence to alter the management and therapy of future patients. This philosophy has led many trialists to be cautious about stopping recruitment prematurely. The ISIS (International Study of Infarct Survival) trials, for example, explicitly state in the protocols that the Data Monitoring Committee will only disclose interim results to the steering committee if there is '**both** (a) "proof beyond all reasonable doubt" that for all, or for some, types of patient one particular treatment is clearly indicated or contraindicated in terms of net difference in mortality, **and** (b) evidence that might reasonably be expected to influence materially the patient management of many clinicians who are already aware of the other main trial results'; the protocol also suggests that this might perhaps correspond to a difference of at least three standard deviations.[10] The need to convince others is also formalized by the drug regulatory process, and a trial that has stopped early may fail to convince regulators; for this reason, too, many trialists are wary of premature termination of patient recruitment.

The concept of the role of a clinical trial being to influence clinical opinion has a number of important consequences. In particular, it implies that statistical significance and statistical stopping rules will not in themselves be sufficient, and that one should additionally consider the prior opinions of clinicians. If clinicians, in general, are sceptical about the merits of a new treatment in terms of its prolonging life or curing patients, the necessary evidence to change that view will have to be substantial; if, on the other hand, most clinicians already expect the new treatment to be an improvement, far less weight of evidence will be necessary to influence them.

In practice, most major trials groups use an independent Data Monitoring Committee to help with the review of trials. One of the functions of a Data Monitoring Committee is to offer advice on whether a particular trial should terminate, and although this is not only a statistical decision, statistical guidelines can help formalize and clarify some of the issues outlined above.[11] There are essentially two schools of thought concerning the statistical procedures and calculations that should be made. The first, adopting a 'classical frequentist' approach, pre-specifies a number of 'looks' at the accumulating data and uses the observed *p*-values of these looks as a basis for stopping. At each look a relatively stringent significance level is used, so that the overall level of significance for the trial is maintained at, say, 5 per cent. The Pocock rule,[8] for example, uses the same significance level at every analysis, whereas the O'Brien/Fleming rule[12] uses extremely stringent criteria at the very earliest visits on the grounds that early observed differences are much more likely to be spurious. These are examples of group-sequential designs. The second, or 'Bayesian', approach formalizes the idea that external or prior evidence or beliefs can be summarized mathematically, and that in stopping the trial one is balancing the evidence from the trial against this other evidence.[1–6,13] When the trial evidence can outweigh this other evidence it is time to stop the trial. Clearly the formalization of this other evidence is critical. This paper examines the practical aspects of specifying prior opinions and the application of a Bayesian approach.

3. DESIGN OF MRC OESOPHAGEAL TRIAL

As an example, we consider the MRC OE02 clinical trial. This aims to evaluate the role of pre-operative chemotherapy for patients with resectable cancer of the oesophagus.

The outlook for patients with oesophageal cancer undergoing surgery remains poor, with only 20 per cent remaining alive at 2 years, and only 5 per cent alive and disease-free at 5 years. However, results from several small, uncontrolled, phase II studies suggest that this cancer may respond to chemotherapy given either pre- or post-operatively. 2-year survival figures of as large as 30 to 40 per cent have been claimed. Two of the more active chemotherapy agents are cisplatin and fluorouracil. Hence OE02 is comparing survival for patients randomized to either pre-operative chemotherapy followed by surgery, or surgery alone. The chemotherapy in OE02 consists of two four-day courses of cisplatin and fluorouracil, with an interval of three weeks between courses. (Copies of the protocol may be obtained from the MRC Cancer Trials Office.)

However, the chemotherapy is expensive. It may sometimes have adverse side-effects including nausea and vomiting, and less frequently diarrhoea, stomatitis, renal disturbance and myelosuppression. Furthermore, since these patients will eventually undergo surgery, most of them would prefer the surgery to take place as soon as possible. Demonstrating equivalence of the two treatment arms is of no interest. Therefore OE02 is testing the hypothesis that pre-operative chemotherapy will improve survival, and that the patients' overall well-being is not impaired. Thus the primary endpoint of interest in OE02 is length of survival, although clearly the general well-being of the patients is also evaluated.

In accordance with MRC standard policy, a Data Monitoring Committee was created. This includes one independent statistician, and two independent clinicians who are experts in oesophageal cancer but are not entering patients into OE02. The trial was launched in 1992, and has a planned sample size of 800 patients. Over 400 patients were recruited by summer 1996. As this is an on-going clinical trial, the true interim results are confidential; the examples that follow are based upon fictitious data.

4. NOTATION

- $\sim \mathrm{N}(\mu, \sigma^2)$ indicates 'is distributed as a normal (Gaussian) distribution with mean μ and standard deviation σ' (that is, a variance of σ^2)
- Φ is the cumulative normal probability, so that $\Phi(1{\cdot}6445) = 0{\cdot}95$.
- $\log(\ldots)$ represents logarithm to the base e.
- $\log(h)$ is the log-hazard ratio, and $\log(h_1)$ the log hazard ratio under the alternative hypothesis, H_1.
- We assume a clinical trial is being carried out, and that at the time of carrying out the interim analysis we have observed O_1 and O_2 deaths or 'events' in the two treatment groups.
- N_d, the total number of deaths observed to date, is given by

$$N_d = O_1 + O_2. \tag{1}$$

- E_1 and E_2 are the 'expected' number of deaths that would have been observed under the null hypothesis; computer programs which calculate survival comparisons usually display E_1 and E_2.

5. HAZARD RATIOS AND SURVIVAL TRIALS

Suppose the clinical trial was designed to compare survival in patients randomized between a standard form of treatment versus a new treatment. Frequently there will be prior knowledge

about the nature of the survival curve for the standard treatment. This may be derived from previous studies or from clinical experience. For example, in the OE02 trial past experience enabled us to expect that 20 per cent of patients receiving standard surgical treatment would still be alive at 2 years after surgery. Thus the 2-year survival rate is 0·20. More generally, we use $surv_1$ and $surv_2$ to represent the survival rates for the two treatment groups, where survival is measured at some fixed time relative to randomization. Thus $surv_1$ might be the pre-study estimate of survival in the standard or control arm of the trial, and would represent the proportion of patients expected to be alive at some specified time point relative to when the patient was randomized. $Surv_2$ would be the survival rate that is hypothesized for the alternative treatment. The trial will have been designed to test a null hypothesis of no treatment difference, against an alternative hypothesis that the treatment difference is at least $surv_2$-$surv_1$.

An estimate of $surv_1$, together with a target value for the alternative hypothesis of a treatment difference of at least $surv_2$-$surv_1$, is usually specified in clinical trial protocols and is used as a basis for sample size estimation (see example 1(a)). The sample size calculations ensure that when the clinical trial has been completed, and provided there has been adequate follow-up of the patients, the trial-based estimates of $surv_1$ and $surv_2$ will be sufficiently precise to enable adequately powerful hypothesis testing; a review of sample size issues is given in Fayers and Machin,[14] and tables for sample size estimation are available.[15,16]

In terms of hazard ratios, this is equivalent to carrying out a trial to detect a log hazard ratio, which we call $\log(h_1)$, of

$$\log(h_1) = \log(\log(surv_1)/\log(surv_2)). \tag{2}$$

Hence we have a null hypothesis that the log hazard ratio is zero, and an alternative hypothesis that the log hazard ratio is $\log(h_1)$.

5.1. Example 1(a)

In OE02 the baseline proportion surviving 2 years, in patients receiving surgery alone, was assumed to be 0·20 (20 per cent of patients remaining alive after 2 years). The alternative hypothesis is that pre-operative chemotherapy produces an absolute improvement of 10 per cent, to 0·30 at 2 years. These values were used as a basis for the sample size estimation, and are specified in the study protocol. From equation (2), this translates to an alternative hypothesis with a log hazard ratio of $\log(h_1) = 0{\cdot}290$.

6. INTERIM ANALYSES

When a trial is being monitored there will be interim, and less precise, estimates of $surv_1$ and $surv_2$ which are based upon the survival experience of patients currently recruited to the trial and followed up until the time of the analysis. These estimates allow calculation of the data-based log hazard ratio $\log(h_d)$, as will be shown later. The role of the interim analysis, and indeed the application of methods described in this paper, is to assess the currently available information from the trial so as to determine whether there is already sufficiently convincing accruing evidence from $\log(h_d)$ for the Data Monitoring Committee to consider terminating patient recruitment to the trial.

7. THE BAYESIAN APPROACH: PRIORS, LIKELIHOODS AND POSTERIORS

The Bayesian approach formalizes the procedure of having pre-study beliefs, which are then influenced by the results from an experiment such as a clinical trial, to yield revised beliefs. These pre-study beliefs are expressed as a 'prior distribution'. If we consider a therapy trial such as OE02, a clinician may start by believing, for example, that the absolute treatment difference is likely to be 10 per cent. Perhaps it is thought possible, although less likely, that the difference could be as large as 15 per cent or as small as 5 per cent; furthermore, it might be thought just about possible, but very unlikely, that it could range from 20 per cent down to 0 per cent. Some clinicians might even fear that the chemotherapy, by delaying surgery, could conceivably confer a survival disadvantage. By persuading the clinician to quantify the terms 'is likely', 'less likely', and 'very unlikely' in terms of probabilities (for example, some clinicians might ascribe the words 'less than 5 per cent of the time' to 'very unlikely'), we can build up a probability distribution for the chance of the treatment effect being of various magnitudes. This is called the 'prior distribution'.

After carrying out the study, we can evaluate the chance that we would have obtained such extreme data if the effect size were of a specified magnitude. This probability is called the 'data likelihood', and can be evaluated across the range of all plausible values for the effect size.

Finally, we have the 'posterior distribution', which is simply the prior distribution modified as a consequence of the observed results. This posterior distribution provides an estimate of what we would expect that same clinician to believe if realistic allowance is made for the information obtained from the trial.

Bayesian methods are becoming increasingly widely employed in clinical trials, although many of these applications remain controversial; a special issue of *Statistics in Medicine*[17] reviewed the state of art. For many statisticians the principal reason for disquiet with Bayesian techniques is the determination of suitable prior distributions. However, in the context of *monitoring* of trials, this becomes a largely specious issue, as we shall show. Thus even a traditional statistician, following the classical frequentist approach, is likely to find Bayesian monitoring not merely satisfactory but more appropriate than alternative approaches.[1,2,13]

8. BAYES' THEOREM

Bayes' theorem puts into mathematical notation the concept of having a prior belief which is then modified according to the observed data, resulting in a revised posterior belief. Formally, it states that the posterior distribution after observing data is proportional to the data 'likelihood' multiplied by the prior distribution. Thus if H is a hypothesis concerning a parameter, such as the treatment effect size (for example, the log hazard ratio of a survival comparison), we have:

$p(H)$ equals the pre-study opinion (prior probability) about the treatment effect size, and

$p(\text{data}|H)$ equals the likelihood of obtaining the observed data, given the effect size.

Then $p(H|\text{data})$ is the revised opinion (posterior probability) about the treatment effect size, given the observed results. This is proportional to the product of the prior and the data-likelihood, which can be written as:

$$p(H|\text{data}) \propto p(\text{data}|H)\, p(H). \tag{3}$$

This is the fundamental equation that underpins the Bayesian approach, and which we shall later apply to the prior distributions which represent clinicians' opinions.

9. PRIOR DISTRIBUTIONS

A prior distribution is chosen to provide an estimate of initial beliefs concerning the size of the potential therapeutic benefit. Thus, if one hopes that the results from a clinical trial will influence the treatment of future patients, the prior distribution should represent the level of scepticism that is expressed by those clinicians that one seeks to influence.

There are three principal types of prior we could use, which are the uninformative, or reference, prior, sceptical prior and enthusiastic prior.[1,2] Other priors that could be considered are a clinical prior (often similar to the enthusiastic prior) and a meta-analysis prior;[1–3] this paper only discusses the first three.

9.1. Uninformative prior

The uninformative, or reference, prior represents a lack of clinical opinion as to the likely treatment difference, and in that sense contains no information about prior beliefs or other prior knowledge; hence the name 'uninformative prior'. It corresponds more or less to the conventional 'frequentist' approach of significance testing. One possible and adequate approximation is to assume that this prior can be represented by a normal distribution with mean 0 (corresponding to the null hypothesis of no difference) and an infinite variance. Although this is an 'improper' distribution (since a normal distribution cannot really have an infinite variance), it serves as a convenient mathematical device which in practice yields sensible posterior distributions. For analogy with the other prior distributions being considered, we will take this infinite variance to be 4/0:

$$\text{Uninformative prior} \sim \mathrm{N}(0, 4/0).$$

9.2. Sceptical prior

The sceptical prior is specified by considering there is only a small probability that the alternative hypothesis (or better than it) is likely to be true. Thus a sceptical prior distribution may be given by considering the best guess to be zero, but allowing that there is a small probability, say γ, of an effect as large as or larger than $\log(h_1)$.

Interestingly, it can be shown that if we take $\gamma = 5$ per cent, and we carry out five interim analyses of the data, then the overall type I error is about 5 per cent (assuming the trial was designed with 90 per cent power, 5 per cent significance level, and a particular stopping rule).[18] Furthermore, if we compare this approach with classical methods, adopting a sceptical prior gives us a procedure which lies between the Pocock rule and the O'Brien/Fleming rule.[2] It should be noted that the role of the sceptical prior is to guard against over-enthusiastic acceptance of a positive and possibly large treatment effect that might be observed by chance at the time of an interim analysis. In this situation emphasis on probabilities in only one direction is appropriate, and this prior is not appropriate if the treatment effect appears to be in the opposite direction. Thus a one-sided probability is used.

Setting $\gamma = 5$ per cent, we can solve for σ_{scep}:

$$1 - \Phi\left(\frac{\log(h_1)}{\sigma_{\text{scep}}}\right) = \gamma = 0{\cdot}05$$

$$\frac{\log(h_1)}{\sigma_{\text{scep}}} = 1{\cdot}6445$$

$$\sigma_{\text{scep}} = \log(h_1)/1{\cdot}6445. \tag{4}$$

Although a normal distribution may not be entirely correct for a sceptical prior, it provides a convenient assumption which has been supported by empirical evidence; studies have been carried out, showing that clinical opinions commonly result in a log hazard ratio with a normal prior distribution.[2]

Therefore we adopt a sceptical prior which is represented by a normal distribution with zero mean and variance σ^2_{scep}. Hence we have the sceptical prior $\sim \text{N}(0, \sigma^2_{\text{scep}})$.

Now for survival data the variance of the log hazard ratio is approximately $4/n$ where n is the total number of events; this is a remarkably good approximation.[19] This approximation enables us to determine an equivalent study size corresponding to this sceptical prior. Equating the sceptical prior variance to $4/n$, we have $\sigma^2_{\text{scep}} = 4/n$. This can be solved for n, and thus the sceptical prior is equivalent to having performed a trial with $N_p = n$ patients, all of whom have died, and in which no difference has been observed between the two arms. Hence:

$$N_p = 4/\sigma^2_{\text{scep}} \tag{5}$$

giving the sceptical prior $\sim \text{N}(0, 4/N_p)$.

9.3. Example 1(b)

For the MRC OE02 clinical trial, we obtained $\log(h_1) = 0{\cdot}290$. Solving equation (4) gives $\sigma_{\text{scep}} = 0{\cdot}176$, and from equation (5) we see that this is equivalent to having performed a trial with $N_p = 129$ patients, all followed to death.

9.4. Enthusiastic prior

An enthusiast might argue that the treatment difference is bound to be greater than zero and that, as a best guess, it is likely to be equal to $surv_2$-$surv_1$. Hence an enthusiastic prior can be taken by considering the best guess to be the alternative hypothesis. We also assume this has the same precision as the sceptical prior; therefore we assume a normal distribution with mean $\log(h_1)$ and variance $4/N_p$:

$$\text{Enthusiastic prior} \sim \text{N}(\log(h_1), 4/N_p).$$

10. JUSTIFICATION FOR OUR CHOICE OF PRIOR DISTRIBUTIONS

We have described three prior distributions, although there are many potential distributions that could be considered; Parmar *et al.*[2] and Spiegelhalter *et al.*[1] discuss a few other possibilities. The difficulty of making an objective and non-controversial selection is, of course, one of the reasons why frequentist statisticians often object to Bayesian techniques. Thus, for example, if the outcome of a clinical trial is being reported using a Bayesian approach, there is always the anxiety that the investigators may have chosen an unduly optimistic prior distribution, and thus may have unfairly claimed that they have confirmation of a treatment effect.

In the context of monitoring of clinical trials, many of these issues which are sometimes levelled against Bayesian methods are largely irrelevant. In particular, for monitoring, one attempts to

guard against premature termination of a clinical trial and thus chooses a prior distribution which reduces the chance of wrongly claiming that the results have already become conclusive. In effect, the role of Bayesian monitoring is to determine whether the early results are already so overwhelmingly convincing that there is no need to continue to collect further confirmatory information.

Hence, in general, the sceptical prior distribution is appropriate for monitoring positive results, when interim analyses appear to be indicating a substantial treatment effect that might lead to stopping the trial. On the other hand, the enthusiastic prior distribution is of greatest importance when the results are leaning towards equivalence, and as a brake against early termination of a trial when initial results suggest a detrimental apparent effect of the new treatment. Choosing these two priors provides a useful brake against premature termination of trials.

It will often be useful to present the results of analyses under several alternative prior distributions, so that the impact of the observed data may be reviewed according to different levels of scepticism. The uninformative prior has the advantage of corresponding roughly to the classical frequentist approach, and so this together with the sceptical and enthusiastic priors gives a broad and useful overview of the implications of terminating a clinical trial.

11. DATA

Having chosen one or more prior distributions, we then consider the observed data.

We can estimate h_d from the observed and expected deaths, using

$$\log(h_d) = \log\left(\frac{(O_1/E_1)}{(O_2/E_2)}\right). \tag{6}$$

We again assume that the log hazard ratio follows a normal distribution, so that the observed data have a normal distribution with mean h_d and variance $4/N_d$. Hence

$$\text{data} \sim \text{N}(\log(h_d), 4/N_d).$$

12. POSTERIOR DISTRIBUTIONS

The prior distribution may be based upon initial or prior beliefs and prejudices, and external evidence such as reports of other (possibly small, possibly non-randomized) studies. When data, in the form of deaths, accrue from the clinical trial, the prior beliefs should be modified according to the weight of evidence that has been collected. Statistically, we use the distribution of the observed data to modify our prior distributions, resulting in a 'posterior' distribution which reflects our revised beliefs in the light of the new data.

12.1. Statistical derivation of posterior distributions

The standard Bayes equation (3) allows a prior distribution to be modified to reflect the data that have been observed, yielding the posterior distribution. Specifically, we can adapt (3) and apply it to the distributions discussed above.

The three prior distributions may be written in the general form of $\text{N}(\mu_x, 4/N_x)$ where μ_x and N_x are 0,0 (uninformative prior), $0, N_p$ (sceptical prior), or $\log(h_1), N_p$ (enthusiastic prior).

Thus, corresponding to $p(H)$ in equation (3), we have

$$\text{Prior} \sim \text{N}(\mu_x, 4/N_x).$$

Similarly, corresponding to $p(\text{data}/H)$, writing $\mu_d = \log(h_d)$, we have

$$\text{Data likelihood} \sim N(\mu_d, 4/N_d).$$

Then by $p(H|\text{data}) \propto p(\text{data}|H)\, p(H)$ we obtain the posterior distribution from the product of the prior and the likelihood. It can be shown that solving the equations gives

$$\text{Posterior} \sim \text{N}\left(\frac{N_x\mu_x + N_d\mu_d}{N_x + N_d}, \frac{4}{N_x + N_d}\right).$$

By applying these equations we can derive the posterior distributions which reflect the impact of the currently observed data upon the uninformative, sceptical and enthusiastic priors. The full set of distributions are listed below.

13. SUMMARY OF EQUATIONS:

We consider the following prior distributions:

$$\text{Uninformative prior} \sim \text{N}(0, 4/0)$$

$$\text{Sceptical prior} \sim \text{N}(0, 4/N_p)$$

$$\text{Enthusiastic prior} \sim \text{N}(\log(h_1), 4/N_p).$$

For the observed data, we have

$$\text{Data 'likelihood'} \sim \text{N}(\log(h_d), 4/N_d).$$

This gives posterior distributions:

$$\text{Uninformative} \sim \text{N}(\log(h_d), 4/N_d) \tag{P1}$$

$$\text{Sceptical} \sim \text{N}\left(\frac{N_d \log(h_d)}{N_p + N_d}, 4/(N_p + N_d)\right) \tag{P2}$$

$$\text{Enthusiastic} \sim \text{N}\left(\frac{N_p \log(h_1) + N_d \log(h_d)}{N_p + N_d}, 4/(N_p + N_d)\right) \tag{P3}$$

14. EVALUATION OF POSTERIOR DISTRIBUTIONS/REVISED BELIEFS

The equations (P1), (P2) and (P3) represent the distributions of the revised beliefs, based upon the pre-study prior beliefs which have been modified in the light of the current data. Now that the equations for these posterior distributions have been derived, we can calculate the probabilities of certain levels of improvement for each of the corresponding prior distributions and the given data.

Let us suppose that the target improvement is δ. From (2), we have

$$\log(h_\delta) = \log(\log(surv_1)/\log(surv_1 + \delta)). \tag{7}$$

This can be substituted into equations (P1), (P2), (P3), to produce a table, as in the following example.

14.1. Example 1(c)

Equations (P1), (P2) and (P3) can be used to construct Table I, showing 0 per cent, 5 per cent and 10 per cent absolute improvement in percentage survival. The reason for selecting these values is that 0 per cent corresponds to the null hypothesis, 10 per cent is the alternative hypothesis for the example trial, and 5 per cent is both mid-way between 0 per cent and 10 per cent and arguably a realistic yet still worthwhile survival benefit; if the survival gain were as low as 1 per cent, it is unlikely that clinicians would use the toxic and expensive chemotherapy, but if it were 5 per cent it might be a sufficiently large improvement to warrant recommending adopting pre-operative chemotherapy as the treatment of choice. Different percentages will be appropriate for other trials. Note that A is the probability that the improvement is greater than 0 per cent when a uniform prior is assumed, and that $1 - A$ corresponds roughly to a conventional p-value (one-sided) where a single test is made (that is, without allowance for multiple looks at the data). Similarly, $1 - B$ can be shown to be equivalent to a significance level corresponding to a monitoring rule lying between the Pocock and the O'Brien/Fleming rules.[2]

Table I

Target improvement δ	Log hazard ratio $\log(h_\delta)$	Probability that improvement is greater than the target value		
		Uninformative prior (P1)	Sceptical prior (P2)	Enthusiastic prior (P3)
0%		A	B	
5%			C	
10%				

15. STOPPING CRITERIA

The table showing the probabilities associated with various target improvements can be reviewed by the Data Monitoring Committee. However, although the Bayesian framework is useful for interpreting the current trial's results in an informal manner, it is still useful to have guidelines for when seriously to consider stopping. If the trial is between two treatment arms, then a reasonable criterion is to demand that the posterior probability of one treatment being better, in the light of a sceptical prior belief, is at least 95 per cent; thus cell B should exceed 95 per cent. Alternatively, if a non-zero target improvement is sought, a reasonable criterion might be to accept a posterior probability of 90 per cent; in Table I, for a 5 per cent improvement this would imply that cell C should be at least 90 per cent.

16. CALCULATIONS: SUMMARY

We start with a table giving the observed and expected number of events in each group, as is readily obtained from most survival analysis programs.

$N_d = O_1 + O_2 =$ the total number of deaths observed to data.

Log(h_d), the log hazard ratio, can be obtained from equation (6)

$$\log(h_d) = \log\left(\frac{(O_1/E_1)}{(O_2/E_2)}\right)$$

Equations (2), (4) and (5) enable N_p to be estimated for the sceptical prior:

$$\log(h_1) = \log(\log(surv_1)/\log(surv_2))$$

where $surv_1$ and $surv_2$ are the hypothesised event rates.

$$\sigma_{\text{scep}} = \log(h_1)/1{\cdot}6445$$

$$N_p = 4/\sigma^2_{\text{scep}}.$$

Then the list of summary equations can be expanded out, although only the posterior distributions (P1), (P2) and (P3) are used for constructing the table of probabilities of various improvements, δ.

Finally, equation (7) allows $\log(h_\delta)$ to be evaluated for the improvements δ; this value of $\log(h_\delta)$ can be substituted for $\log(h_d)$ in (P1), (P2) and (P3), giving the required table of probabilities.

17. EXAMPLES

Three worked examples are given, illustrating monitoring of clinical trials such as the oesophageal trial where survival is the main endpoint. These illustrate the calculations for three different arios, and indicate the suggested interpretation of the results. In particular, example 1 (which is a continuation of the example already used in the text) is a trial which might be regarded as beginning to show emerging treatment differences, but is at an early stage of patient recruitment; this trial should be continued. Example 2 is the same trial at a later stage, when early termination would be recommended, and example 3 is a trial which could be terminated because early results suggest there is unlikely to be any treatment difference.

17.1 Worked example 1

The following is an example of what might happen if early results for the OE02 trial were to suggest that there might be a treatment effect, and the Data Monitoring Committee wished to consider the possibility of early termination. We shall adopt the position of a sceptic and mainly focus upon the sceptical prior.

We assume that routine interim analyses are being carried out, say annually. At the time of the first analysis perhaps 200 patients have been entered into the trial, of whom 100 have died (60 in group 1, and 40 in group 2), out of those so far recruited and entered into the trial. Standard survival-analyses techniques allow computation of the expected number of deaths in each group; these values are calculated and printed out by most survival analysis software.

Suppose the following results were obtained at the first interim analysis of the MRC OE02 trial:

Data:

Group	Observed	Expected	Obs/Exp
1	60	48·00	1·250
2	40	52·00	0·769

Applying the equations in Section 16, we obtain the log hazard ratio:

$$\text{Hazard ratio} = 1{\cdot}25/0{\cdot}769 = 1{\cdot}625$$

$$\text{Log hazard ratio} = 0{\cdot}486.$$

From parts 1(a) and 1(b) of this example, already given in the text, we have $surv_1 = 0{\cdot}20$ and $surv_2 = 0{\cdot}30$, giving

$$\log(h_1) = \log(\log(0{\cdot}20)/\log(0{\cdot}30)) = 0{\cdot}290.$$

$$\text{Hence } \sigma_{\text{scep}} = \log(h_1)/1{\cdot}6445 = 0{\cdot}176, \text{ and } N_p = 4/0{\cdot}176^2 = 129 \text{ patients.}$$

Priors: From the summary of equations, we have the following priors, data likelihood and posteriors:

$$\text{Uninformative prior} = \text{N}(0, 4/0)$$

$$\text{Sceptical prior} = \text{N}(0, 4/129)$$

$$\text{Enthusiastic prior} = \text{N}(0{\cdot}290, 4/129)$$

$$\text{Data likelihood} = \text{N}(0{\cdot}486, 4/100)$$

$$\text{Uninformative posterior} = \text{N}(0{\cdot}486, 4/100) \quad \text{from (P1)}$$

$$\text{Sceptical posterior} = \text{N}(0{\cdot}212, 4/229) \quad \text{from (P2)}$$

$$\text{Enthusiastic posterior} = \text{N}(0{\cdot}376, 4/229) \quad \text{from (P3).}$$

The table showing the probabilities of various improvements can be constructed by estimating the log hazard ratio for the target improvement, and using the three posterior distributions.

For example, for a target improvement of 0·100 upon the assumed baseline survival rate of 0·20 in OE02, we use equation (7) to obtain an equivalent log hazard of

$\log(\log(0{\cdot}20)/\log(0{\cdot}20 + 0{\cdot}100)) = 0{\cdot}290$. We then apply the enthusiastic posterior of N(0·376, 4/299) giving $1 - N\{(0{\cdot}290 - 0{\cdot}376)/\sqrt{(4/229)}\} = 0{\cdot}741$, as in the bottom right-hand cell.

Probabilities of improvement being greater than:

Target improvement	Log hazard	Uninformative	Sceptical	Enthusiastic
0·000	(0·000)	0·992	0·946	0·998
0·050	(0·149)	0·954	0·683	0·957
0·100	(0·290)	0·836	0·277	0·741

Interpretation: Our interpretation is that, starting from a sceptical position, there is reasonable (94·6 per cent) evidence of some improvement, but only modest evidence (68 per cent) of a difference as large as a 5 per cent improvement in survival, and only slim evidence (28 per cent) of a difference as big as 10 per cent.

Recommendations: Casting a sceptical eye over the data available, we would recommend that:

(i) the trial should continue at present;
(ii) the Data Monitoring Committee should consider recommending termination of the trial if the sceptical posterior probability of a 5 per cent difference exceeds 90 per cent.

17.2. Worked example 2

Suppose the study in example 1 continued recruiting patients, with the same observed hazard ratio, until 350 patients had entered and 200 deaths had occurred. The Data Monitoring Committee might now consider that there is increasing weight of evidence in favour of a treatment effect, and would again wish to re-examine the impact of the data upon someone adopting a sceptical stance.

Repeating all the calculations would yield the following:

Design: P1 = 0·20, P2 = 0·30, log hazard = 0·290, gamma = 0·05, SD scep = 0·176 (equivalent to study of 129 patients).

Data:

Group	Observed	Expected	Obs/Exp
1	120	96·00	1·250
2	80	104·00	0·769

Hazard ratio = 1·625

Log hazard ratio = 0·486.

Priors:

Uninformative prior = N(0, 4/0)

Sceptical prior = N(0, 4/129)

Enthusiastic prior = N(0·290, 4/129)

Data likelihood = N(0·486, 4/200)

Uninformative posterior = N(0·486, 4/200)

Sceptical posterior = N(0·295, 4/329)

Enthusiastic posterior = N(0·409, 4/329).

Probabilities of improvement being greater than:

Target improvement	Log hazard	Uninformative	Sceptical	Enthusiastic
0·000	(0·000)	1·000	0·996	1·000
0·050	(0·149)	0·991	0·907	0·991
0·100	(0·290)	0·916	0·518	0·859

Interpretation: We now see that a person who demanded an absolute change of 5 per cent in survival as the minimal worthwhile improvement, and who had a sceptical prior belief, would be likely to be convinced by the observed data, having a posterior probability of 90·7 per cent.

Recommendations: Casting a sceptical eye over the data available, we would recommend that the trial should cease recruitment of patients.

17.3. Worked example 3

Another situation might arise in which early results make it appear as if there is no difference between the treatments. For example, one might have entered about half the patients and might have observed death rates which suggest that the two treatments are either equal or that the new treatment is inferior. Would it be worth continuing the trial? In the case of the OE02 trial, one should remember that we are seeking to establish whether there is a treatment advantage (equivalence is not of interest). Thus we might decide that we would consider early termination if even an enthusiast would agree that it is very unlikely that the new treatment could provide clinically worthwhile treatment benefit.

Suppose that the following situation were to occur during the monitoring of the clinical trial. In this case we have recruited 450 patients and observed 300 deaths, and these are divided almost equally between the two treatment groups. Is it worth continuing the trial?

Here, again, we assume that a 5 per cent difference in survival would be considered worthwhile.

Data:

Group	Observed	Expected	Obs/Exp
1	153	150·00	1·020
2	147	150·00	0·980

$$\text{Hazard ratio} = 1{\cdot}041$$

$$\text{Log hazard ratio} = 0{\cdot}040.$$

Priors:

$$\text{Uninformative prior} = \text{N}(0, 4/0)$$

$$\text{Sceptical prior} = \text{N}(0, 4/129)$$

$$\text{Enthusiastic prior} = \text{N}(0{\cdot}290, 4/129)$$

$$\text{Data likelihood} = \text{N}(0{\cdot}040, 4/300)$$

$$\text{Uninformative posterior} = \text{N}(0{\cdot}040, 4/300)$$

$$\text{Sceptical posterior} = \text{N}(0{\cdot}028, 4/429)$$

$$\text{Enthusiastic posterior} = \text{N}(0{\cdot}115, 4/429).$$

Probabilities of improvement being greater than:

Target improvement	Log hazard	Uninformative	Sceptical	Enthusiastic
0·000	(0·000)	0·636	0·614	0·884
0·050	(0·149)	0·172	0·105	0·362
0·100	(0·290)	0·015	0·003	0·035

Interpretation: Adopting the stance of an enthusiast, we find that there is only a small chance (36·2 per cent) that there could be a difference of 5 per cent or greater. Someone with a more open mind might prefer to use the 'uninformative' prior, which suggests that the chance is even smaller, at 17 per cent. A sceptic, of course, would feel that a probability of 10 per cent tends to confirm his prior beliefs. The probability of a 10 per cent improvement is especially unlikely.

Recommendations: Even an enthusiast would have to agree that the chance of a treatment advantage of 10 per cent is exceedingly slim, and that there is but a small chance of a 5 per cent improvement. Probably it would be a waste of resources if more patients were entered into the trial. The trial should be discontinued.

However, there may be reasons beyond the calculations presented here for the trial to continue. For example, if another related trial has recently reported a large effect in favour of chemotherapy then it may be decided to continue the current trial so as to refute (or confirm) this other result. Although in principle this could be modelled by introducing the data (likelihood) for this second trial as well, and amending the posterior distribution, in practice it might be felt that the decisions should be reached by discussion and the application of informed value judgement by the Data Monitoring Committee. As always, other considerations, whether wider statistical ones or non-statistical, might lead to a decision other than that suggested by a formal statistical stopping rule.

18. CONCLUSION

Bayesian monitoring, as illustrated here, is very simple to implement. It helps to put into perspective one major inherent problem in the early termination of trials, namely the risk that the results will be regarded by sceptical clinicians as inconclusive. Thus it should help to ensure that trials only stop early when the results to date are sufficiently conclusive. Furthermore, our experience has been that in this context clinicians find Bayesian concepts intuitively appealing; the idea of collecting sufficient data to convince not only enthusiasts but both those with open minds, and sceptics too, accords with most clinicians' experience of research and the introduction of new drugs.

The methods described in this paper were initially tested retrospectively on the MRC randomized trials of misonidazole for head and neck cancer,[1] neutron therapy for pelvic cancer,[1] and chemotherapy for osteosarcoma.[20] Following the successful application of the methodology, the Bayesian approach for monitoring was adopted for prospective use on the two MRC trials of continuous hyperfractionated accelerated radiotherapy (CHART) for head and neck cancer and bronchus cancer,[1,2] and the on-going MRC oesophageal cancer trial (OE02) that provided basis of the hypothetical worked examples. Our experience to date leads us to recommend these procedures for more general application in monitoring of randomized controlled trials.

It is, perhaps, also worth finishing on a cautionary note. Bayesian analyses, like any other statistical calculations relating to stopping rules, provide information which should be considered within the more general context of the impact of early termination of a trial. There may be many reasons for deciding to ignore the calculations of posterior probabilities, and to continue recruiting patients to the trial. For example, there many be considerations concerning other endpoints such as toxicity, quality of life, and cost. Also, if a trial is stopped prematurely, some readers may regard the results as dubious and less convincing whatever approach is used; in effect, they may have an additional scepticism which they apply to all trials which are stopped early. Perhaps this trial may contribute to meta-analyses or overviews. If a trial is terminated permaturely, it may become difficult to later launch a confirmatory study. Thus one should never blindly and naively use Bayesian or any other monitoring in isolation and without full consideration of the implications.

Nevertheless, we believe that the Bayesian approach presented here makes explicit many of the issues involved in the monitoring of trials, and because of this it deserves to be more widely used.

ACKNOWLEDGEMENTS

This paper was motivated by the MRC Data Monitoring Committee for the oesophageal trial (OE02), and we would like to thank the clinical members Professor W. Duncan and Dr. J. Dark, and the clinical coordinator Dr. D. J. Girling.

REFERENCES

1. Spiegelhalter, D. J., Freedman, L. S. and Parmar, M. K. B. 'Bayesian approaches to randomised trials', *Journal of the Royal Statistical Society, Series A*, **157**, 357–416 (1994).
2. Parmar, M. K. B., Spiegelhalter, D. J. and Freedman, L. S. 'The CHART trials: Bayesian design and monitoring in practice', *Statistics in Medicine*, **13**, 1297–1312 (1994).
3. Abrams, K., Ashby, D. and Errington, D. 'Simple Bayesian analysis in clinical trials: a tutorial', *Controlled Clinical Trials*, **15**, 349–359 (1994).
4. Freedman, L. S. and Spiegelhalter, D. J. 'Application of Bayesian statistics to decision making during a clinical trial', *Statistics in Medicine*, **11**, 23–25 (1992).
5. Freedman, L. S. and Spiegelhalter, D. J. 'Comparison of Bayesian with group sequential methods for monitoring clinical trials', *Controlled Clinical Trials*, **10**, 357–367 (1989).
6. Freedman, L. S. and Spiegelhalter, D. J. 'The assessment of subjective opinion and its use in relation to stopping rules for clinical trials', *Statistician*, **32**, 153–160 (1983).
7. Pocock, S. J. *Clinical Trials: A Practical Approach*, Wiley, Chichester, 1983.
8. Pocock, S. J. 'Interim analyses for randomised clinical trials: the group sequential approach', *Biometrics*, **38**, 153–162 (1982).
9. Fayers, P. M., Cook, P. A., Machin, D., Donaldson, N., Whitehead, J., Ritchie, R., Oliver, R. T. D. and Yuen, P. 'On the development of the Medical Research Council trial of alpha-interferon in metastatic renal carcinoma', *Statistics in Medicine*, **13**, 2249–2260 (1994).
10. ICRF Clinical Trial Service Unit. *ISIS 3 Protocol*, CTSU, Radcliffe Infirmary, Oxford UK, 1989.
11. Parmar, M. K. B. and Machin, D. 'Monitoring clinical trials – experience of and proposals under consideration by the Cancer Therapy Committee of the British Medical Research Council', *Statistics in Medicine*, **12**, 495–504 (1993).
12. O'Brien, P. C. and Fleming, T. R. 'A multiple testing procedure for clinical trials', *Biometrics*, **35**, 549–556 (1979).
13. Berry, D. A. 'A case for Bayesianism in clinical trials', *Statistics in Medicine*, **12**, 1377–1393 (1993).
14. Fayers, P. M. and Machin, D. 'Sample size: how many patients are necessary?', *British Journal of Cancer*, **72**, 1–9 (1995).
15. Machin, D. and Campbell, M. J. *Statistical Tables for the Design of Clinical Trials*, Blackwell Scientific Publications, Oxford, 1987.

16. Machin, D., Campbell, M. J., Fayers, P. M. and Pinol, A. *Sample Size Tables for Clinical Studies*, Blackwell Science, Oxford, 1997.
17. Ashby, D. 'Preface: Papers from the conference on Methodological and Ethical Issues in Clinical Trials', *Statistics in Medicine*, **12**, 1373–1374 (1993).
18. Grossman, J., Parmar, M. K. B., Spiegelhalter, D. J. and Freedman, L. S. 'A unified method for monitoring and analysing controlled trials', *Statistics in Medicine*, **13**, 1815–1826 (1994).
19. Tsiatis, A. A. 'The assymptotic joint distribution of the efficient scores test for the proportional hazards model calculated over time', *Biometrika*, **68**, 311–315 (1981).
20. Spiegelhalter, D. J., Freedman, L. S. and Parmar, M. K. B. 'Applying Bayesian thinking in drug development and clinical trials', *Statistics in Medicine*, **12**, 1501–1511 (1993).

