6

Randomised Controlled Trials

A Bayesian: one who asks you what you think before a clinical trial in order to tell you what you think afterwards. (Senn, 1997b)

6.1 INTRODUCTION

Randomised controlled trials are traditionally considered the 'gold standard' for evaluation of health-care interventions, and have provided fertile territory for arguments between alternative statistical philosophies. In this chapter we consider a number of specific issues in which a distinct Bayesian approach is identifiable: these include the role of decision theory, ethics of randomisation, use of historical controls, selection of sample size, monitoring sequential studies, subset analysis, alternative designs and so on. Some of the strongest arguments for the Bayesian approach have been made in this context, with notable examples being Cornfield (1976), Berry (1993) and Kadane (1995). Each of these authors has emphasised the internal consistency of the Bayesian approach, and welcomed the need for explicit prior distributions and loss functions as producing scientific openness and honesty: see Section 6.13 for additional references by these and other authors.

The issues in this chapter are largely common to trials both in the public sector and in the pharmaceutical industry. For industry-sponsored trials we shall use the standard language of drug development: phase I studies deal with identifying a safe dose, usually on healthy volunteers; phase II studies are concerned with finding an effective dose; phase III studies are intended to prove treatment benefit over an appropriate control; and phase IV studies monitor the use and possible side-effects of a drug in routine use. This structure is necessarily rather simplistic, and there are increasing moves toward hybrid studies in order to speed up the drug development process. Parallel phases of development can be given for complex

Bayesian Approaches to Clinical Trials and Health-Care Evaluation D. J. Spiegelhalter, K. R. Abrams and J. P. Myles
 ISBN: 0-471-49975-7

public health interventions (Campbell *et al.*, 2000): in phase I an intervention is developed possibly through a theoretical model; in phase II explanatory trials in tightly controlled situations seek to demonstrate the potential efficacy of the intervention; in phase III pragmatic trials evaluate its costs and effectiveness in practice; and in phase IV the intervention is rolled out into routine use.

We shall begin by considering the basic issue of whether a trial is for inference or decision (Section 6.2), and then investigate the role of null hypotheses and their relation to the demands set of a new intervention (Section 6.3). The ethics of randomisation are then viewed from a Bayesian perspective (Section 6.4). A substantial section explores a number of ways in which prior opinion can be incorporated into sample-size calculations (Section 6.5), followed by a full discussion of the many ways to tackle the important issue of trial monitoring (Section 6.6), and the possible use of sceptical priors in deciding whether a confirmatory trial is necessary (Section 6.7). Apart from repeated looks at the data, 'multiplicity' features in many aspects of trial design and analysis, and we briefly discuss multiple subsets, outcomes, centres and trial arms (Section 6.8). The use of historical control groups fits naturally into a Bayesian perspective and is treated in some detail (Section 6.9); different trial designs are then examined, for example data-dependent allocation (Section 6.10) and multiple *N*-of-1 studies (Section 6.11). We only briefly consider phase I and II studies (Section 6.12), and discussion about the regulatory context is left until we consider policy decisions (Chapter 9).

6.2 USE OF A LOSS FUNCTION: IS A CLINICAL TRIAL FOR INFERENCE OR DECISION?

There has been a heated dispute about whether a clinical trial should be considered as a *decision* problem, with an accompanying loss function, or as an *inference* problem in which no explicit loss function is developed and conclusions are based solely on the posterior distributions of quantities of interest. This has been a point of clear distinction between different schools of Bayesianism (Section 3.20). Here we briefly review the arguments.

1. *A clinical trial should be a decision.* Lindley (1994) categorically states that 'Clinical trials are not there for inference but to make decisions', while Berry (1994) states that 'deciding whether to stop a trial requires considering why we are running it in the first place, and this means assessing utilities'. Healy (1978) considers that 'the main objective of almost all trials on human subjects is (or should be) a decision concerning the treatment of patients in the future'. The potential role for explicit statement of a loss function is a running theme throughout discussions on sample size (Section 6.5), sequential analysis (Section 6.6.4), adaptive allocation (Section 6.10) and payback from research programmes (Section 9.10), and many would argue that the

eventual decision is inseparable from the design and analysis of a study. From an economic perspective, it is claimed that a utility approach to clinical trial design and analysis is necessary in order to prevent conclusions based on inferential methods leading to health or monetary losses. This perspective derives from the observation made in Section 3.14 that only the expected utility of a decision is relevant, and expressions of uncertainty are, theoretically, of no concern except when deciding whether to collect further evidence. This echoes the original work on pragmatic clinical trials by Schwartz *et al.* (1980), in which it was argued that *P*-values and interval estimates are irrelevant to trials that guide decisions. The role for decision theory in health policy and regulation will be covered in Section 9.11.

The explicit use of utility functions within the design and monitoring of clinical trials is controversial but has been explored in a number of contexts: for example, Berry and Stangl (1996a) discuss the problems of whether to stop a phase II trial based on estimating the number of women in the trial and who will respond in the future; whether to continue a vaccine trial by estimating the number of children who will contract the disease; and the use of adaptive allocation in a phase III trial such that at each point the treatment which maximises the expected number of responders is chosen.

2. *A clinical trial provides an inference*. Armitage (1985), Breslow (1990), DeMets and Lan (1994), Simon (1977) and Orourke (1996) all describe how it is unrealistic to place clinical trials within a decision-theoretic context, primarily because the impact of stopping a trial and reporting the results cannot be predicted with any confidence: Peto (1985), in the discussion of Bather (1985), states that 'Bather, however, merely assumes . . . ''it is implicit that the preferred treatment will then be used for all remaining patients'' and gives the problem no further attention! This is utterly unrealistic, and leads to potentially misleading mathematical conclusions'. Peto goes on to argue that a serious decision-theoretic formulation would have to model the subsequent dissemination of a treatment.
3. *It depends on the context*. Whitehead (1997b, p. 208) points out that the theory of optimal decision-making only exists for a single decision-maker, and that no optimal solution exists when making a decision on behalf of multiple parties with different beliefs and utilities. He therefore argues that internal company decisions at phase I and phase II of drug development may be modelled as decision problems, but that phase III trials cannot (Whitehead, 1993).

Our personal view is that the context of evaluation often means that the investigators who design and carry out a study are generally not the same body who make decisions on the basis of the evidence (Section 3.1), and so, taking a pragmatic rather than ideological perspective, our general separation of inference and decision appears reasonable.

6.3 SPECIFICATION OF NULL HYPOTHESES

Attention in a trial usually focuses on the null hypothesis of treatment equivalence expressed by $\theta = 0$, but realistically this is often not the only hypothesis of interest. Increased costs, toxicity and so on may mean that a certain improvement would be necessary before the new treatment could be considered clinically superior, and we shall denote this value θ_S. Similarly, the new treatment might not actually be considered clinically inferior unless the true benefit were less than some threshold denoted θ_I. The interval between θ_I and θ_S has been termed the 'range of equivalence' (Freedman *et al.*, 1984); often θ_I is taken to be 0.

This is not a specifically Bayesian idea (Armitage, 1989) and can be considered as representing an interval null hypothesis. Figure 6.1 shows the

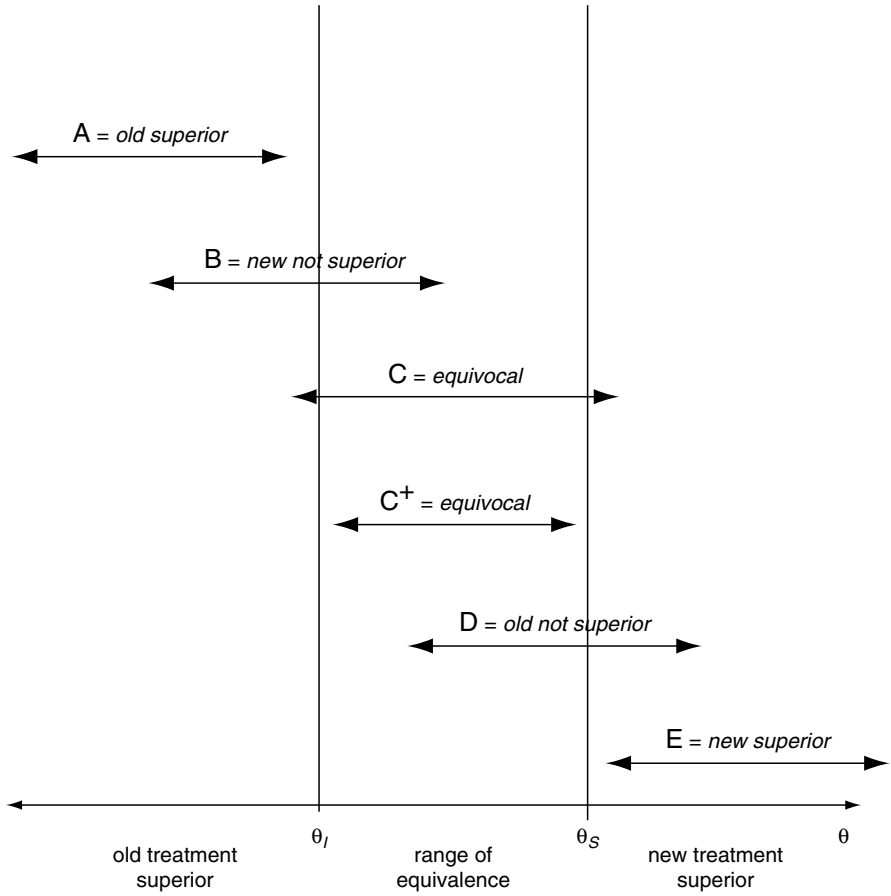

Figure 6.1 Possible situations at any point in a trial's progress, derived from superimposing an interval estimate (say, 95%) on the range of equivalence.

possible situations one could be in at any stage of a trial when calculating a 95% interval for a treatment benefit.

A: We are confident that the old treatment is clinically superior.
B: The new treatment is not superior, but the treatments could be clinically equivalent.
C: We are substantially uncertain as to the two treatments – this is essentially a position of 'equipoise'.
C^+: We are confident the two treatments are clinically equivalent – as applied to equivalence studies (Section 6.11).
D: The old treatment is not superior, but the treatments could be clinically equivalent.
E: We are confident that the new treatment is clinically superior.

It could be argued that if one really wants to convince people of the clinical superiority of a treatment, then one should aim for conclusion E in design and monitoring, even though this demands increased sample sizes and requires a highly significant (in the traditional sense) result.

Example 6.1 *CHART (continued): Clinical demands for new therapies*

References: Parmar *et al*. (1994, 2001) and Spiegelhalter *et al*. (1994). See Example 5.1 for details of the trials and the elicitation process.

Loss function or demands: No formal loss function was elicited, but a pre-trial survey was carried out of 11 clinicians participating in the trials. The clinicians were given the following instructions (Parmar *et al*., 1994):

> Suppose you had been told on good authority the exact absolute improvement [in 2-year survival rates] you would obtain by treating patients with the CHART regimen. If this was exactly zero improvement you would presumably use your standard radical radiotherapy in the future. If there was an absolute improvement of 20% you would presumably use CHART. Somewhere in between these figures there is likely to be a difference where you would change from standard therapy to CHART. There may be a range of differences where the decision would not be clearcut, i.e. a range where you feel the two regimens are approximately equivalent. Please mark your change-over point or the range on the scale of treatment differences shown below.

The upper and lower values for the ranges were averaged and the following results were obtained.

Lung trial. The participants would be willing to use CHART routinely if it conferred at least 13.5% improvement in 2-year survival (from a baseline of 15%), and unwilling if less than 11% improvement. Thus the range of equivalence is from 11% to 13.5%: from (2.33) this is equivalent to hazard ratios (HR) from 0.66 to 0.71, or log(HR) from -0.41 to -0.34.

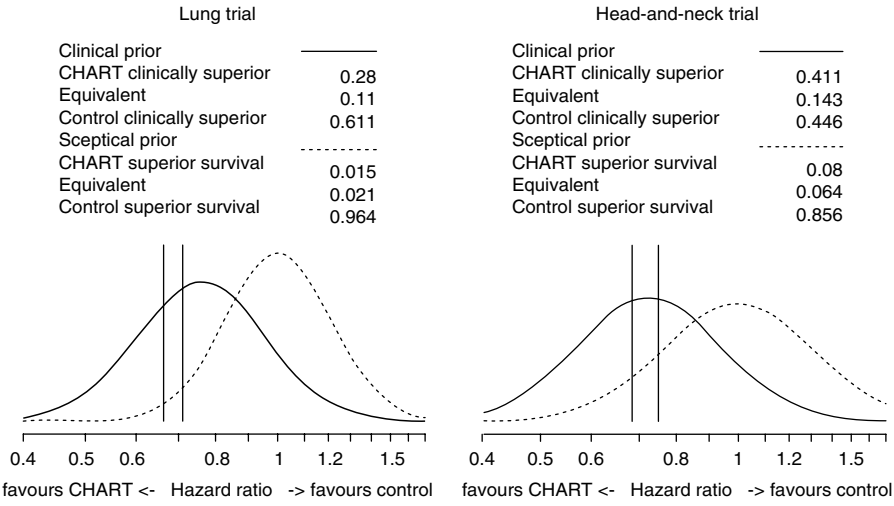

Figure 6.2 Clinical and sceptical priors superimposed on an assessed average clinical range of equivalence. Probabilities of lying below, within and above the range of equivalence are given both for clinical and sceptical priors. The juxtaposition of the clinical priors and ranges of equivalence suggests a reasonable basis for randomisation.

Head-and-neck trial. The participants would be willing to use CHART routinely if it conferred a 13% improvement in 2-year recurrence-free rate (from a baseline of 45%), and unwilling if less than 10% improvement. Thus the range of equivalence is from 10% to 13%, equivalent to HR from 0.68 to 0.75, or log(HR) from -0.38 to -0.29. The average ranges of equivalence are shown in Figure 6.2, with the clinical and sceptical priors derived previously. The average range of equivalence is reasonably central to the clinical prior, suggesting, on average, a reasonable basis for randomisation.

One advantage of the Bayesian approach is that the posterior distribution can be juxtaposed to the clinical demands being made in order to graphically display the current probabilities concerning the status of treatments. There is also no reason why the 'goalposts' shown in Figure 6.1 should not change as a study progresses and more is learnt about, for example, the side-effects of treatments. However, in order to prevent subjective bias, it may be better for those responsible for specifying the 'range of equivalence' to be blind to the data. Elicitation of such intervals can be carried out at the same time as elicitation of prior beliefs (Section 5.2) and uses very similar techniques: see Example 6.1. The crucial aspect is that those whose opinions are being elicited must be very clear in their distinction between *demands*, as expressed in their range of equivalence, and their *expectation or beliefs*, as

represented by the prior distribution. Two factors increase the potential for confusion: demands and beliefs are often quantitatively similar (indeed, we argue below that this is the ethical basis for randomisation), and the loose usage of words such as 'the difference hoped for', which carries connotations both of desire and expectation. It follows that such terms must be strictly avoided!

6.4 ETHICS AND RANDOMISATION: A BRIEF REVIEW

6.4.1 Is randomisation necessary?

Randomisation has two traditional justifications: it ensures treatment groups are directly comparable (up to the play of chance), and it provides a fundamental basis for the probability distributions underlying conventional statistical procedures. Since Bayesian probability models are derived from subjective judgement, and hence do not require any underlying physical justification for a randomisation mechanism, the latter requirement is irrelevant. This has led some to question the need for randomisation at all, provided alternative methods of balancing groups can be established. For example, Urbach (1993) argues that a 'Bayesian analysis of clinical trials affords a valid, intuitively plausible rationale for selective controls, and marks out a more limited role for randomisation than it is generally accorded'. It has even been claimed that 'Randomised trials are inherently unethical' (Berry, 1989a). Papineau (1994) refutes Urbach's position and claims that, despite it not being essential for statistical inference, experimental randomisation forms a vital role in drawing causal conclusions (Rubin, 1978). The relationship between randomisation and causal inferences is beyond the scope of this book, but in general the need for sound experimental design appears to dominate philosophical statistical issues (Hutton, 1996). In fact, Berry and Kadane (1997) suggest that if there are several parties who make different decisions and observe different data, randomisation may be a strictly optimal procedure since it enables each observer to draw their own appropriate conclusions.

The extent to which careful analysis of high-quality databases can complement or even replace randomised trials is a delicate issue: for example, Howson and Urbach (1989) and Hlatky (1991) argue in favour of databases, while Byar (1980) puts an opposing view. Although a full discussion is outside the scope of this book, we nevertheless point out that Bayesian methods provide a natural basis for synthesising data from randomised and non-randomised studies: see the discussion on the use of historical data (Section 3.16), historical controls (Section 6.9) and cross-design synthesis (Section 8.4).

6.4.2 When is it ethical to randomise?

If we agree that randomisation is useful, then the issue arises of when it is ethical to randomise. This is closely associated with the process of deciding

when to stop a trial (Section 6.6) and is often represented as a balance between *individual* and *collective* ethics (Pocock, 1992; Palmer and Rosenberger, 1999): individual ethics would suggest that it is inappropriate to randomise a patient to a treatment near the end of a trial in which one could be reasonably confident as to another treatment's superiority, while collective ethics could argue that such a benefit will only be available for future patients if the current trial runs long enough for the findings to be convincing to a wide range of clinical opinion. See Edwards *et al.* (1998) for a full review of issues concerning the ethics of randomisation in clinical trials.

Freedman (1987) introduced the idea of *professional equipoise*, in which disagreement among the medical profession makes randomisation ethical. The trial design of Kadane (1996) is an expression of this principle, in that only a treatment that at least one clinician thought optimal could be given to a patient (although unfortunately a programming error meant that some patients were allocated to treatments that *all* clinicians felt were sub-optimal). Perhaps a more appealing approach is the 'uncertainty principle' which is often argued as a basis for ethical randomisation (Byar *et al.*, 1990): this may be thought of as 'personal equipoise' in which the clinician was uncertain as to the best treatment for the patient in front of them. However, a quantified degree of uncertainty is not specified. Senn (2002) argues that it is reasonable for a society to restrict new interventions to trials, and in those trials it is ethical to randomise even when one believes in the superiority of the new treatment.

The Bayesian approach can be seen as formalising the uncertainty principle by explicitly representing, in theory, the judgement of an individual clinician that a treatment may be beneficial – this could be provided by superimposing the clinician's posterior distribution on the range of equivalence (Section 6.3) relevant to a particular patient (Spiegelhalter *et al.*, 1994). It has been argued that a Bayesian model naturally formalises the individual ethical position (Lilford and Jackson, 1995; Palmer, 1993), in that it explicitly confronts the personal belief in the clinical superiority of one treatment. Berry (1993), however, has suggested that if patients were honestly presented with numerical values for their clinician's belief in the superiority of a treatment, then few might agree to be randomised. One option might be to randomise but with a varying probability that is dynamically weighted towards the currently favoured treatment (Section 6.10).

Chaloner and Rhame (2001) consider the roles of professional and individual equipoise, and suggest scenarios which indicate different bases for ethical randomisation. Fifty-eight opinions elicited before a trial showed a wide range of responses, and the acknowledged variability in clinical opinion suggests that a suitable aim in conducting a trial is to bring disparate opinions into agreement: Chaloner and Rhame (2001) quote Byar as saying 'We may reasonably ask, if we do a study that convinces us but convinces no one else and is then ignored or requires confirmation by yet another study, whether we have really acted in the most ethical fashion in the long run'. Pocock and White (1999) consider the

situation in which one has a 'significant' effect in a trial, when further randomisation is 'unethical, but only if the statistically significant difference is genuine (in many cases it is not) and if the new treatment would indeed be given to future patients (which is by no means inevitable)'. We largely agree with the advice of Kass and Greenhouse (1989), who claim that 'the purpose of a trial is to collect data that bring to conclusive consensus at termination opinions that had been diverse and indecisive at the outset' and go on to state that 'randomisation is ethically justifiable when a cautious reasonable sceptic would be unwilling to state a preference in favour of either the treatment or the control'. This approach leads naturally to the development of sceptical prior distributions (Section 5.5.2) and their use in monitoring sequential trials (Section 6.6.2).

6.5 SAMPLE SIZE OF NON-SEQUENTIAL TRIALS

In this section we consider the Bayesian contribution to selecting the sample size of a clinical trial which will not be subject to interim monitoring: there is particular emphasis on 'hybrid' methods in which prior information is formally used but the final analysis is carried out in a classical framework. In some contexts this may be quite appropriate, as there may be substantial prior information that cannot be included in the final report for, say, regulatory purposes.

This section does contain a number of rather complex expressions for quantities of interest, but the content appears too important for this to be a 'starred' section. On a technical note, the formulae we present follow the traditional formulation in which interest focuses on a parameter θ and $\theta > 0$ indicates benefit of the experimental treatment. We recognise that in many of our examples $\theta < 0$ has represented such benefit, and furthermore in other cases we might be using thresholds other than 0. Care must therefore be taken when using the formulae in this chapter – it may be best to first transform the particular problem being analysed into the standard formulation adopted here. Details of these transformations are given in Section 6.5.4.

It could be argued that elicitation of prior beliefs and demands from a broad community of stakeholders is necessary not only in order to undertake a specifically Bayesian approach to design and analysis, but also more generally as part of good research practice. A potential consequence of ignoring this source of judgement is that trials may be designed on the basis of over-enthusiastic beliefs and demands, and hence fail to convince others and modify health-care policy or practice.

6.5.1 Alternative approaches to sample-size assessment

In Section 4.1 we described a taxonomy of six broad statistical approaches to the evaluation of health-care interventions. Here we focus on how the four main

viewpoints (ignoring the Bayesian hypothesis-testing and classical decision-theory approaches) deal with selecting the sample size of a fixed-size experiment: the design and monitoring of sequential studies will be covered in Section 6.6. A hybrid philosophy is also included.

Fisherian. In principle there is no need for preplanned sample sizes, but a choice may be made by selecting a particular precision of measurement and informally trading that off against the cost of experimentation.

Neyman–Pearson. The first stage is to set up a null hypothesis (Section 6.3), and then specify an alternative hypothesis H_A: $\theta = \theta_A$ that the trial is being designed to detect. A variety of opinions have been expressed about the interpretation of θ_A (Spiegelhalter *et al.*, 1994), including a 'minimum clinically significant difference', a 'worthwhile difference' and a difference 'thought likely to occur'. These ideas tend to conflate the demands made of the new treatment and the expectations of its benefit (Section 6.3), and this combined role of the alternative is reflected in its common definition as a difference that is 'both realistic and important' (within a Bayesian framework these properties are clearly separated). The sample size is then selected to have reasonable power to detect this alternative hypothesis. Power is generally set to 80% or 90%: formula (2.38) can be used to derive the necessary sample size in simple circumstances. In practice the choice of alternative may be influenced by available resources.

Hybrid classical and Bayesian. Considerable attention has been paid to a hybrid approach in which it is assumed that a traditional analysis will take place at the end of the trial, and the prior distribution is used solely for the design.

It may be helpful to consider the joint probability distribution of hypotheses and outcomes displayed in Table 6.1. In a traditional framework these are point hypotheses and the study is designed around the Type I error $\alpha = p((D_1|H_0)$, and the power $1 - \beta = p(D_1|H_1)$. However, if we are prepared to acknowledge prior

Table 6.1 Joint probability distribution of hypotheses and outcomes of a hypothesis test.

		Truth		
		H_0	H_1	
Outcome	D_0 : do not reject H_0	$p(D_0, H_0) =$ P(correct negative)	$p(D_0, H_1) =$ P(false negative)	$p(D_0)$
	D_1 : reject H_0	$p(D_1, H_0) =$ P(false positive)	$p(D_1, H_1) =$ P(correct positive)	$p(D_1)$
		$p(H_0)$	$p(H_1)$	1

probabilities for the hypotheses, then it would appear reasonable to focus also on the probability of rejecting H_0 *and* this being the correct decision, *i.e.* the joint probability $p(D_1,H_1)$. Since $p(D_1,H_1) = p(D_1|H_1)\ p(H_1) = (1-\beta)\ p(H_1)$, this simply means adjusting the power by the initial probability of H_1: the problem with using only the conditional power $p(D_1|H_1)$ is that no account is taken of the plausibility of the alternative and hence there is a temptation to delude oneself into designing trials to detect implausible hypotheses.

The unconditional probability of getting a 'positive' conclusion can be expressed as

$$p(D_1) = p(D_1, H_0) + p(D_1, H_1),$$

and the first term, which is the probability $p(D_1,\ H_0) = p(D_1|H_0)\ p(H_0)$ of a false positive result, will generally be very small provided that $\alpha = p(D_1|H_0)$ is small and the prior opinion is substantially supportive of H_1 (as will often be the case preceding a trial). Thus

$$p(D_1) \approx p(D_1|H_1)\ p(H_1); \tag{6.1}$$

and so the 'prior-adjusted power' $(1-\beta)\ p(H_1)$ will often also be close to the unconditional probability of the trial getting a 'significant' result.

Things get a little more complicated in the more general case when the hypotheses are composite, for example H_0: $\theta < 0$ and H_A: $\theta > 0$. Here the classical power is given by a curve $p(D_1|\theta)$, and we wish to make use of a continuous prior distribution $p(\theta)$.

A number of means of incorporating the prior are possible.

1. One can plot the conditional power curve and superimpose the prior distribution as an informal guide to the relative plausibility of alternative hypotheses. This might prevent a study being designed around an alternative that was clearly grossly optimistic.
2. The prior mean μ might simply be taken as a point alternative hypothesis θ_A, representing a 'plausible and worthwhile difference', although this does not acknowledge the current uncertainty about θ expressed by the prior.
3. The whole classical power curve $p(D_1|\theta)$ can be averaged with respect to the prior distribution to obtain an 'expected' or 'average' classical power $p(D_1) = \int p(D_1|\theta)\ p(\theta)\ d\theta$. This will give the unconditional probability of rejecting H_0. From the discussion above, we might expect this to be a reasonable approximation to the prior-adjusted power $p(D_1,H_1)$ if $p(\theta)$ does not give substantial probability to values of $\theta < 0$.
4. The classical power curve can be averaged with respect to the prior distribution $p(\theta|H_1) = p(\theta|\theta > 0)$, *i.e.* conditional on H_1 being true (since $p(\theta|\theta > 0) = p(\theta,\ \theta > 0)/p(\theta > 0)$, this can be obtained by restricting the prior to $\theta > 0$ and renormalising it to have total probability 1). Brown *et al.*

(1987) recommend this technique as predicting the chance of *correctly* detecting a positive improvement, rather than the overall chance $p(D_1)$ of getting a positive result regardless of the truth. But this method suffers from the same difficulty as the original classical power calculation, in that no account is taken of the plausibility of H_1.

5. The predictive distribution over the possible powers could be displayed as an aid to deciding appropriate sample sizes.

We shall illustrate these options in the following sections, using normal likelihoods and priors.

Prior distributions might be from any of the sources described in Chapter 5, for example subjective assessments (Ten Centre Study Group, 1987), a single previous study (Brown *et al.*, 1987), or a meta-analysis of previous results (DerSimonian, 1996): Example 6.4 illustrates the use of subjective opinion. Most of the applications have assumed a conventional analysis, although Bryant and Day (2000) suggest that a suitable Bayesian perspective is for a trial to be large enough to enable a sceptic and an enthusiast to be brought into consensus.

Finally, it is natural to express a cautionary note on projecting from previous studies (Korn, 1990), and possible techniques for discounting past studies are very relevant (Section 5.4).

Proper Bayesian. As in the Fisherian approach, there is in principle no need for preplanned sample sizes (Lilford *et al.*, 1995). Alternatively, it is natural to focus on the eventual precision of the posterior distribution of the treatment effect: for normal assumptions this is straightforward to calculate. There is an extensive literature on non-power-based Bayesian sample-size calculations (Joseph *et al.*, 1997).

When working within a hypothesis-testing framework, all the above discussion on hybrid classical and Bayesian methods holds, except that the final conclusion of whether the result is 'significant' or not will be based on a posterior distribution rather than a classical analysis. One is still faced with a variety of means of incorporating the prior distribution, although since the conclusions are going to include that prior it seems natural to use its full form and calculate expected power. The necessary formulae for normal likelihoods and priors are provided in Section 6.5.3.

Lee and Zelen (2000) propose a method based on obtaining a high posterior probability of an effective treatment after a 'significant' result, using the analysis described in Section 3.10, *i.e.* by trying to fix $p(H_1|D_1)$. This has been criticised by Simon (2000) and Bryant and Day (2000) as ignoring the actual data observed and hence violating the likelihood principle.

Decision-theoretic Bayesian. If we are willing to express a utility function for the cost of experimentation and the potential benefit of the treatment, then

sample sizes can be chosen to maximise the expected utility. Lindley (1997) and discussants argue strongly for this position. Detsky (1985) conducted an early attempt to model the impact of a trial in terms of future lives saved, which required modelling beliefs about the future number to be treated and the true benefit of the treatment, while Claxton *et al.* (2000) and Gittins and Pezeshk (2000), for example, show how sample sizes could be explicitly determined by a trade-off between the cost of the trial and the expected future benefit: for further references, see Section 6.13. This approach also attempts to answer the question 'what is the expected net benefit from carrying out the trial?' (Section 9.10). An intermediate 'information-theoretic' position is taken by Lindley (1997) who does not attempt to model the future benefit of a trial, and instead trades off the information in the posterior distribution against the cost of sampling.

6.5.2 'Classical power': hybrid classical–Bayesian methods assuming normality

We now assume we have a prior distribution to use in our study design, but that the conclusions of the study will be entirely classical and will not make use of the prior, perhaps because of submission to a regulatory authority. Suppose we have a normal prior $\theta \sim \mathrm{N}[\mu,\ \sigma^2/n_0]$ and our future data Y_n have distribution $Y_n \sim \mathrm{N}[\theta,\ \sigma^2/n]$, and we wish to calculate the predictive probability of obtaining a classically 'significant' result when testing the null hypothesis $\theta < 0$. Under a classical analysis (Section 2.5), H_0 will be rejected when the parameter estimate Y_n obeys

$$Y_n > -\frac{1}{\sqrt{n}} z_\epsilon \sigma; \tag{6.2}$$

this event, denoted S_ϵ^C, will occur with probability

$$P(S_\epsilon^C|\theta) = \Phi\left[\frac{\theta\sqrt{n}}{\sigma} + z_\epsilon\right], \tag{6.3}$$

which is the classical power curve previously given in (2.37).

We can plot (6.3) superimposed on the prior $p(\theta)$, which can reveal the relative plausibility of the potential alternative hypotheses and suggest whether the trial is based on over-optimistic assumptions (see Example 6.2). If we wish to calculate the overall unconditional probability of a 'significant' result S_ϵ^C we can integrate (6.3) with respect to the prior. However, it is analytically more straightforward to use the the predictive distribution (3.23)

$$Y_n \sim \mathrm{N}\left[\mu, \sigma^2\left(\frac{1}{n_0}+\frac{1}{n}\right)\right]$$

to directly evaluate the chance of the critical event (6.2) occurring, which can be shown to be

$$P(S_\epsilon^C) = \Phi\left[\sqrt{\frac{n_0}{n_0+n}}\left(\frac{\mu\sqrt{n}}{\sigma}+z_\epsilon\right)\right]. \tag{6.4}$$

The relationship to the power curve (6.3) is clear. As $n_0 \to \infty$, the prior tends to a lump on μ and $P(S_\epsilon^C)$ tends to the classical power evaluated at the prior mean μ. However, finite n_0 will mean that the expected power is less than the classical power evaluated at the prior mean μ, provided the classical power is greater than 50%. This may be a more realistic assessment of the chance that the trial will yield a positive conclusion.

We note that Table 6.1 can be extended to allow 'equivocal' decisions, and that the necessary probabilities can be calculated using tail areas of the bivariate normal distribution (Spiegelhalter and Freedman, 1986).

Example 6.2 *Bayesian power: Choosing the sample size for a trial*

We revisit Example 2.6, in which a trial for a new cancer treatment is designed to have 80% power to detect a log(hazard ratio) $\theta_A = 0.56$, requiring 100 events when assuming a two-sided α of 0.05. Consider an archetypal enthusiastic prior (Section 5.5.3) centred on the alternative hypothesis and with 5% prior probability that $\theta < 0$. Hence $\theta \sim \mathrm{N}\,[\mu, \sigma^2/n_0]$ where $\mu = 0.56, \sigma = 2$ and $\mu - 1.645\sigma/\sqrt{n_0} = 0$, so that $n_0 = 1.645^2\sigma^2/\mu^2 = 34.5$. The classical power curve and the prior are shown on Figure 6.3: the power at the prior mean is 80% as designed, the expected power (6.4) averaging over the entire prior distribution is 0.66, showing the decline from the conditional value of 0.80. If we took the approach recommended by Brown *et al*. (1987) we would average the power curve with respect to the conditional prior $p(\theta|H_1) = p(\theta|\theta > 0)$; this is not straightforward to calculate and is perhaps easiest to evaluate using Monte Carlo methods (Section 3.19.1), from which we find, using the notation of Table 6.1, that $p(D_1|H_1) = 0.70$. Such a value might have been predicted, since we know that $p(H_1) = 0.95$, $p(D_1) = 0.66$, and from (6.1) that $p(D_1) \approx p(D_1|H_1)p(H_1)$.

6.5.3 'Bayesian power'

Suppose we have the same normal prior and likelihood as in Section 6.5.2 but now wish to carry out a fully Bayesian analysis in which the prior will be

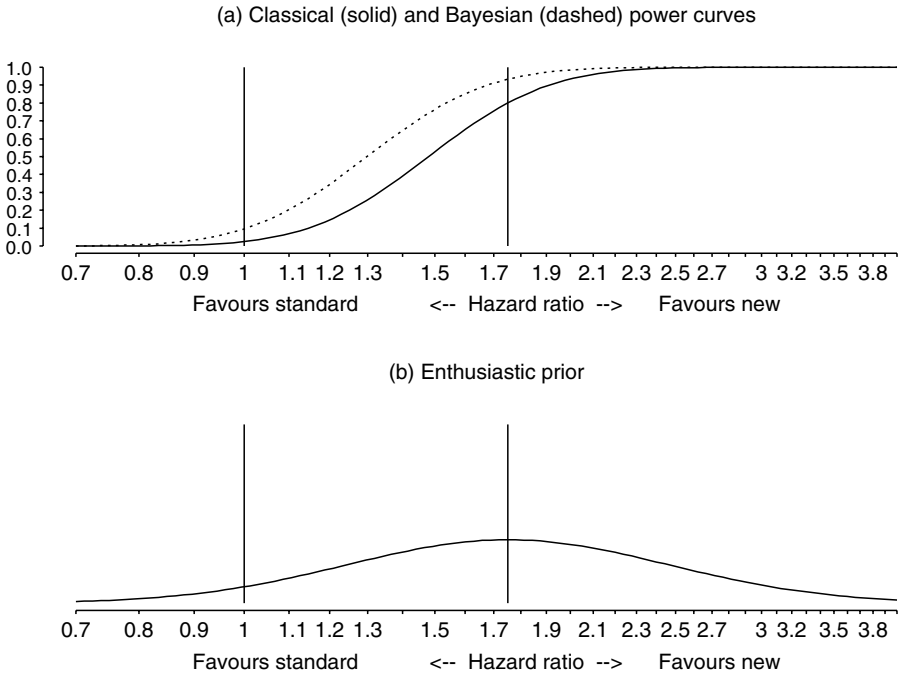

Figure 6.3 Power curves (a) for testing $H_1: \theta > 0$, designed to have classical power of 80% at $\theta_A = 0.56$ (HR = 1.75). The Bayesian power curve in (a) assumes that the enthusiastic prior shown in (b) is to be included in the analysis.

incorporated. We wish to calculate the predictive probability of obtaining a 'significant' Bayesian result when testing the null hypothesis $\theta < 0$ against an alternative $\theta > 0$, and we shall denote such 'Bayesian significance' as $S_\epsilon^B \equiv P(\theta < 0|\text{data}) < \epsilon$.

Assuming a future parameter estimate Y_n, we will obtain the posterior distribution

$$\theta|Y_n \sim \mathrm{N}\left[\frac{n_0\mu + nY_n}{n_0 + n},\ \frac{\sigma^2}{n_0 + n}\right],$$

and so S_ϵ^B will occur when the parameter estimate Y_n obeys

$$Y_n > \frac{-\sqrt{n_0 + n}\ z_\epsilon\ \sigma - n_0\mu}{n}. \tag{6.5}$$

For a particular true value of θ, $Y_n \sim \mathrm{N}[\theta, \sigma^2/n]$, and hence it can be easily shown that this event will occur with probability

$$P(S_\epsilon^B|\theta) = \Phi\left[\frac{\theta\sqrt{n}}{\sigma} + \frac{\mu n_0}{\sigma\sqrt{n}} + \sqrt{\frac{n_0+n}{n}}z_\epsilon\right]. \tag{6.6}$$

With vague prior opinion, $n_0 \to 0$ and we are left with the standard classical power curve given in (2.37).

Just as in Section 6.5.2, we can plot (6.6) superimposed on the prior $p(\theta)$. To calculate the overall unconditional probability of a 'significant' result S_ϵ^B it is again analytically more straightforward to use the the predictive distribution of Y_n to evaluate the chance of the critical event (6.5) occurring:

$$\begin{aligned} P(S_\epsilon^B) &= P\left(Y_n > \frac{-\sqrt{n_0+n}z_\epsilon\sigma - n_0\mu}{n}\right) \\ &= \Phi\left[\frac{\mu\sqrt{n_0+n}\sqrt{n_0}}{\sigma\sqrt{n}} + \sqrt{\frac{n_0}{n}}z_\epsilon\right]. \end{aligned} \tag{6.7}$$

Example 6.3 *Bayesian power (continued): Choosing the sample size for a trial*

If we are willing to include the prior distribution in the analysis then we obtain the Bayesian power curve (6.6) shown as a dashed line in Figure 6.3(a), which is substantially higher than the classical power curve due to the prior giving a 'head start'. The power at the alternative hypothesis $\theta_A = 0.56$ is 0.93, while the chance of a false rejection of $\theta = 0$ has risen from 0.025 to 0.10 – this inflated chance of a Type I error illustrates the danger of getting the prior 'wrong'. The expected Bayesian power (6.7), averaged with respect to the prior distribution in Figure 6.3(b), is 0.78.

6.5.4 Adjusting formulae for different hypotheses

All the formulae provided so far have assumed that $\theta > 0$ indicates superior performance of the innovative treatment and therefore is the alternative hypothesis of interest – this has simplified the exposition but clearly will not hold in all situations. One option is to redefine the outcome measures and parameters so that θ has the required properties. Alternatively, one can transform the formulae provided, and we now consider the necessary transformations when different hypotheses are being considered.

- ***Non-zero threshold***. Suppose the null hypothesis is $H_0: \theta < \theta_0$ and the alternative $H_1: \theta > \theta_0$. Each of the previous formulae can be transformed by subtracting θ_0 from the prior mean μ, the observed statistic y_m and, in conditional power calculations, the parameter θ. For example, suppose in

Example 6.2 that the threshold of interest was changed to $\theta = 0.2$, *i.e.* the posterior interval would need to lie wholly above a log(hazard ratio) of $0.2\,(\mathrm{HR} = 1.22)$ before H_0 is rejected. The conditional power at the alternative hypothesis $\theta_A = 0.56$ is now only 0.56, obtained from transforming (6.6), while the expected power is found from (6.7) to be 0.53.

- ***Reversal of hypotheses***. As we have seen in most of our examples, it is common to express benefit from the new intervention as a reduction in risk, and hence on a logarithmic scale to set $H_1: \theta < 0$. Thus a 'significant' result will be obtained if a final interval lies wholly below 0. If, for example, we were adopting a fully Bayesian approach this would be equivalent to the event $P(\theta > 0|\mathrm{data}) < \epsilon$, which we shall denote S_ϵ^{B-}. Now

$$S_\epsilon^{B-} \equiv [P(\theta > 0|\mathrm{data}) < \epsilon] \equiv [P(\theta < 0|\mathrm{data}) > 1 - \epsilon]$$

and hence, for example,

$$P(S_\epsilon^{B-}) = 1 - P(S_{1-\epsilon}^{B}).$$

Therefore the formulae provided can be transformed by substituting $1 - \epsilon$ for ϵ, and subtracting the result from 1.

For example, suppose in Example 6.2 that the threshold of interest was changed to $\theta = 0.69$, $\mathrm{HR} = 2$, and furthermore we were interested in the expected power to reject the null hypothesis $H_0: \theta > \theta_0$, *i.e.* we are interested in values of θ with an odds ratio less than 2. Using both transformations on (6.7) leads to

$$P(S_\epsilon^{B-}) = 1 - P(S_{1-\epsilon}^{B}) = 1 - \Phi\left[\frac{(\mu - \theta_0)\sqrt{n_0 + n}\sqrt{n_0}}{\sigma\sqrt{n}} + \sqrt{\frac{n_0}{n}} z_{1-\epsilon}\right]. \qquad (6.8)$$

Then from (6.8) we find the expected power is 0.24: such a low value might be anticipated from the substantial prior support for H_0.

Example 6.4 *Gastric: Sample size for a trial of surgery for gastric cancer*

Reference: Fayers *et al*. (2000).

Intervention: Radical (D2) compared to conventional (D1) surgery for gastric cancer.

Aim of study: Evidence from Japan suggested that more radical surgery was a possible explanation for the better survival rates of patients with gastric cancer, and the UK Medical Research Council initiated a randomised trial to compare survival following radical and conventional surgery.

Study design: Two-group parallel RCT.

Outcome measure: Hazard ratio of death (HR > 1 favours radical treatment).

Planned sample size: The trial was designed under the assumption that the minimum clinically significant difference was a 13.5% improvement in 5-year survival from 20% to 33.5% in patients undergoing conventional surgery – this value for the alternative hypothesis was based on the opinion of the trial team. This is equivalent to a hazard ratio of $\log(0.20)/\log(0.335) = 1.47$ (Section 2.4.2), or $\log(\text{HR}) = 0.39$. For the trial to be able to detect a 13.5% difference at the 5% significance level with 90% power, the necessary number of events (*i.e.* deaths) is $n = \sigma^2(1.96 + 1.28)^2/0.39^2 = 276$, when taking $\sigma = 2$ (Section 2.4.2 and (2.38)). The trial was designed to have 200 patients per arm which was predicted to yield this number of events.

Statistical model: For planning purposes, the normal approximation of Section 2.4.2 was adopted, while for analysis a full Cox regression was used to obtain a likelihood for log(HR).

Prospective analysis?: Yes.

Prior distribution: In addition to the three surgical members of the trial steering committee, a further 23 surgeons had their beliefs regarding the likely benefit/harm of radical compared to conventional surgery elicited, both at the start of the trial and later when the trial had stopped but had not yet been published. Fayers *et al*. (2000) shows each individual's prior distribution on a scale representing improvement in 5-year survival, elicited using a similar questionnaire to that of Parmar *et al*. (1994); see Example 5.1. The average distribution had a prior mean of 9.4% improvement over their average assessed control 5-year survival of 21%, although skewness in the distributions gives rise to a median of around 4%. Assuming a baseline survival of 21%, the distribution for an improvement p can be transformed to a log(HR) scale by $\log(\text{HR}) = \log(\log(0.21)/\log(0.21 + p))$ as in Example 5.1: fitting a normal distribution to the transformed histogram yields a prior with mean $\mu = 0.12$ and standard deviation $\sigma/\sqrt{n_0} = 0.19$, and so $n_0 = 4/0.19^2 = 111$. This corresponds to a hazard ratio of 1.13 (95% interval from 0.78 to 1.64). This distribution is shown in Figure 6.4(a), revealing that the probability of exceeding the alternative hypothesis of HR $= 1.47$ is 8%. Hence, the overall prior beliefs for the surgeons reveal the trial has been designed around a rather optimistic target.

Figure 6.4(b) shows the power curve (6.3) for the trial based on an expected $n = 276$ events, with 90% power at the alternative hypothesis of 1.47. Juxtaposing with Figure 6.4(a) shows that the surgeons' belief is

concentrated in an area of rather low power. Indeed, (6.4) shows that the expected power is only 30%, which rises marginally to 31% if a Bayesian final analysis is undertaken (6.7). Even if the surgeons were considerably more optimistic, and their prior mean was set to the alternative hypothesis of HR $= 1.47$, then the expected power would rise to only 45%.

Loss function or demands: No, but as well as eliciting the beliefs of the surgeons, the authors elicited their demands for radical surgery: around a 10% improvement was judged to be necessary before wishing to routinely implement the more radical surgery, which is more extensive and has extra risk of complications and resource usage.

Computation/software: Conjugate normal model.

Evidence from study: The trial recruited the full 200 patients on each arm, and eventually 281 events were observed (137 under D1, 144 under D2), with a result slightly in favour of the conventional surgery. The observed hazard ratio, based on a Cox regression, was 0.91 (95% CI from 0.72 to 1.15), equivalent to a log(HR) of -0.09 (standard error 0.11, equivalent to an effective number of events of $m = \sigma^2/0.11^2 = 278$, almost exactly the same as the actual number of events observed). The 5-year survival rate in those patients undergoing conventional surgery was 30%, considerably higher than the 20% expected before the trial started. This likelihood is displayed in Figure 6.4(c).

Bayesian interpretation: Figure 6.4(d) displays the predictive distribution for the observed hazard ratio, derived using the methods described in Section 3.13. The probability of observing a result as extreme as that observed is 0.32, twice the shaded area shown in Figure 6.4(d). From Section 5.8 this is Box's measure of conflict between prior and likelihood, and is not particularly extreme even though the prior expectation of a benefit from D2 conflicted with the observed hazard ratio.

Comments: Fayers *et al*. (2000) carried out a second elicitation exercise when the trial was complete but before the results were announced, and found there was still considerable optimism among the clinical collaborators. They conclude that although opinions change over time, those involved in a clinical trial tend to be optimistic and if their prior expectations are used as a naive basis for sample-size calculations, the trial could result in too small a sample size. Nevertheless, in this example the alternative hypothesis was judged to be optimistic even by the participants. A more realistic assessment of the trial's chances of success might be made by taking into account their full uncertainty.

It is also important to monitor such a trial so that it does not continue unnecessarily – in this example the trial might have been stopped and

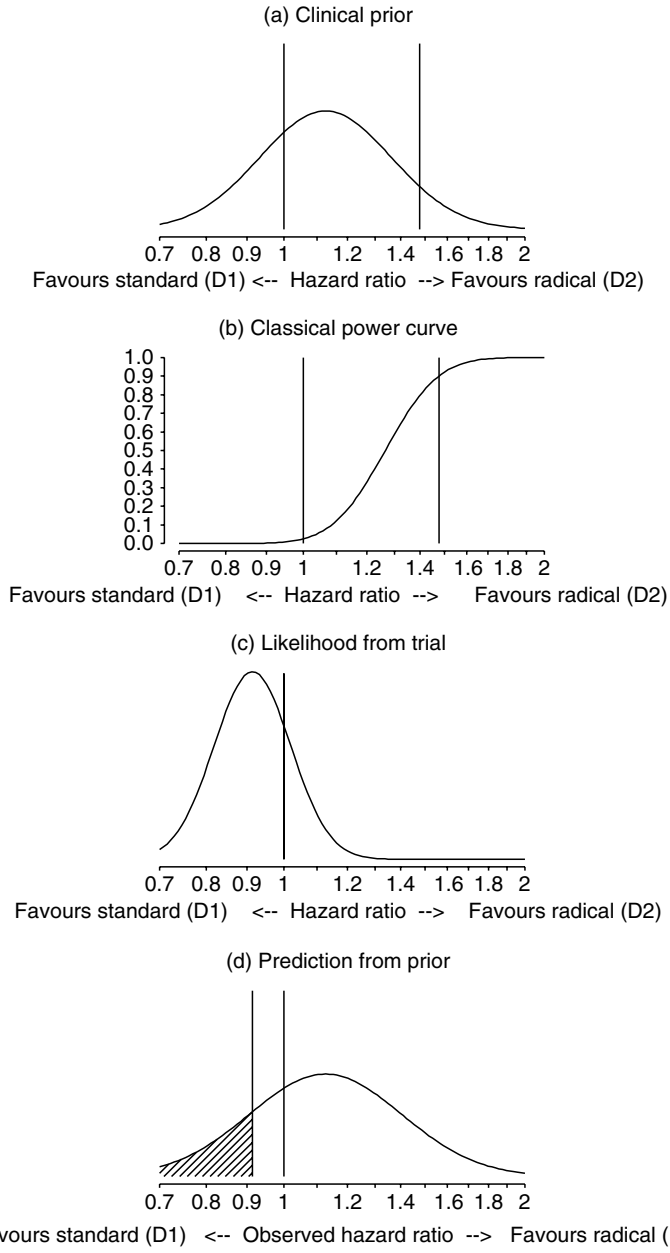

Figure 6.4 The prior assessment (a) for D2 trial in gastric cancer surgery shows some expectation of benefit, but the alternative hypothesis of 1.47 around which the trial has been designed is clearly very optimistic (b). The eventual trial result (c) showed no clear evidence for benefit. The predictive distribution derived from the prior (d) shows that the observed result (HR = 0.91) was not particularly surprising, given the prior opinion as expressed by (a).

rejected an 'important difference' some time before the eventual conclusion. However, as we shall see in Section 6.6.2, it may be more appropriate to monitor using the clinical prior, in order to ensure that the negative finding is convincing even to enthusiasts.

6.5.5 Predictive distribution of power and necessary sample size

Consider the classical power formula given in (2.37). If we express uncertainty over the parameters as a prior distribution, then the power can be considered as an unknown quantity with a distribution induced by this prior. This *predictive* distribution over the power can best be obtained by simulation methodology: essentially the unknown parameters are simulated from their prior distribution, plugged into the formula for the power, and the result recorded. After many iterations of this procedure a distribution over possible powers is obtained. This is essentially a *Monte Carlo* procedure (Section 3.19.1) and is illustrated in Example 6.5.

Example 6.5 *Uncertainty: Predictive distribution of power*

Assume that a randomised trial is planned with n patients in each of two arms, using a response with standard deviation $\sigma = 1$; hence, the variance of a contrast between two patients is $2\sigma^2$. The trial is aimed to have Type I error (two-sided α) of 5%, and 80% power to detect a true difference of $\theta = 0.5$ in mean response between the groups.

From (2.38) the necessary sample size per group is

$$n = \frac{2\sigma^2}{\theta^2}(z_{0.8} - z_{0.025})^2$$

where $z_{0.8} = 0.84$, $z_{0.025} = -1.96$; note that this differs slightly from (2.38) as here σ is the standard deviation of a single response.

The necessary sample size is $n = 63$. Suppose, however, that we wish to express uncertainty concerning both θ and σ. For θ we assess a prior mean of 0.5 and prior standard deviation of 0.1, while for σ we assume a prior mean of 1 and standard deviation of 0.3. θ and σ are assumed to be independent and normally distributed (subject to the constraint of σ being positive).

Using Monte Carlo methods we simulate values of θ and σ from their prior distributions, substitute them in the sample-size formula above, and so obtain a predictive distribution over n. This distribution has the properties shown in Table 6.2 and is plotted in Figure 6.5 – it is clear that there is huge uncertainty as to the appropriate sample size.

Table 6.2 Properties of predictive distributions of necessary sample size n for fixed power of 80%, and power for fixed sample size $n = 63$.

	Median	95% interval
n	62.5	9.3 to 247.2
Power (%)	80	29 to 100

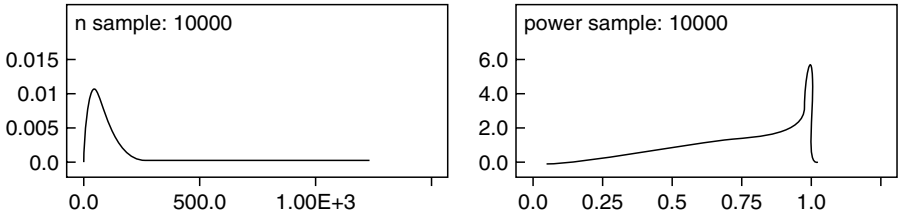

Figure 6.5 Predictive distributions from WinBUGS for necessary sample size n to achieve 80% power, and power for $n = 63$ patients per group.

For fixed n, the power is

$$\text{power} = \Phi\left(\sqrt{\frac{n\theta^2}{2\sigma^2}} + z_{0.025}\right).$$

If we decide to use 63 patients per group, we can simulate potential values for the power using the same methodology. The results are again presented in Table 6.2 and plotted in Figure 6.5, and show that although the median power is 80%, a trial of 63 patients per group could be seriously underpowered. We can calculate other quantities that could give insight into the planned sample size: for example, that there is a 37% chance that the power is less than 70%.

6.6 MONITORING OF SEQUENTIAL TRIALS

6.6.1 Introduction

Whether or not to stop a trial early is a complex ethical, financial, organisational and scientific issue, in which statistical analysis plays a considerable role. Section 4.3 has already demonstrated that sequential analysis might be considered the 'front line' between Bayesian and frequentist approaches, and the monitoring of sequential trials has been said to reach 'to the very foundations of the two paradigms' (Etzioni and Kadane, 1995).

Recommendations concerning early stopping or changes in the conduct of trials increasingly rest in the hands of independent committees known as data and safety monitoring boards or data monitoring committees (DMC). We shall adopt the latter term. In Section 6.6.6 we shall discuss the relevance of the Bayesian perspective to the deliberations of a DMC, where we shall emphasise the ability to incorporate external evidence and formally account for the desire to bring the trial to a conclusive result.

Four main statistical approaches can be identified, again corresponding to the four main entries in Table 4.1:

- *Fisherian*. This is perhaps best exemplified in trials influenced by the Clinical Trial and Services Unit in Oxford, in which protocols generally state (Collins *et al.*, 1995) that the DMC should only alert the steering committee to stop the trial on efficacy grounds if there is '*both* (a) ''proof beyond reasonable doubt'' that for all, or for some, types of patient one particular treatment is clearly indicated . . . *and* (b) evidence that might reasonably be expected to influence the patient management of many clinicians who are already aware of the results of other main studies'. There is no formal expression of what evidence is required to establish 'proof beyond reasonable doubt' (although $2P < 0.001$ is mentioned as a possible criterion). We also note the explicit, though again informal, appeal to the idea that the results should be convincing to a broad spectrum of opinion, and its close relation to the quote by Kass and Greenhouse (1989) on the need for trials to bring 'conclusive consensus' (Section 6.4.2).
- *Neyman–Pearson*. This classical method attempts to retain a fixed Type I error through prespecified stopping boundaries or guidelines which may be used at prespecified analysis times ('group-sequential methods') or with continuous monitoring. Group-sequential methods boundaries include those of O'Brien and Fleming, which are very conservative at early interim analyses, and Pocock, which have constant nominal 'significance', while continuous methods include alpha-spending functions and triangular boundaries. See Whitehead (1997a) for a detailed review. DeMets (1984) states that 'while they are not stopping rules, such methods can be useful in the decision-making process', although regulatory authorities require good reasons for not adhering to such boundaries (International Conference on Harmonisation E9 Expert Working Group, 1999).

  Objections to this approach from both Fisherian and Bayesian perspectives have already been covered in Section 4.3. In addition, there is no agreed method of estimation following a sequential trial (Freedman, 1996), although frequentist sequential rules are 'prone to exaggerate magnitude of treatment effect' (Pocock and Hughes, 1989) since they would tend to stop when on a random high; Pocock and White (1999) term the tendency for early extreme results to become less impressive as 'regression to the truth'. Armitage (1991a) agrees that adjusted *P*-values are 'too tenuous to be quoted in an

authoritative analysis of the data', but still considers frequency properties of stopping rules may be useful guides for 'mental adjustment'.

In practice, a DMC will need to take into account multiple sources of evidence when making its judgement and, if working within the traditional Neyman–Pearson paradigm, classical sequential analysis may be a useful warning against over-interpretation of naive *P*-values. Freidlin *et al.* (1999) provide a useful analysis, pointing out that the role of a trial is to change practice and warning of over-strict adherence to formal stopping procedures.

- *Proper Bayesian*. Probabilities derived from a posterior distribution may be used for monitoring, without formally prespecifying a stopping criterion or even prespecifying a sample size (Berry, 1993). It is natural to use the posterior probabilities of hypotheses of interest as a basis for monitoring (Section 6.6.2), although this may be supplemented by making predictions of the possible consequences of continuing (Section 6.6.3). As for trials with fixed sample size, a hybrid strategy is possible in which prior distributions may be used at the design stage but assuming a Neyman–Pearson analysis (McPherson, 1982). However, if external evidence becomes available during a clinical trial it can be argued that this should be incorporated into a prior distribution.

  There is no direct implication of the Bayesian approach on trial size. Matthews (1995) and Edwards *et al.* (1997) have suggested that small, open trials fit well into a Bayesian perspective in which all evidence contributes and there is no demand for high power to reject hypotheses. Alternatively, monitoring with a sceptical prior may demand larger than standard sample sizes in order to convince an archetypal sceptic about treatment superiority.

- *Decision-theoretic Bayesian*. This assumes we are willing to explicitly assess the losses associated with consequences of stopping or continuing the study, and therefore the trial requires a full specification of the 'patient horizon', the allocation rule and so on. This approach also quantifies the expected benefit of the trial and therefore helps decide whether to conduct the trial at all – see Sections 6.6.4 and 9.10.

6.6.2 Monitoring using the posterior distribution

Following the 'proper Bayesian' approach, it is natural to consider terminating a trial when one is confident that one treatment is better than the other, and this may be formalised by assessing the posterior probability that the treatment benefit θ lies above or below some boundary, such as the ends of the range of equivalence described in Figure 6.1. For example, when comparing two treatments in which θ represents success rates, we might consider stopping in favour of the new treatment and concluding $\theta > 0$ when the posterior probability that $\theta < 0$ is less than some threshold ϵ (we note we are not using α to denote our tail area in order to avoid confusion with expressions for Type I error). In

Section 6.5.3 we denoted this event S_ϵ^B, and for normal prior and likelihood this will occur if the parameter estimate y_m obeys

$$y_m > \frac{-\sqrt{n_0 + m}\; z_\epsilon\; \sigma - n_0\mu}{m}; \tag{6.9}$$

this is equivalent to (6.5) but seen as a retrospective assessment of observed data y_m rather than a prospective view of future data Y_n. Applications of this procedure have been reported in a wide variety of trials (Section 6.13).

We have already discussed how a well-designed trial should contain sufficient evidence to bring both a sceptic and an enthusiast to broadly the same conclusions (Section 6.4.2) as to whether the treatment is effective or not. This idea may be formalised in the following way, using the concept of sceptical and enthusiastic priors (Section 5.5).

- First, stopping with a 'positive' result (*i.e.* in favour of the new treatment) might be considered if a posterior based on a *sceptical* prior suggested a high probability of treatment benefit.
- Second, stopping with a 'negative' result (*i.e.* that is equivocal or in favour of the standard treatment) may be based on whether the results were sufficiently disappointing to make a posterior based on an *enthusiastic* prior rule out a treatment benefit.

In other words, we should stop if we have convinced a reasonable adversary that they are wrong. Fayers *et al.* (1997) provide a tutorial on such an approach, and Example 6.6 describes its application by a DMC for two cancer trials. In addition, Example 6.7 considers a trial in which the data overwhelmed an optimistic prior centred on a 40% risk reduction, and hence justified assuming a negative result and early stopping with a conclusion of no treatment benefit.

It is worth considering in more detail the use of a sceptical prior as a basis for monitoring, particularly as it encourages an explicit comparison with classical sequential methods. Suppose we assume a sceptical prior for a treatment difference

$$\theta \sim \mathrm{N}\left[0, \frac{\sigma^2}{n_0}\right],$$

and we would consider stopping the trial when the event S_ϵ^B occurs, *i.e.* $P(\theta < 0|\text{data}) < \epsilon$, or equivalently when a symmetric $100(1 - 2\epsilon)\%$ interval lies wholly above 0. From (6.9) this will occur when

$$y_n > \frac{-\sqrt{n_0 + m}\; z_\epsilon\; \sigma}{m}. \tag{6.10}$$

Let $z_m = y_m\sigma/\sqrt{m}$ be the standardised classical test statistic. Then (6.10) can be rearranged as

$$z_m > -z_\epsilon \sqrt{1 + \frac{n_0}{m}} \,. \tag{6.11}$$

The term $\sqrt{1 + n_0/m}$ is a multiplier of the 'naive' critical value $-z_\epsilon$, and demonstrates how the sceptical prior opinion introduces conservatism through increasing the critical value.

Suppose $2\epsilon = 0.05$ and hence $-z_\epsilon = 1.96$, and the maximum intended sample size of the trial is n. In Section 5.5.2 we argued that a reasonable 'handicap' might be $n_0/n = 0.26$, based on a trial with 90% power to detect an 'optimistic' difference. Substituting into (6.11), we stop and reject H_0 when

$$z_m > 1.96 \sqrt{1 + 0.26 \frac{n}{m}} \,. \tag{6.12}$$

The boundary is a function solely of the proportion m/n of the trial that has been completed, and is shown in Figure 6.6. Assuming a sceptical prior thus

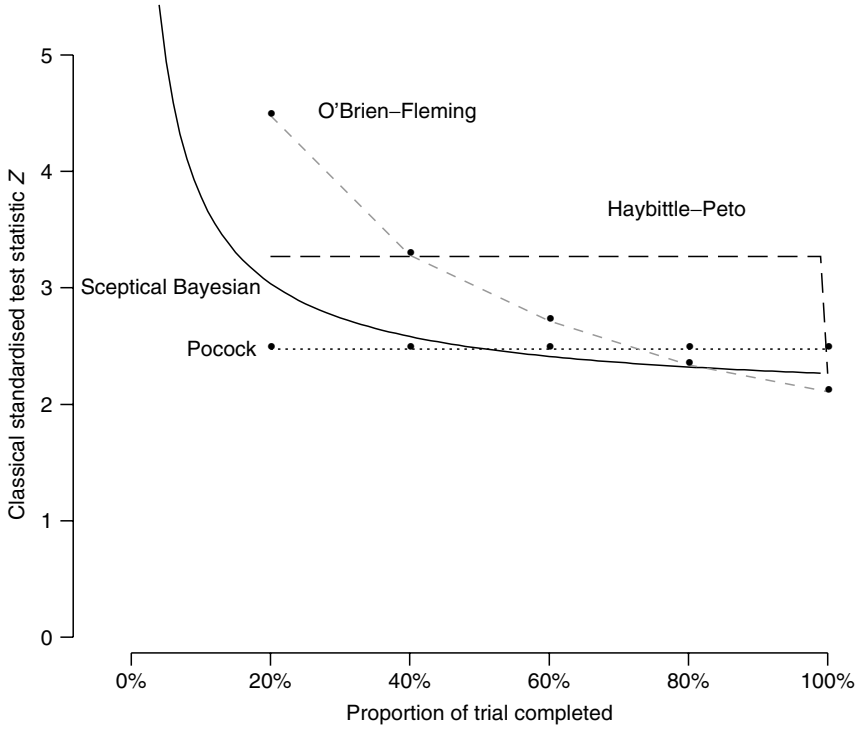

Figure 6.6 Monitoring boundaries for a sceptical prior opinion with $2\epsilon = 0.05$ and handicap 0.26. This is compared to Pocock and O'Brien–Fleming boundaries assuming five equally spaced analyses, and the Haybittle–Peto boundary in which a difference of three standard errors is sought at all interim analyses, and then an unadjusted *P*-value adopted at the end of trial.

provides a handicap to early stopping: explicit comparison with boundaries obtained by classical sequential methods is made in Figure 6.6 and the qualitative similarity is clear, while a quantitative investigation is made in Section 6.6.5. Other comparisons with frequentist procedures have been carried out by Freedman and Spiegelhalter (1989), DerSimonian (1996) and Freedman *et al.* (1994).

It is also possible to use 'robust priors' (Section 5.6) in which the set of prior distributions leading to a specific conclusion are identified at each interim analysis (Greenhouse and Wasserman, 1995; Carlin and Sargent, 1996). In addition, posterior probabilities of two responses can be monitored jointly and stopping considered when an event of interest, such as either outcome occurring (Etzioni and Pepe, 1994), exceeds a certain threshold. This monitoring scheme has also been proposed for single arm studies and for phase I and II trials (Section 6.12).

Although monitoring using posterior distributions appears intuitive, criticisms of this procedure include its lack of explicit loss function (Section 6.6.4), its sampling properties, and its dependence on the prior (Section 6.6.5).

Example 6.6 *CHART (continued): Monitoring trials using sceptical and enthusiastic priors*

Reference: Parmar *et al*. (1994, 2001) and Spiegelhalter *et al*. (1994). This example has previously been considered in Examples 5.1, 5.3 and 6.1.

Evidence from study: For the lung cancer trial, the data reported at each of the annual meetings of the independent DMC is shown in Table 6.3: the final row is that of the published analysis. Recruitment stopped in early 1995 after 563 patients had entered the trial. It is clear that the extremely beneficial early results were not retained as the data accumulated, although a clinically important and statistically significant difference was eventually found. Perhaps notable is that the DMC recommended continuation of the trial even when the two-sided *P*-value was 0.001, *i.e.* when the data had crossed the Haybittle–Peto boundary.

Table 6.3 Summary data reported at each meeting of the CHART lung trial DMC. Under a proportional hazards assumption with hazard ratio HR, the 2-year survival improvement, s, over a baseline of 15%, obeys $\text{HR} = \log(0.15 + s) / \log(0.15)$, which can be rearranged to $s = 0.15^{\text{HR}} - 0.15$.

Date	No. patients	No. deaths	Hazard ratio		2-year % survival improvement		Two-sided *P*-value
			Estimate	(95% CI)	Estimate	(95% CI)	
1992	256	78	0.55	(0.35 to 0.86)	20	(5 to 36)	0.007
1993	380	192	0.63	(0.47 to 0.83)	15	(6 to 26)	0.001
1994	460	275	0.70	(0.55 to 0.90)	12	(4 to 20)	0.003
1995	563	379	0.75	(0.61 to 0.93)	9	(3 to 16)	0.004
1996	563	444	0.76	(0.63 to 0.90)	9	(3 to 15)	0.003

Table 6.4 Summary data reported at each meeting of the CHART head-and-neck trial DMC. Two-year survival improvements are based on a baseline of 45% disease-free survival.

Date	No. patients	No. events	Hazard ratio		2-year % survival improvement		Two-sided *P*-value
			Estimate	(95% CI)	Estimate	(95% CI)	
1992	531	188	0.91	(0.68, 1.21)	3	(−7, 11)	0.50
1993	674	293	0.92	(0.73, 1.16)	3	(−5, 11)	0.16
1994	791	387	0.89	(0.72, 1.09)	4	(−3, 11)	0.20
1995	918	464	0.92	(0.76, 1.11)	3	(−4, 10)	0.33
1996	918	485	0.95	(0.79, 1.14)	2	(−5, 8)	0.52

For the head-and-neck cancer trial, the data reported at each meeting of the independent DMC are shown in Table 6.4. There was no strong evidence of benefit shown at any point in the study.

Bayesian interpretation: For the lung trial, the DMC was presented with survival curves, and posterior distributions and tail areas arising from a reference prior (uniform on a log(HR) scale). In view of the positive findings, the posterior distribution resulting from the sceptical prior derived in Example 5.3 was presented, in order to check whether the evidence was sufficient to persuade a reasonable sceptic.

Figure 6.7 shows the sceptical prior distributions at the start of the lung cancer trial, and the likelihood (essentially the posterior under the reference prior) and posterior for the results available in subsequent years. Under the reference prior there is substantial reduction in the estimated effect as the extreme early results are attenuated, while the sceptical results are remarkably stable and the initial estimate in 1992 is essentially unchanged as the trial progresses. The detailed results under the sceptical prior are shown in Table 6.5. Before the trial the clinicians were demanding a 13.5% improvement before changing treatment: however, the inconvenience and toxicity were found to be substantially less than expected and so probabilities of improvement are shown for 0% and 7%, around half the initial demands. Such 'shifting of the goalposts' is entirely reasonable provided it is not based on the primary outcome results.

Table 6.5 Estimates presented to CHART DMC in successive years (apart from 1996, which are the final published data) for lung cancer trial, obtained under a sceptical prior distribution. Posterior probabilities are presented for 'no improvement from CHART' (analogous to one-sided *P*-values), and for 'practically significant improvement from CHART'.

Date	No deaths	Estimated hazard ratio (HR)	2-year % survival improvement (95% CI)		*P* (imp. $< 0\%$) *i.e.* HR > 0	*P* (imp. $> 7\%$) *i.e.* HR < 0.80
1992	78	0.79	7	(−1 to 17)	0.048	0.56
1993	175	0.73	10	(3 to 18)	0.006	0.73
1994	275	0.78	8	(2 to 15)	0.009	0.60
1995	379	0.80	7	(1 to 13)	0.010	0.48
1996	444	0.81	7	(2 to 12)	0.003	0.52

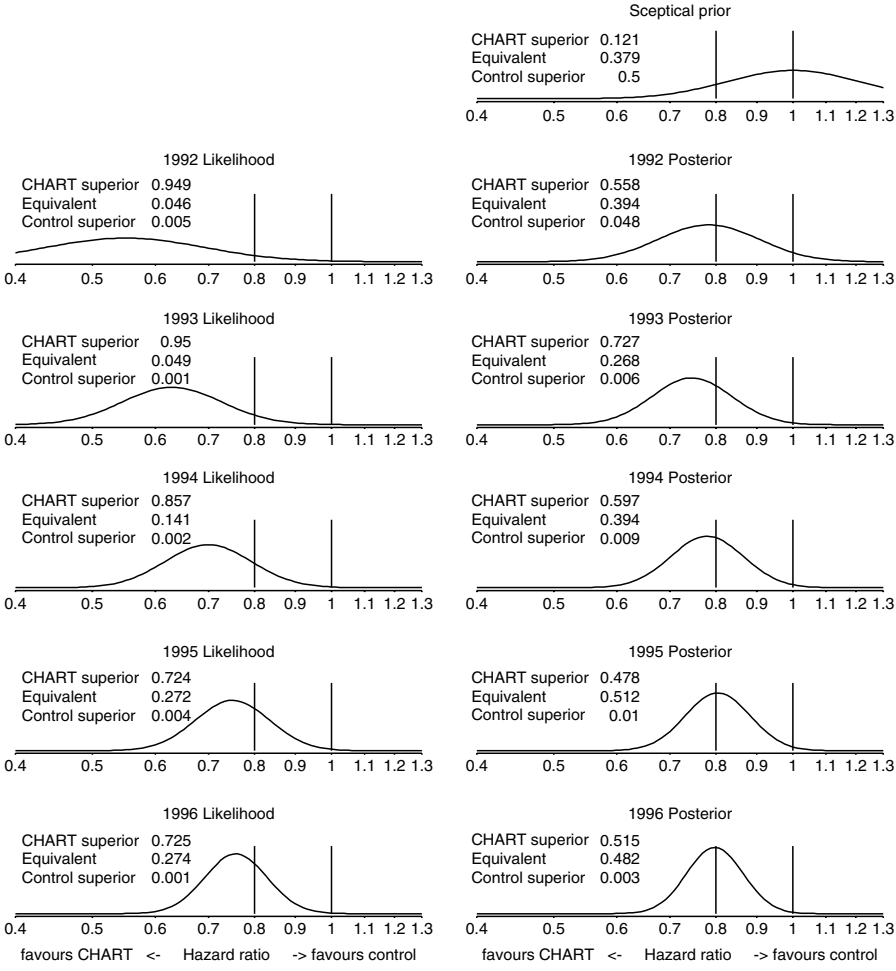

Figure 6.7 Prior, likelihood and posterior distributions for the CHART lung cancer trial assuming a sceptical prior. The likelihood becomes gradually less extreme, providing a very stable posterior estimate of the treatment effect when adopting a sceptical prior centred on a hazard ratio of 1. Demands are based on a 7% improvement from 15% to 22% 2-year survival, representing a hazard ratio of 0.80.

The sceptical posterior distribution is centred around these clinical demands, showing that these data should persuade even a sceptic that CHART both improves survival and, on balance, is the pragmatic treatment of choice.

Since the results for the head-and-neck trial were essentially negative, it is appropriate to monitor the trial assuming a enthusiastic prior in order to see if it is sufficiently convincing even to optimists. The results are shown in

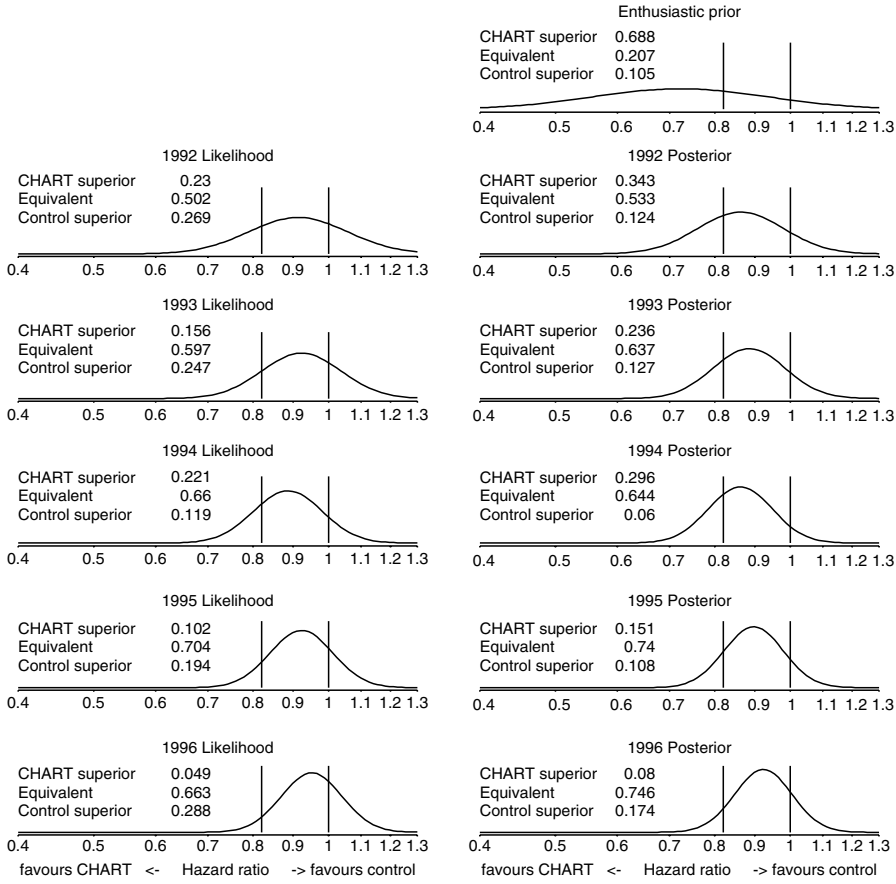

Figure 6.8 Prior, likelihood and posterior distributions for the CHART head-and-neck cancer trial assuming an enthusiastic prior, and clinical demands of a 7% improvement from 45% to 52% 2-year survival, equivalent to a hazard ratio of 0.82.

Figure 6.8, using the clinical prior derived in Example 5.1. The initial clinical demands were a 13% improvement in survival from 45% to 58%, but in parallel with the lung trial we have reduced this to a 7% improvement. The results remain equivocal, and should be sufficient to convince a reasonable enthusiast that, on the basis of the trial evidence, CHART is not of clinical benefit in head-and-neck cancer.

Sensitivity analysis: The three priors provide the sensitivity analysis.

Comments: There are two important features of the prospective Bayesian analysis of the CHART trial. First, while classical stopping rules may well have led the DMC to stop the lung trial earlier, perhaps in 1993 when the two-sided *P*-value was 0.001, this would have overestimated the benefit.

The DMC allowed the trial to continue, and consequently produced a strong result that should be convincing to a wide range of opinions. Second, after discovering that the secondary aspects of the new treatment were less unfavourable than expected, the DMC is allowed to 'shift the goalposts' and not remain with unnecessarily strong clinical demands.

6.6.3 Monitoring using predictions: 'interim power'

Investigators and funders are often concerned with the question – given the data so far, what is the chance of getting a 'significant' result? This is closely related to the concept of 'futility', and the traditional approach to this question is 'stochastic curtailment' (Halperin *et al.*, 1982) which calculates the conditional power of the study, given the data so far, for a range of alternative hypotheses: this might also be termed 'interim power'.

The following formulae assume we are interested in predicting whether future data will result in a posterior probability, or a one-sided P-value, for the null hypothesis H_0: $\theta < 0$, being less than ϵ, *i.e.* either the event S_ϵ^B or S_ϵ^C. One can make the appropriate adjustments for H_0: $\theta > 0$ and non-zero thresholds using the methods described in Section 6.5.4.

'Hybrid' predictions: using a prior and current data to predict a future classical analysis. It is straightforward to calculate predictive probabilities of eventual classical conclusions if we assume a normal likelihood. Suppose we have observed a parameter estimate y_m based on our current sample size m, and are considering a further n observations which will yield a parameter estimate Y_n. Then, since

$$\frac{my_m + nY_n}{m+n} \sim \mathrm{N}\left[\theta,\ \frac{\sigma^2}{m+n}\right],$$

after these observations we shall have a classically 'significant' result S_ϵ^C provided that

$$Y_n > \frac{-\sqrt{m+n}\ z_\epsilon\ \sigma - my_m}{n}. \tag{6.13}$$

Since $Y_n \sim \mathrm{N}[\theta,\ \sigma^2/n]$, the probability of this occurring, as a function of the observed data and unknown θ, is

$$P(S_\epsilon^C | y_m, \theta) = \Phi\left[\frac{\sqrt{n}\ \theta}{\sigma} + \frac{m\ y_m}{\sigma\sqrt{n}} + \sqrt{\frac{m+n}{n}} z_\epsilon\right]; \tag{6.14}$$

we note that this is exactly the form of the pre-trial Bayesian power curve (6.6) but replacing the 'imaginary' prior data with the observed real data. Equation (6.14) is known as the 'conditional power curve' and forms the basis for a stochastic curtailment procedure, in which this curve may be plotted and its value examined at the null, alternative and other values of θ.

It does not, however, seem reasonable to condition on a hypothesis that is no longer tenable (Spiegelhalter *et al.*, 1986; Dignam *et al.*, 1998). From a Bayesian perspective it is natural to average such conditional powers with respect to the current posterior distribution, just as the pre-trial power was averaged with respect to the prior to produce the average or expected power (Section 6.5). By again using the predictive distribution (3.24) of Y_n we can calculate the probability of S_ϵ^C to be

$$p(S_\epsilon^C|y_m,\ \text{prior}) = \Phi\left(\sqrt{\frac{n_0 n}{(n_0+m)(n_0+m+n)}}\frac{\sqrt{n_0}\mu}{\sigma} + \sqrt{\frac{m(n_0+m+n)}{n(n_0+m)}}\frac{\sqrt{m}y_m}{\sigma} + \sqrt{\frac{(m+n)(n_0+m)}{n(n_0+m+n)}}\ z_\epsilon\right). \tag{6.15}$$

We note that if $m = 0$ there are no current data and (6.15) can be shown to reduce to the pre-trial average classical power given by (6.4).

Bayesian predictions: using a prior and current data to predict a future Bayesian analysis. In a fully Bayesian analysis the posterior distribution will eventually be

$$\theta|y_m,\ Y_n \sim \mathrm{N}\left[\frac{n_0\mu + my_m + nY_n}{n_0+m+n},\ \frac{\sigma^2}{n_0+m+n}\right].$$

Having observed Y_n, we shall assume that we are interested in a 'significant' result S_ϵ^B which we have defined as the event $p(\theta < 0|y_m,\ Y_n) < \epsilon$, *i.e.* the tail area of the posterior is less than ϵ. This result will occur if

$$Y_n > \frac{-\sqrt{n_0+m+n}\ z_\epsilon\ \sigma - (n_0\mu + my_m)}{n}. \tag{6.16}$$

Since $Y_n \sim \mathrm{N}[\theta,\ \sigma^2/n]$, the probability of this event occurring, as a function of the observed data and unknown θ, is

$$P(S_\epsilon^B|y_m,\ \theta) = \Phi\left[\frac{\sqrt{n}\theta}{\sigma} + \frac{my_m}{\sigma\sqrt{n}} + \frac{n_0\mu}{\sigma\sqrt{n}} + \sqrt{\frac{n_0+m+n}{n}}\ z_\epsilon\right]. \tag{6.17}$$

Equation (6.17) can be thought of as a general form of all the other conditional power curves we have previously derived: if $n_0 = 0$ we have no prior input and we obtain the classical conditional power curve in (6.14); if $m = 0$ we obtain the Bayesian power curve in (6.6); while if $n_0 = 0$, $m = 0$ we obtain the standard power curve in (6.3).

Expression (3.24) gives the predictive distribution of Y_n, and from this we can calculate the unconditional probability of S_ϵ^B to be

$$p(S_\epsilon^B|y_m,\ \text{prior}) = \Phi\left[\frac{\sqrt{n_0+m+n}}{\sqrt{(n_0+m)n}}\frac{(n_0\mu + my_m)}{\sigma} + \sqrt{\frac{n_0+m}{n}}\, z_\epsilon\right]. \qquad (6.18)$$

Classical predictions: using only current data to predict a future classical analysis. If we wish to ignore prior opinion both in the prediction and in the reporting then we can set $n_0 = 0$ in either (6.15) or (6.18) and obtain a predictive probability of a significant result as

$$p(S_\epsilon^C|y_m) = \Phi\left[\frac{\sqrt{m+n}}{\sqrt{n}}\frac{\sqrt{m}y_m}{\sigma} + \sqrt{\frac{m}{n}}z_\epsilon\right]. \qquad (6.19)$$

This can be expressed solely in terms of the current standardised test statistic $z = \sqrt{m}y_m/\sigma$ and the fraction $f = m/(m+n)$ of the trial so far completed, to give the probability that the future tail area below 0 is less than ϵ as

$$p(S_\epsilon^C|y_m) = \Phi\left[\frac{z + \sqrt{f}z_\epsilon}{\sqrt{1-f}}\right]. \qquad (6.20)$$

Values of this quantity are plotted in Figure 6.9, which reveals that predicted probabilities of success are often surprisingly low.

The technique has been used with results that currently show approximate equivalence between treatments to justify the 'futility' of continuing a trial (Ware *et al.*, 1985), and may be particularly useful for DMCs and funders when accrual or event rates are lower than expected (Korn and Simon, 1996; Abrams, 1998). Example 6.7 provides a practical illustration of its use by a DMC. The method does not, strictly speaking, require a Bayesian justification, since the predictions can be based on a 'pivotal quantity' that does not depend on the parameter (Armitage, 1989): the '*B*-value' of Lan and Wittes (1988) enables calculation of the predictive probability of significance. Frei *et al.* (1987) and Hilsenbeck (1988) provide practical examples of stopping studies due to the futility of continuing; see Section 6.13 for further references.

In spite of the attraction of making such predictions at interim analyses, we follow Armitage (1991b) in warning against using this predictive procedure as any kind of formal stopping rule. It gives an undue weight to 'significance', and makes strong assumptions about the direct comparability of future data with

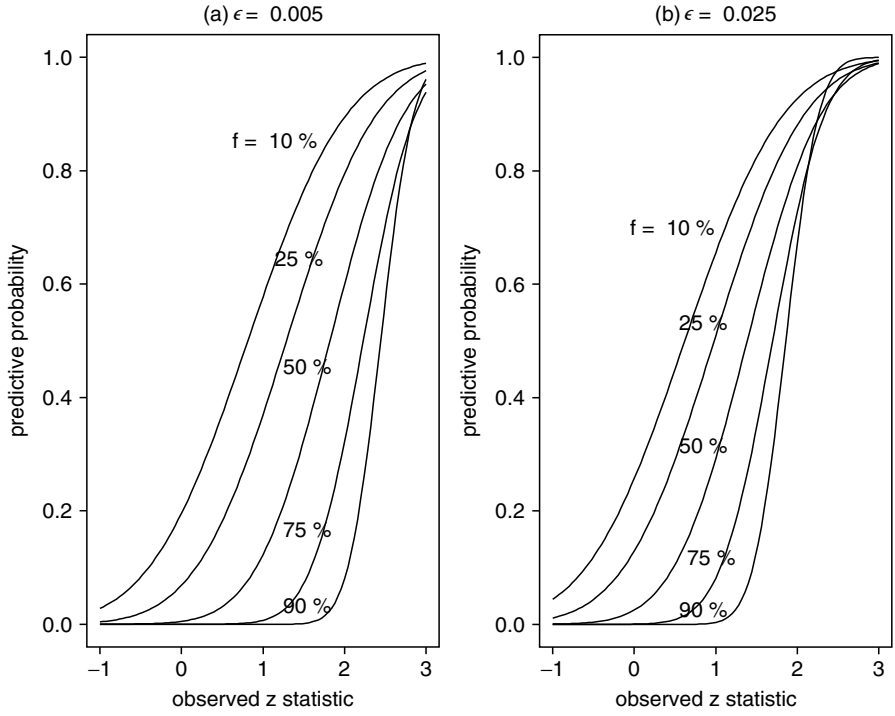

Figure 6.9 Predictive probability $\Phi[(z + \sqrt{f}z_\epsilon)/\sqrt{1-f}\,]$ of obtaining a classically significant result (two-sided $P = 0.01$ or 0.05, *i.e.* $\epsilon = 0.005$ or 0.025), given a fraction f of the study completed ($f = 10\%$, 25%, 50%, 75% and 90%) and current standardised test statistic z. For example, if one is half-way through a study ($f = 50\%$), and the treatment effect is currently one standard error away from 0 ($z = 1$), then based on this information alone there is only a 29% chance that the trial will eventually show a significant (two-sided $P = 0.05$) benefit of treatment.

those data already observed – for example, if future data involve extended follow-up there may be undue reliance on an assumption of proportional hazards.

Example 6.7 *B-14: Using predictions to monitor a trial*

Reference: Dignam *et al*. (1998).

Intervention: Long-term tamoxifen therapy for prevention of recurrence of breast cancer.

Aim of study: To estimate disease-free survival benefit from tamoxifen over placebo, in patients who already have had 5 years of taking tamoxifen without a recurrence.

Study design: Sequential randomised controlled study (National Surgical Adjuvant Breast and Bowel Project (NSABP) B-14) using O'Brien–Fleming stopping boundaries. Interim analyses were planned at intervals of approximately 1–1.5 years beginning in the fourth year of the study.

Outcome measure: Disease-free survival.

Planned sample size: To detect a 40% reduction in annual risk associated with tamoxifen (hazard ratio $= 0.6$), with 85% power and a one-sided tail area of 5%, 115 events were required. It had been planned that 624 patients were to be randomised, but eventually 1172 were recruited due to a lower than expected event rate.

Statistical model: Proportional hazards regression model, with summary using the approximate hazard ratio analysis. Following Section 2.4.2, if there are O_T events on treatment, and O_C events on control, then $2(O_T - O_C)/m$ is an approximate estimate of the log(hazard ratio) θ, with mean θ and variance $4/m$.

Prospective Bayesian analysis?: No, the DMC used conditional power and current data in order to make decisions.

Prior distribution: An 'enthusiastic' (or optimistic) prior was centred on a 40% hazard reduction and a 5% chance of a negative effect, *i.e.* HR > 1, equivalent on the log(HR) scale to a normal prior with mean -0.51 and standard deviation 0.31 ($\sigma = 2$, $n_0 = 41.4$). Also a sceptical prior was adopted with the same standard deviation as the enthusiastic prior but centred on 0, thus displaying a 5% chance of the true difference exceeding the alternative hypothesis of 40% hazard reduction.

Loss function or demands: No explicit loss function or range of equivalence.

Computation/software: Conjugate normal analysis.

Evidence from study: The DMC was presented with the data in Table 6.6. Unexpectedly, the results favoured the control treatment. At the third analysis in June 1995, there was a nominal two-sided $P = 0.01$ using the full survival data; this was not sufficient to cross the O'Brien–Fleming stopping boundary which demands two-sided $P < 0.003\,46$. Eighty-eight of the planned 115 events had been observed, and the DMC calculated that even if all 27 remaining events occurred in the control arm, the final results would still not 'significantly' favour tamoxifen. The DMC also considered the conditional power if the trial was extended until 229 events were observed – this was less than 50% for HR $= 0.5$ in favour of tamoxifen, and 15% for HR $= 0.6$. Since these hazard ratios were implausible in the light of the current data, the DMC recommended stopping the trial since the data favoured the control treatment and there

Table 6.6 Summary data from B-14 trial, with hazard ratios and *P*-values estimated using approximate normal analysis based only on the total number of events.

Date	No. events (O_C) on placebo	No. events (O_T) on tamoxifen	Estimated log(HR) (SD)	Estimated hazard ratio (95% CI)	Two-sided *P*-value
Sept. 1993	18	28	0.435 (0.295)	1.54 (0.87 to 2.75)	0.140
Sept. 1994	24	43	0.567 (0.244)	1.76 (1.09 to 2.85)	0.020
June 1995	32	56	0.545 (0.213)	1.72 (1.14 to 2.62)	0.010
Dec. 1995	36	66	0.588 (0.198)	1.80 (1.22 to 2.65)	0.003
Dec. 1996	50	85	0.519 (0.172)	1.68 (1.20 to 2.35)	0.003

was negligible chance of the conclusions being reversed. Further events were subsequently observed and are shown in Table 6.6.

Bayesian interpretation: Figure 6.10 shows the consequences of assuming the sceptical and enthusiastic (optimistic) priors considered by Dignam

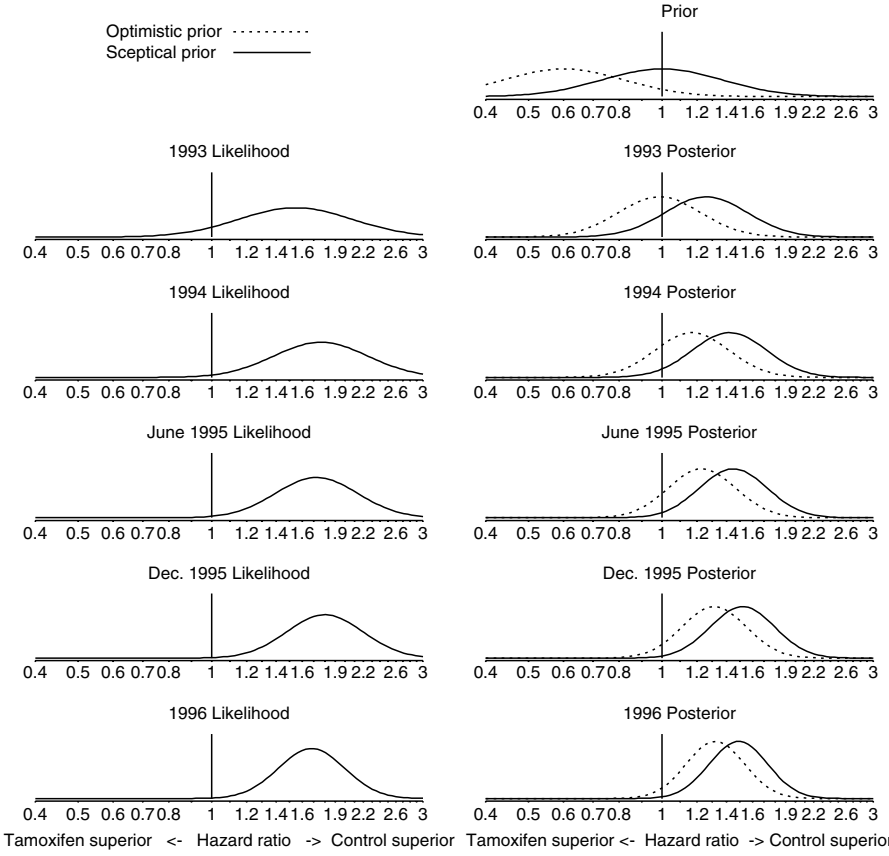

Figure 6.10 Sceptical and 'optimistic' prior distributions, likelihoods and posterior distributions at meetings of the DMC for the B-14 trial. The strong likelihood brings sceptics and enthusiasts into agreement.

et al. (1998). At the first interim analysis the evidence against tamoxifen is sufficient to bring an 'optimist' into a situation of equipoise, with a posterior mean of almost exactly 0. It is clear that by the end of the trial the likelihood is sufficiently in favour of control to bring the two extremes of opinion substantially into agreement.

We may use the results in Section 6.6.3 to calculate the predictive probability of the consequences of continuing the trial up to 115 events, based on the data observed at each of the five interim analyses. We first consider the situation after the first interim analysis in 1993 when 46 events had been observed. Three prior assumptions are examined: a reference analysis (essentially a classical analysis with no adjustment for repeated looks at the data), and sceptical and 'optimistic' analyses using the priors derived above. Each column in Figure 6.11 is headed by the posterior distribution under each assumption, and below are shown the conditional probability of obtaining different conclusions at the planned end of the trial, *i.e*. after a further $115 - 46 = 69$ events have occurred. The conclusions are: 'tamoxifen superior', defined as a 95% posterior interval for the hazard ratio lying wholly below 1; 'equivocal', defined as a 95% posterior interval including 1; and 'control superior', defined as a 95% posterior interval lying wholly above 1. Conditional on each value of $\theta = \log(\text{HR})$, the probabilities of these outcomes can be obtained from (6.17) by substituting the appropriate values for the prior distribution.

Under the reference analyses, the chance of concluding in favour of control is fairly substantial for true hazard ratios greater than 1.5, and such values are supported by the current posterior distribution. The chance of finding in favour of tamoxifen is negligible unless the true hazard ratio is as low as 0.4, which is essentially ruled out by the reference posterior. Integrating the power curves with respect to the reference posterior provides the expected powers shown in the first column of Table 6.7. These probabilities can be obtained as follows. The current z statistic in favour of control is $0.435/0.295 = 1.475$, the fraction of the trial completed is $f = 46/115 = 0.4$, and $\epsilon = 0.025$. From Figure 6.9 we can read off that the expected power is approximately 0.6, and substituting in (6.20) gives the exact value of 0.619. For the expected power to find in favour of tamoxifen, we can take one minus the expected power for control when $\epsilon = 0.975$, which is 0. The unconditional probability of finishing with an equivocal result is simply one minus the other expected powers.

The sceptical analysis has a greater tendency to find an equivocal result as the sceptical prior will be included in the final analysis, and this is reflected in both the conditional power curves and the expected powers

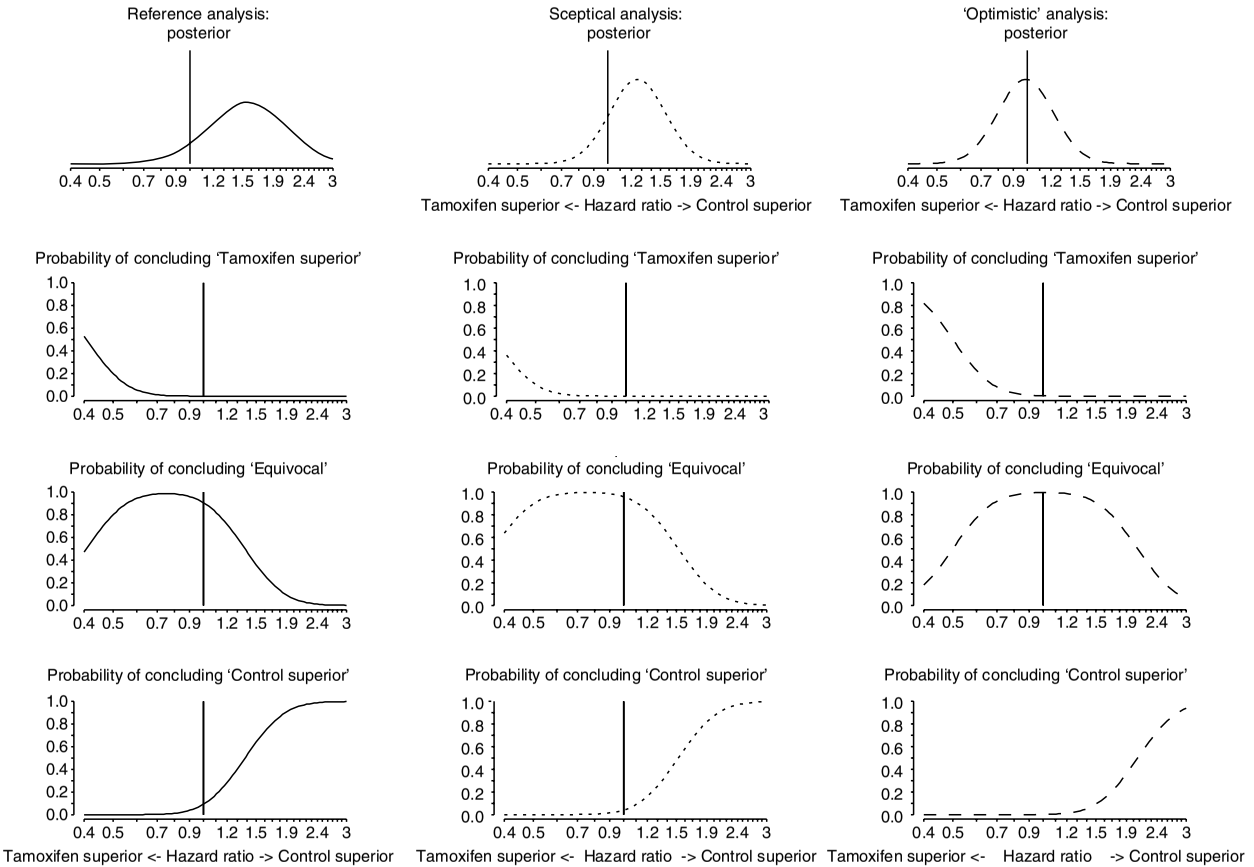

Figure 6.11 Alternative predictions that could be made at the first interim analysis in 1993. The reference analysis uses the data alone, and the power curves are the standard conditional power curves for each of the three possible conclusions after a further $115 - 46 = 69$ events are observed. The sceptical and optimistic analyses show the conditional power for each possible conclusion assuming the prior is to be used in the analysis.

Table 6.7 Probabilities of eventual conclusions for the B-14 trial after the first interim analysis in 1993. Three different prior assumptions are considered, first with the prior to be used in the analysis as well as the predictions, and then with the prior not being used in the final analysis.

Final conclusion	Reference	When using prior in analysis		When *not* using prior in analysis	
		Sceptical	'Optimistic'	Sceptical	'Optimistic'
'Tamoxifen superior'	0.000	0.000	0.017	0.000	0.003
'Equivocal'	0.380	0.724	0.972	0.610	0.846
'Control superior'	0.619	0.276	0.011	0.390	0.151

shown in Table 6.7. The optimistic analysis is even more reluctant to draw a firm conclusion given its current balanced opinion, and firmly (and wrongly, with hindsight) predicts an equivocal result at the end of the trial.

In practice it is likely that the final analysis of the trial would be classical, and therefore it is of interest to carry out a 'hybrid' or mixed prediction in which the prior is used for prediction but not for analysis. This essentially means that the classical conditional power curves shown in the first column of Figure 6.11 are averaged with respect to the sceptical or optimistic posterior distributions. The results are shown in the last two columns of Table 6.7. The chance of finding a result in favour of control is strengthened.

The consequences of making mixed predictions at each interim analysis are shown in Figure 6.12; only the chances of obtaining a conclusion in favour of control are shown, as the chance of finding in favour of tamoxifen is less than 0.003 in all cases.

Sensitivity analysis: Dignam *et al*. (1998) considered a range of prior distributions with means varying between optimistic and sceptical – we have just illustrated the extremes of this range.

Comments: A predictive calculation suggests that continued follow-up would almost certainly not lead to evidence of benefit for tamoxifen. However, when the DMC recommended stopping at the third interim analysis, Figure 6.10 shows that an optimist could still have 13% belief in a benefit from tamoxifen, and therefore would not rule out further trials. Dignam *et al*. (1998) defend the decision to stop and state that 'even an advocate of continued testing of the question might argue that we should have closed and reported the B-14 study, if for no other reason than to make way for a confirmatory trial in which participants could be adequately consented'.

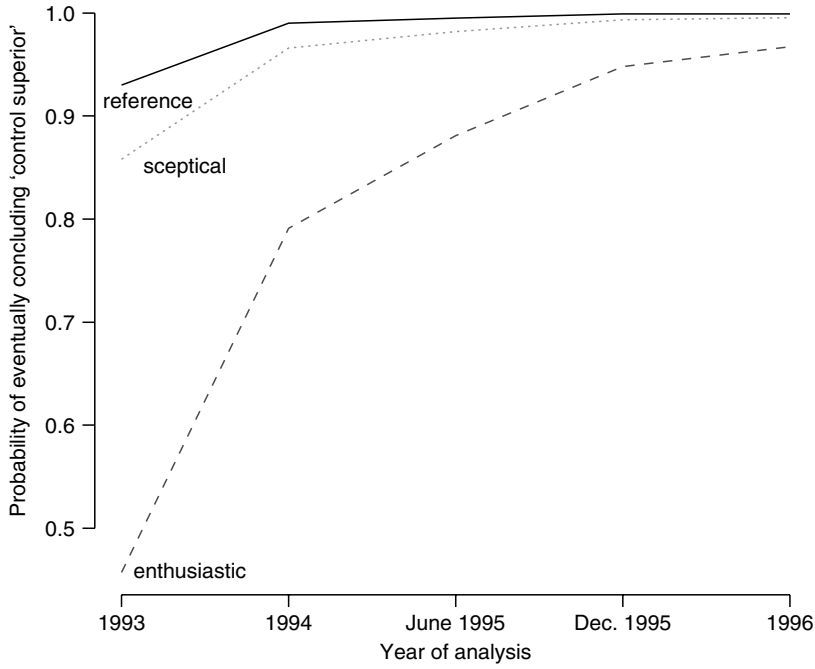

Figure 6.12 Predictive probability of reaching the conclusion 'control superior' at the end of the trial, under different prior assumptions but assuming a classical analysis. The predictive probability of a 'significant' result in favour of tamoxifen is negligibly small and is not shown. At the third interim analysis (June 1995), even an enthusiast would admit only a 16% chance of eventually drawing any conclusion except that control was superior.

6.6.4 Monitoring using a formal loss function

The full Bayesian decision-theoretic approach requires the specification of losses associated with all combinations of possible true underlying states and all possible actions. The decision whether to terminate a trial is then, in theory, based on whether termination has a lower expected loss than continuing, where the expectation is with respect to the current posterior distribution, and the consequences of continuing have to consider all possible future actions. This 'backwards induction' requires the computationally intensive technique of 'dynamic programming' and typically makes practical implementation troublesome. There is also an extensive theoretical literature on sequential trials designed from a non-Bayesian decision-theoretic perspective (Bather, 1985).

However, reasonably straightforward solutions can be found in some somewhat idealised circumstances. For example, Anscombe (1963) considers n pairs of patients randomised equally to two groups, a total patient horizon of N, a uniform prior on true treatment benefit, and a loss function proportional to the

number of patients given the inferior treatment times the size of the inferiority. He concludes it is approximately optimal to stop and give the 'best to the rest' when the standard one-sided P-value is less than n/N – half the proportion of patients already randomised.

Berry and Pearson (1985) and others have extended such theory to allow for unequal stages and so on, while Carlin *et al.* (1998) claim backwards induction is computationally feasible using Markov chain Monte Carlo methods, in which forward sampling is used as an approximation to the optimal strategy.

As an illustrative (but retrospective) example, Berry *et al.* (1994) consider a trial of influenza vaccine for Navajo children. They construct a theoretical model consisting of priors for the effectiveness of the vaccine and the placebo treatment, the probability of obtaining regulatory approval and the time taken to obtain it, and the probability of a superior vaccine appearing in the next 20 years and the time taken for it to appear. After each month the expected number of cases of the strain amongst Navajo children in the next 20 years is calculated in the case of stopping the trial and of continuing the trial (the latter being calculated by dynamic programming). The trial is stopped when the former exceeds the latter.

As already discussed in Section 6.2, the level of detail required for such an analysis has been criticised as being unrealistic (Breslow, 1990), but it has been argued that trade-offs between benefits for patients within and outside the trial should be explicitly confronted (Etzioni and Kadane, 1995) and decision theory used to decide whether a trial is worth embarking on in the first place (Section 9.10).

6.6.5 Frequentist properties of sequential Bayesian methods

Although the long-run sampling behaviour of sequential Bayesian procedures is irrelevant from the strict Bayesian perspective, a number of investigations have taken place which generally show good sampling properties (Rosner and Berry, 1995). In particular, Grossman *et al.* (1994) explore the sampling properties of the boundaries described in (6.11) arising from assuming a sceptical prior (Section 5.5) centred on zero and with 'sample size' n_0, and a planned maximum experimental sample size n. They estimate by simulation and interpolation the values for the 'handicap' n_0/n that would give rise to an overall Type I error of 5% and 1% for different numbers of equally spaced interim analyses. The results in Table 6.8 show the required handicap is fairly stable over a range of designs: in particular, the boundaries displayed in Figure 6.6, based on an 'imaginary' prior trial of around 26% of the planned sample size, will have Type I error around 5% for five interim analyses. Grossman *et al.* (1994) also show this boundary has good power and expected sample size. Thus an 'off-the-shelf' Bayesian procedure assuming a sceptical prior essentially mirrors the conservative behaviour of the Neyman–Pearson approach. The sampling properties of Bayesian designs has been particularly investigated in the context of phase II trials (Section 6.12).

Table 6.8 Handicaps to fix Type I error rate when monitoring using a sceptical prior for different number of analyses: the handicap is n_0/n, the ratio of the prior 'sample size' to the maximum intended sample size.

Number of analyses	'Handicap' for two-sided $\alpha = 0.05$	'Handicap' for two-sided $\alpha = 0.01$
1	0	0
2	0.16	0.11
3	0.22	0.15
4	0.25	0.17
5	0.27	0.18
6	0.29	0.20
7	0.30	0.21
8	0.32	0.22
9	0.33	0.22
10	0.33	0.23

One contentious issue is 'sampling to a foregone conclusion' (Armitage *et al.* 1969). This mathematical result proves that repeated calculation of posterior tail areas will, even if the null hypothesis is true, eventually lead a Bayesian procedure to reject that null hypothesis. This does not, at first, seem an attractive frequentist property of a Bayesian procedure. Nevertheless, Cornfield (1966) argued that 'if one is seriously concerned about the probability that a stopping rule will certainly result in the rejection of a null hypothesis, it must be because some possibility of the truth of the hypothesis is being entertained', and if this is the case then one should be placing a lump of probability on it, as discussed in Section 5.5, and so fit within the Bayesian hypothesis-testing framework (Section 3.3). He shows that if such a lump, however small, is assumed then the problem disappears in the sense that the probability of rejecting a true null hypothesis does not tend to one. Armitage (1990) is not persuaded, claiming that even with a continuous prior distribution with no lump at the null hypothesis, one might still be interested in Type I error rates at the null as giving a bound to those at non-null values.

A somewhat more subtle objection, well described by Rosenbaum and Rubin (1984), is that the properties of a Bayesian stopping rule based on posterior tail areas may be over-dependent on the precise prior distribution (Jennison, 1990). A possible response is that Bayesian stopping should not be based on a strict rule derived from a single prior, and instead a variety of reasonable perspectives investigated and a trial stopped only if there is broad convergence of opinion.

6.6.6 Bayesian methods and data monitoring committees

A DMC is charged with both safeguarding the patients involved in a trial, and ensuring the quality of a trial's conduct and conclusions. The principles and

practice of DMCs are fully discussed in Ellenberg *et al.* (2002), and here we restrict ourselves to the possible impact of Bayesian methods on a DMC's deliberations. Perhaps the most relevant elements are the ability to use external evidence as a basis for prior opinion in any analysis, and the formalisation through sceptical and enthusiastic priors of the wide range of clinical opinion that it may be necessary to convince before a trial's results have the appropriate impact. As outlined in Section 6.6.4, a full decision-theoretic approach would be attractive but difficult to put into practice in a convincing manner, although Kadane *et al.* (1998) report an intention to elicit prior distributions and utilities from members of the DMC for a large collaborative cancer trials group (NSABP), and use the forward sampling approach to solve the dynamic programming problem. Their success in this ambitious venture remains to be seen.

At an interim analysis of trial data, a DMC may be faced with a variety of possible recommendations that it can make concerning the future conduct of the trial. Using the structure of Altman *et al.* (2004), these may include the following:

- *The study should stop completely*. We have already seen in Example 6.6 how a DMC might use Bayesian methods in order to inform a recommendation whether to stop in favour of an apparent benefit of the new intervention on a primary outcome measure, possibly through using a sceptical prior to assess the degree to which the results would be convincing to a wide range of opinion. Similarly, in Example 6.7 we saw how an enthusiastic prior can be used to temper claims for apparent benefit in the control group. The DMC might also recommend stopping because of safety concerns on secondary outcomes, although these may not be so amenable to formal stopping procedures. A recommendation to stop could also be influenced by a 'futility' argument which assesses the chance of ever reaching a particular conclusion were the trial to continue, and this naturally falls into the framework outlined in Section 6.6.3. Finally, there may be convincing evidence of equivalence or non-inferiority: while a frequentist framework requires prespecification of this as an objective of the trial with pre-chosen limits, a Bayesian analysis allows the 'goalposts' to change as the trial progresses and hence a DMC can make such a recommendation on the basis of all currently available evidence. In all these deliberations the DMC is free to incorporate external evidence, such as recently published studies, into a prior opinion.
- *Part of the study should stop*. A recommendation could be made for randomisation to cease for a subgroup of patients or one of many arms in a multi-arm trial. Hierarchical models may be useful in these contexts: again stopping might be based on posterior tail areas to assess the extent to which available evidence would convince a wide body of clinical opinion.
- *The study should continue with modifications*. Design changes such as additional interim analysis, extending recruitment or extending follow-up time can have serious implications for frequentist designs that have pre-set

criteria for assigning statistical significance based on pre-set design characteristics. A Bayesian analysis is completely unaffected by such decisions and so a DMC is given considerably more freedom to adapt trial designs.

Of course, a DMC that adopts a Bayesian approach must do so in full recognition of any regulatory issues, and in such a context it would currently be unwise not to carry out such an analysis in parallel with a traditional analysis – see Section 9.12 for future discussion of regulatory acceptance of Bayesian analyses.

6.7 THE ROLE OF 'SCEPTICISM' IN CONFIRMATORY STUDIES

After a clinical trial has given a positive result for a new therapy, there remains the problem of whether a confirmatory study is needed. Fletcher *et al.* (1993) argue that the first trial's results might be treated with scepticism, and Berry (1996b) claims that using a sceptical prior is a means of dealing with 'regression to the mean', in which early extreme results tend to return to the average over time. Example 6.8 illustrates the potential value of this approach.

Example 6.8 *CALGB: Assessing whether to perform a confirmatory randomised clinical trial*

Reference: Parmar *et al*. (1996).

Intervention: Adjunct chemotherapy for non-small-cell lung cancer.

Aim of study: To compare adjunct chemotherapy with radiotherapy alone.

Study design: A RCT conducted by the Cancer and Leukemia Group B (CALGB) between 1984 and 1987 planned to enrol 240 patients with locally advanced stage III non-small-cell lung cancer and to observe approximately $n = 190$ deaths. From (2.38), this design has 80% power to detect at the 5% level a log(hazard ratio) of $\theta_A = (z_{0.8} - z_{0.025})\,\sigma/\sqrt{n}$ where $\sigma = 2$ (Section 2.4.2). Thus $\theta_A = 0.405$, corresponding to a hazard ratio (HR) of $\exp(-0.405) = 0.67$, where HR < 1 favours new over standard therapy.

Outcome measure: Full survival data were available, with results presented in terms of estimates of HR, the 2-year survival improvement, and the median improvement in survival in months. From Section 2.4.2, the relation between these quantities is as follows. Let the 2-year survival probability under the standard and new therapies be p_S and p_N, respectively. Then, assuming proportional hazards, $\text{HR} = \log(p_N)/\log(p_S)$. Further, let the median survival time under the standard and new therapies be s_S and s_N, respectively. If we assume an exponential survival distribution (constant hazard rate), then $\text{HR} = s_S/s_N$.

Statistical model: Proportional hazards model, providing an approximate normal likelihood for $\theta = \log(\text{HR})$ (Section 2.4.2).

Prospective analysis?: The Bayesian analysis was carried out retrospectively.

Prior distribution: A default reference (uniform on the log(HR) scale) prior was termed 'enthusiastic' by Parmar *et al*. (1996). They also derived a sceptical prior by the method described in Section 5.5.2, with mean 0 and standard deviation $\sigma/\sqrt{n_0}$. The original alternative hypothesis was $\theta_A = \log(0.67) = -0.405$, and a prior centred at zero and with 5% chance of exceeding this value would have standard deviation $0.405/1.645 = 0.246$. Using $\sigma = 2$, this is equivalent to a 'prior sample' of size $n_0 = (2/0.246)^2 = 66$. Figure 6.13 shows this sceptical prior distribution with a median HR of 1, which is equivalent to an 'imaginary' trial in which 33 patients died on each treatment.

Loss function or demands: Parmar *et al*. (1996) argue that it might be reasonable to demand an improvement equal to the alternative hypothesis of a hazard ratio of 0.67, or an additional 5 months' median survival. The sceptical prior expresses a probability of 45% that the true benefit lies in the range of equivalence.

Evidence from study: The trial stopped early after enrolling 156 patients and observing the data shown in Table 6.9. These results suggested a substantial improvement – the two-sided *P*-value adjusted for covariates was 0.0075. The results show an estimated log (hazard ratio) $y_m = -0.489$ with standard error $(-0.489 + 0.846)/1.96 = 0.183$, which from the likelihood above is equivalent to $m = (\sigma/0.183)^2 = 120$ deaths.

Computation/software: Conjugate normal analysis.

Bayesian interpretation: The likelihood plot shows the inferences to be made from the reference prior, essentially equivalent to those in Table 6.9. The probability that the new treatment is actually inferior is 0.004 (equivalent to the one-sided *P*-value 0.0075/2.) The probability of clinical superiority is 68%, which might be considered sufficient to change treatment policy. The posterior plot shows the impact of the sceptical prior, in that the chance of clinical superiority is reduced to 27% – hardly sufficient to change practice.

Comments: In fact, Parmar *et al*. (1996) report that the NCI Intergroup Trial investigators were unconvinced by the CALGB trial due to their previous negative experience, and so carried out a further confirmatory study. They found a significant median improvement but of only 2.4 months, from 11.4 to 13.8 months. Under an exponential assumption this corresponds to a hazard ratio of 0.83, suggesting the sceptical approach might have given a more reasonable estimate than the likelihood based on the CALGB trial alone.

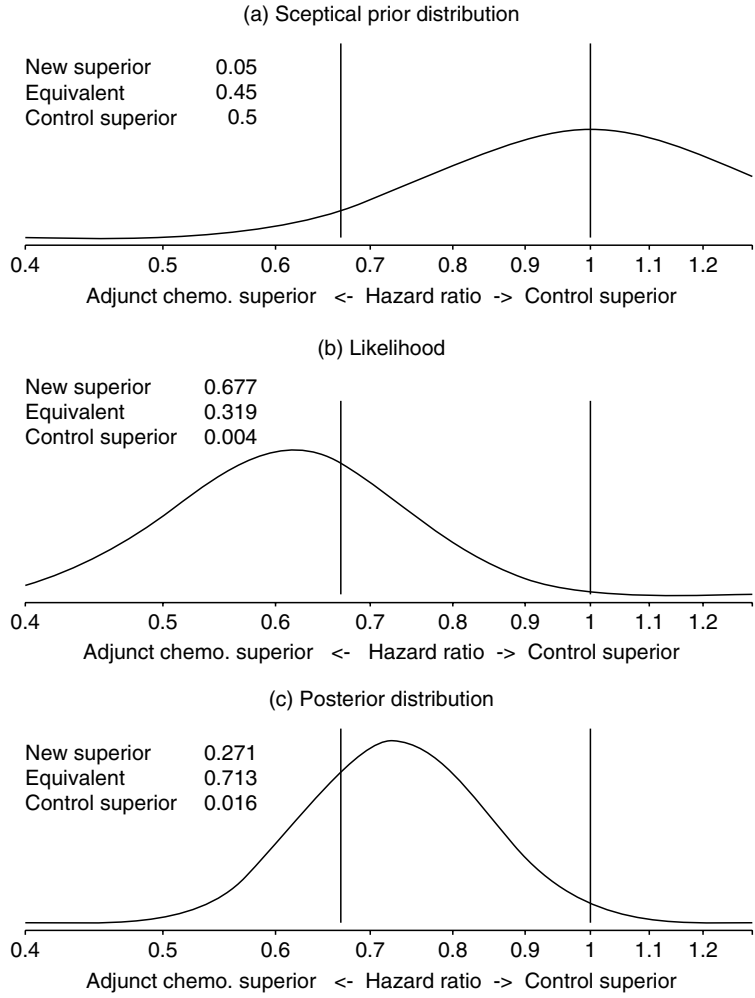

Figure 6.13 Prior, likelihood and posterior distributions arising from CALGB trial of standard radiotherapy versus additional chemotherapy in advanced lung cancer. The vertical lines give the boundaries of the range of clinical equivalence. Probabilities of lying below, within and above the range of equivalence are shown.

Table 6.9 Results of CALGB trial comparing adjunct chemotherapy with radiotherapy alone in advanced non-small-cell lung cancer.

Outcome	Estimate of improvement	95% CI
Median survival (mo)	6.3	1.4 to 13.3
2-year survival (%)	16	4 to 29
Hazard Ratio HR	0.61	0.43 to 0.88
$\theta = \log(\text{HR})$	-0.489	-0.846 to -0.131

6.8 MULTIPLICITY IN RANDOMISED TRIALS

6.8.1 Subset analysis

The discussion on multiplicity in Section 3.17 has already described how multiple simultaneous inferences may be made by assuming a common prior distribution with unknown parameters, provided an assumption of exchangeability is appropriate, *i.e.* the prior does not depend on the units' identities. Within the context of clinical trials this has immediate relevance to the issue of estimating treatment effects in subgroups of patients.

A reasonable model might be to assign a reference (uniform) prior for the overall treatment effect, and then assume the subgroup-specific deviations from that overall effect have a common prior distribution with zero mean. This prior expresses scepticism about widely differing subgroup effects, although the variability allowed by the prior is usually estimated from the data: this procedure 'leads to 1) pooling subgroups if the differences among them appear small, 2) keeping them separate if differences appear large, and 3) providing intermediate results for intermediate situations.' (Cornfield, 1976). This specification avoids the need for detailed subjective input, which may be seen as an attractive feature. Many applications consider this an empirical Bayes procedure which gives rise to traditional confidence intervals which are not given a Bayesian interpretation. Donner (1982) sets out the basic ideas, and Dixon and Simon (1991), Simon (1994b) and Simon *et al.* (1996) have elaborated the techniques in a number of examples.

6.8.2 Multi-centre analysis

Methods for subset analysis (Section 6.8.1) naturally extend to multi-centre analysis, in which the centre-by-treatment interaction is considered as a random effect drawn from some common prior distribution with unknown parameters. Explicit estimation of individual institutional effects may be carried out, which in turn relates strongly to the methods used for institutional comparisons of patient outcomes (Section 7.4).

There have been numerous examples of this procedure (Section 6.13), generally adopting Markov chain Monte Carlo techniques due to the intractability of the analyses. Recent case studies include Gould (1998) who provides WinBUGS code (Section 3.19.3), and Jones *et al.* (1998) who compare estimation methods. Senn (1997b, p. 199) discusses when a random-effects model for centre-by-treatment interaction is appropriate, emphasising the possible difficulty of interpreting the conclusions particularly in view of the somewhat arbitrary definition of 'centre'.

6.8.3 Cluster randomisation

Rather than randomising individual patients, some trials randomise *clusters* of patients, grouped (say) by their general practitioner, both for administrative

convenience and because some interventions, for example those involving education or organisation, are applied at the cluster level. A Bayesian approach to the analysis of such trials has been considered by Spiegelhalter (2001) with respect to continuous responses, and Turner *et al.* (2001) for binary responses. In each situation they assume exchangeable clusters, and discuss the appropriate choice of priors on between-cluster variances. Of particular interest is the growing body of empirical evidence on the magnitude of intra-class correlation coefficients observed in different clinical trial contexts, and its value in deriving appropriate prior distributions.

6.8.4 Multiple endpoints and treatments

Multiple endpoints in trials can often be of interest when dealing with, say, simultaneous concern with toxicity and efficacy. This tends to occur in early phase studies, and a Bayesian approach allows one to create a two-dimensional posterior distribution over toxicity and efficacy (Etzioni and Pepe, 1994; Dominici, 1998; Thall and Sung, 1998). General random-effects models for more complex situations can be constructed (Legler and Ryan, 1997). Naturally, a two-dimensional prior is required and particular care must be taken over the dependence assumptions.

A similar situation arises with many treatments: if one is willing to make exchangeability assumptions between treatment effects, then a hierarchical model can be constructed to deal with the multiple-comparison problem. This was proposed long ago by Waller and Duncan (1969). Brant *et al.* (1992) update this procedure by assuming exchangeable treatments and setting the critical values for the posterior probabilities of treatment effects by using a decision-theoretic argument based on specifying the relative losses for Type I to Type II error.

Both multiple endpoints and treatments are also common in meta-analysis of randomised controlled trials (Chapter 8).

6.9 USING HISTORICAL CONTROLS*

A Bayesian basis for the use of historical controls in clinical trials, generally in addition to some contemporaneous controls, is based on the idea that it is wasteful and inefficient to ignore all past information on control groups when making a new comparison. Pocock (1976) argued that careful use of historical controls may allow fewer controls in current studies and give more accurate effect estimates, and methods have since been developed particularly within the field of carcinogenicity studies (Ryan, 1993).

The crucial issue is the extent to which the historical information can be considered similar to contemporaneous data: Pocock (1976) suggests somewhat

stringent criteria for use of historical controls, demanding that, in comparison to contemporaneous controls, they should have the same treatment, the same eligibility, the same evaluation, the same baseline characteristics, and the same organisation and investigators, and that there should be no reason to suspect systematic differences. These issues are essentially indistinguishable from those to be taken into account when using any historical evidence, such as when basing prior opinion on past data. We can therefore place the possible approaches within the structure laid out in Sections 3.16 and 5.4, keeping in mind that here we are concerned with past evidence concerning a single (control) arm of a trial, whereas in Section 5.4 we were concerned with past data on a treatment effect. However from an analytic perspective there is little difference between these two contexts. Possible approaches include the following:

(a) *Ignore the historical control data.* This is the standard option in which each trial uses only its own control group.
(b) *Assume the historical control groups are exchangeable with the current control group*, and hence build or assume a hierarchical model for the response within each group (Tarone, 1982; Dempster *et al.*, 1983). Pocock's criteria, described above, seem a natural basis for making a subjective judgement of exchangeability, and such an assumption leads to a degree of pooling between the control groups, depending on their observed or assumed heterogeneity – a classical random-effects formulation of this approach is also possible (Thall and Simon, 1990). Gould (1991) suggests using past trials to augment current control group information, assuming exchangeable control groups. Rather than directly producing a posterior distribution on the contrast of interest, he uses this historical information to derive predictive probabilities of obtaining a significant result were a full trial to have taken place (Section 6.5); his example is treated in Example 8.4.
(c) *Assume the historical controls are a biased sample.* With only one group of historical controls, Pocock (1976) adopts the model in Section 5.4 in which one assumes an additional bias with prior mean 0 – we shall give details of this method and illustrate its use in Example 6.9. Let y_t, y_c and y_h be the observed response in the randomised treated, randomised control and historical control groups respectively, where we assume

$$y_t \sim \mathrm{N}[\theta_t, \sigma_t^2], \tag{6.21}$$

$$y_c \sim \mathrm{N}[\theta_c, \sigma_c^2], \tag{6.22}$$

$$y_h \sim \mathrm{N}[\theta_c + \delta, \sigma_h^2], \tag{6.23}$$

and the degree of bias δ in the historical control evidence is assumed to be

$$\delta \sim \mathrm{N}[0, \sigma_\delta^2]. \tag{6.24}$$

From (6.23) and (6.24) we find the marginal distribution of y_h to be

$$y_h \sim \mathrm{N}[\theta_c, \sigma_h^2 + \sigma_\delta^2]. \tag{6.25}$$

Both (6.22) and (6.25) provide evidence concerning θ_c, and a combined likelihood for θ_c is obtained by weighting the two estimates of θ_c inversely by their variances:

$$\frac{y_c + W y_h}{1 + W} \sim \mathrm{N}\left[\theta_c, \left(\frac{1}{\sigma_c^2} + \frac{1}{\sigma_h^2 + \sigma_\delta^2}\right)^{-1}\right], \tag{6.26}$$

where $W = \sigma_c^2/(\sigma_h^2 + \sigma_\delta^2)$. (6.26) can also be obtained in a somewhat convoluted way by assuming a uniform prior for θ_c, doing two Bayesian updates using the likelihoods (6.22) and (6.25), and then seeing what likelihood would have given rise to the resulting posterior.

The parameter of interest is the treatment effect $\theta = \theta_t - \theta_c$, and we can obtain a likelihood for θ from (6.21) and (6.26), giving

$$y_t - \frac{y_c + W y_h}{1 + W} \sim \mathrm{N}\left[\theta, \sigma_t^2 + \left(\frac{1}{\sigma_c^2} + \frac{1}{\sigma_h^2 + \sigma_\delta^2}\right)^{-1}\right]. \tag{6.27}$$

The likelihood (6.27) can then be combined with a prior for θ in the standard manner.

In addition to the assumptions above, values or estimates are also required for σ_e^2, σ_c^2 and σ_h^2. Finally, prior opinion regarding σ_δ^2 also has to be specified.

(d) *Discount the size of the historical control group*. This is essentially the 'power' prior described in Section 5.4, but applied solely to the control arm.

(e) *Functional dependence*. This would be relevant if, for example, the historical controls were considered entirely compatible with current controls, but needed to be adjusted for imbalance in covariates.

(f) *Assume the historical control individuals are exchangeable with those in the current control group*, which leads to a complete pooling of historical with experimental controls.

Various combinations of these assumptions are possible: Berry and Stangl (1996a) assume a parameter representing the probability that any past individual is exchangeable with current individuals, while Racine *et al.* (1986) assume a certain prior probability that the entire historical control group exactly matches the contemporaneous controls and hence can be pooled. It is also possible to use such models as a basis for designing future studies and deciding the number of patients to be allocated in each arm.

Example 6.9 *ECMO: incorporating historical controls*

Reference: Ware (1989) and the subsequent discussion.

Intervention: Extracorporeal membrane oxygenation (ECMO), an invasive technique for blood oxygenation in newborn babies.

Aim of study: Until the advent of ECMO, conventional medical therapy (CMT) for infants with severe persistent pulmonary hypertension of the newborn (PPHN) achieved less than a 20% survival rate. Early experiences with ECMO were promising, and by 1985 survival rates of over 80% were being reported. Following a review of the evidence of CMT prior to 1985, an RCT was undertaken at two hospitals at Harvard between 1986 and 1988, in order to evaluate the use of ECMO compared to CMT in this extremely poor prognosis patient population.

Study design: Adaptive two-phase RCT. Phase I randomised patients to either ECMO or CMT, while in phase II patients were to be allocated to whichever was the superior treatment in phase I. We consider here an evaluation of the effectiveness of ECMO based on the evidence from the first, randomised, phase of the trial, including information from historical control patients.

Outcome measure: Odds ratio (OR) of death (OR < 1 favours ECMO).

Planned sample size: The study was designed so that when stopped with at most four deaths in each arm, the study would have approximately 77% power to detect an odds ratio of 1/16 at the 5% significance level corresponding to mortality rates of 20% and 80% in the ECMO and CMT groups, respectively.

Statistical model: A normal likelihood based on the observed log(odds ratio) is adopted: more accurate methods would make use of the full binomial likelihood and MCMC methods (Section 3.19.2).

Prospective analysis?: No.

Prior distribution: Following the approach of Kass and Greenhouse (1989), we shall investigate the use of a sceptical prior distribution for the treatment effect, and historical evidence for survival in the control group. As prior evidence of survival under CMT, we shall follow Ware (1989) in restricting attention to cases of severe PPHN treated with CMT in the specific Harvard hospitals immediately preceding the trial: 13 patients were thus identified as 'historical controls', of whom 11 died. Table 6.10 shows the resulting estimated odds of death, log-odds of death and its variance (Section 2.4). Whilst the use of such historical data may be discounted totally or simply used at 'face-value', it may also be reasonable to discount it in some manner, such as assuming exchangeability,

Table 6.10 Historical and observed data for Harvard ECMO study showing notation for estimates and variances of log-odds of death.

Trial	ECMO deaths/ cases	CMT deaths/ cases	Odds	log(odds)	Variance of log(odds)
Historical data		11/13	4.60	1.53(y_h)	0.49 (σ_c^2)
Harvard phase I	0 / 9		0.05	−2.94(y_t)	2.11 (σ_t^2)
		4/10	0.69	−0.37(y_c)	0.38 (σ_c^2)

Table 6.11 Use of historical controls in assessing odds ratio of death for patients receiving ECMO compared to conventional treatment: OR < 1 favours ECMO. For example, a fourfold relative bias corresponds to a 95% chance that the odds ratio between historical and current control mortality lies between 0.25 and 4.

Potential relative bias assumed in historical controls	σ_δ	Posterior distribution of odds ratio			
		Mean	95% interval	P(OR<1)	P(OR<0.4)
0	0.000	0.033	0.0017 to 0.658	98.7%	94.9%
1.1	0.048	0.033	0.0017 to 0.659	98.7%	94.9%
1.5	0.207	0.035	0.0017 to 0.686	98.6%	94.6%
2	0.354	0.037	0.0018 to 0.741	97.7%	92.1%
4	0.707	0.045	0.0022 to 0.929	97.1%	90.3%
8	1.061	0.053	0.0025 to 1.113	96.8%	89.8%
16	1.415	0.055	0.0026 to 1.166	96.7%	89.4%
Not using historical controls		0.076	0.0035 to 1.673	94.9%	85.4%

bias or simply discounting its sample size (Section 6.9). For a single historical source, and assuming normal likelihoods, all these methods lead to essentially the same model (Section 5.4), and here we shall illustrate the use of the bias model (Pocock, 1976).

Assuming a model such as (6.27) requires prior opinion concerning the potential extent of the bias as measured by σ_δ. For example, if it were thought that in fact the historical controls may over- or underestimate the odds of death in the randomised controls by a factor of 2, then $\exp(1.96\sigma_\delta) = 2$, or $\sigma_\delta = (\log(2)/1.96) = 0.35$: this is similar to the analysis in Section 5.7.3 for interpreting the standard deviation of random effects. Table 6.11 gives a variety of values for σ_δ corresponding to beliefs which range from acceptance of the historical evidence at 'face value', *i.e.* $\sigma_\delta = 0$, to stating that the potential bias could be such that the historical controls could over- or underestimate the odds of death in the randomised controls by a factor of 16.

The choice of a suitable value for σ_δ will depend on the circumstances and the extent to which Pocock's criteria are met (Section 6.9). In this

instance the historical controls seem reasonable in that they came from the same centre and were treated in a similar way, except they were not involved in a clinical trial which is known can have an impact on outcomes.

Loss function or demands: No, but an OR of 0.4 was taken to be of clinical importance by Kass and Greenhouse (1989).

Computation/software: Conjugate normal model.

Evidence from study: The results of phase I of the ECMO study are shown in Table 6.10: of the ten patients randomised to conventional therapy four died, whilst of the nine randomised to ECMO none died. The estimates and variances of the log-odds of death were obtained using the adjustments given in Section 2.4. We note the apparent contrast between the mortality rates under CMT before and during the trial: it is generally felt that all participants in a randomised trial get superior treatment. Using the randomised evidence alone, the treatment effect θ would be estimated by $-2.94 + 0.37 = -2.57$, with variance $2.11 + 0.38 = 2.49$. A traditional standardised test statistic, ignoring the sequential nature of the design, is therefore $-2.57/\sqrt{2.49} = 1.63$, corresponding to a one-sided *P*-value of 0.052; Fisher's exact test yields a one-sided *P*-value of 0.054 (Ware, 1989).

Bayesian interpretation: We first consider an analysis with a reference prior on the treatment effect. If the historical evidence is totally discounted ($\sigma_\delta = \infty$) then it can be seen from Table 6.11 that the posterior mean of the odds ratio is 0.076, and the posterior probablity of ECMO being inferior is 5.1%; the posterior probability of ECMO not being clinically superior, *i.e*. an odds ratio above 0.4, is 14.6%. However, treating the historical controls as exchangeable with the randomised controls, *i.e*. at 'face value' ($\sigma_\delta = 0$), gives a posterior mean for the odds ratio of 0.033, but now the probability of ECMO being inferior is only 1.3%, and of it not being clinically superior is 5.1%.

Sensitivity analysis: Table 6.11 displays a range of intermediate results between the extremes of totally accepting and totally ignoring the historical controls. A 95% posterior interval for the odds ratio will exclude 1 provided σ_δ is less than around 8, corresponding to a relative bias of around 5. The probability of the odds ratio being less than 0.4 is only around 95% provided that the historical controls are accepted at near face value.

We might also consider a sceptical prior on the treatment effect: the original alternative hypothesis in the Harvard trial was a reduction of the mortality rate from 80% to 20%, equivalent to an odds ratio of 1/16 or $\log(\text{OR}) = -2.77$. Using the argument in Section 5.5.2, we might assume a prior centred on 0 and with 5% of its probability below this alternative of -2.77 – this corresponds to a prior standard deviation of

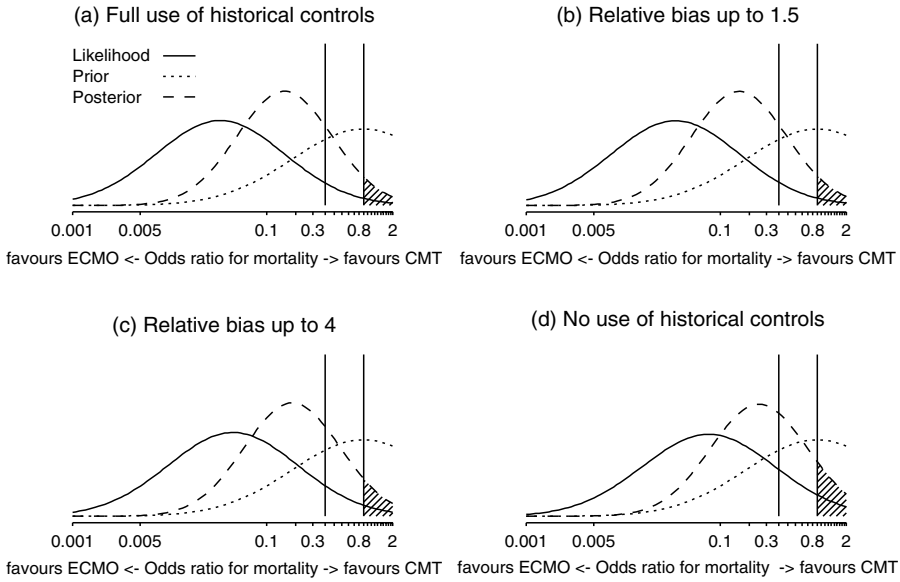

Figure 6.14 Sensitivity analysis of different choices of potential bias in historical controls in the ECMO trial, assuming a sceptical prior with mean 0 (on the log(OR) scale), and a 5% chance of an odds ratio less than 0.0625.

$-2.77/-1.64 = 1.69$. The consequences of using such a sceptical prior are shown in Figure 6.14 for a range of choices of potential bias in the historical controls. As Kass and Greenhouse (1989) conclude, a reasonable sceptic, even taking account of the historical data, is not going to be completely convinced by the ECMO trial.

Comments: This trial presents a number of interesting challenges which are fully argued in the discussion of Ware (1989) and in subsequent publications. For example, there are other historical data available, including some which show good survival on CMT, and there is a database of outcomes on ECMO. Other statistical models for this trial, including and discounting historical data, have been considered by Kass and Greenhouse (1989), Greenhouse and Wasserman (1995) and Berry and Stangl (1996a). Berry (1989b) also considers the inclusion of evidence from an RCT using a play-the-winner design which was also conducted before 1985. Such information could be included, if assumed to be exchangeable with the study reported by Ware (1989), using either a meta-analytic approach (Section 8.2) or by using this historical trial evidence to derive a prior distribution for the intervention effect (Section 5.4).

The discussants of Ware (1989) also have opposing views concerning the ethics of randomisation (Section 6.4): Royall and Berry (1989) say

the trial should never have been started since it was unethical to randomise given the available evidence, whereas Begg (1989) takes the completely conflicting view that the Harvard trial was stopped too early since, as we have seen in the analysis above, the result was not convincing to a wide range of opinion.

It is notable that the evidence concerning ECMO was not considered sufficient to prevent a further large trial. After ECMO was introduced in the UK in 1989, it was agreed to organise a randomised trial involving 55 referral hospitals, in which patients randomised to ECMO were referred to one of five specialist centres (Field *et al*., 1996). This pragmatic trial was designed to randomise 300 babies, but the DMC stopped the trial after 185 cases when the mortality rate was 30/93 on ECMO and 54/92 on CMT, with an odds ratio of 0.55 (95% interval from 0.39 to 0.77). Long-term follow-up of the patients over 4 years (Bennett *et al*., 2001) revealed only one additional death (in the ECMO arm) but a high rate of disability and impairment: overall only 16% of survivors were without abnormal signs or disability, but with no significant excess in the ECMO group. Treatment was, however, confounded with hospital and the trial was of a referral service rather than ECMO being carried out in direct competition to conventional treatment.

6.10 DATA-DEPENDENT ALLOCATION

So far we have only covered standard randomisation designs in which patients are allocated 50:50 or in some other constant ratio to alternative treatments. However, a full decision-theoretic approach to trial design would consider data-dependent allocation so that, for example, in order to minimise the number of patients getting the inferior treatment, the proportion randomised to the apparently superior treatment could be increased as the trial proceeded. Such 'adaptive' designs are claimed to satisfy ethical considerations for the patients under study (Section 6.4). They can be called 'bandit' designs, as they are analogous in theory to a gambler deciding which arm of a two-armed bandit to pull in order to maximise the expected return: both Bayesian and non-Bayesian approaches are available. An extreme example is Zelen's (1969) 'play-the-winner' rule in which the next patient is given the currently superior treatment, and randomisation is dispensed with entirely; Palmer and Rosenberger (1999) review non-standard trial designs and suggest circumstances where they may be appropriate. Palmer (2002) claims that many of the current difficulties faced in carrying out trials could be relieved by using adaptive designs, and Berry (2001) provides a recent argument for their use.

Nevertheless, there has been considerable criticism of these ideas as not being practically rooted in the realities of clinical trials; see, for example, Byar *et al.* (1976), Simon (1977), Armitage (1985) and Peto (1985). Objections to adaptive allocation include the following:

1. Responses have to be observed without delay.
2. Adaption depends on a one-dimensional response.
3. Sample sizes may have to be bigger.
4. Patients may not be homogeneous throughout the trial.
5. Clinicians may be unhappy with adaptive randomisation.
6. Informed consent may be more difficult to obtain.
7. The trial will be complex and may deter recruitment.
8. Estimation of the treatment contrast will lose efficiency.
9. Potential inflation of Type I error.
10. Treatment assignments may be biased as clinicians may guess which treatment is 'in the lead'.

A careful analysis of two-armed trials has been carried out by Berry and Eick (1995), who conclude that balanced allocation is appropriate if the condition is reasonably common, but adaptive designs may yield a substantial improvement in the expected number of successful treatments when a large proportion of patients with the disease are likely to be in the trial. This is echoed by Senn (1997b, p. 88), who points out that future patients, who in general will greatly outnumber those in the trial, would value a more precise treatment estimate and therefore would prefer large trials with balanced allocation. The ECMO studies discussed in Example 6.9 provide one of the few examples of adaptive allocation, and the subsequent controversy did little to encourage the use of such designs; other examples include an adaptive trial in patients with depressive disorder (Tamura *et al.*, 1994), while the trial described in Kadane (1996) also adapts its allocation rules, in a somewhat complex way, to the current evidence.

A recent example has proved, however, that it is possible to carry out a large and complex adaptive trial. Berry *et al.* (2001a) describe the design of a phase II/III dose-finding study in acute stroke, in which 15 different doses were to be given at random at the start of randomisation, with steady adaptation to the range of doses around the ED95, i.e. the minimum dose that provides 95% of the maximum efficacy. This trial has now been completed. Various characteristics may have contributed to the success of the methodology: only short-term (90-day) outcomes were considered, modern communication technology was used to ensure rapid updating of the current posterior distribution of the dose–response curve, a minimum of 15% of patients given placebo dose ensured that the imbalance did not become too acute, the ability to completely blind clinicians as to the dose provided, the replacement of the original decision-theoretic stopping criterion with one based on posterior tail areas being less than a certain value, and classical estimation of the size and power of the study based on pre-trial simulations.

We may conclude that adaptive designs, which are not a specifically Bayesian issue, may be better accepted when there are many arms in the trial and not just an imbalanced randomisation between two arms. In addition, formulation of a trial as a decision rather than an inference problem leads to many objections (Section 6.2), and adaptation may be better based on posterior distributions.

6.11 TRIAL DESIGNS OTHER THAN TWO PARALLEL GROUPS

Equivalence trials. There is a large statistical literature on trials designed to establish equivalence between therapies. From a Bayesian perspective the solution is straightforward: define a region of equivalence (Section 6.3) and calculate the posterior probability that the treatment difference lies in this range – a threshold of 95% or 90% might be chosen to represent strong belief in equivalence. Several examples of this remarkably intuitive approach have been reported (Section 6.13), which tend to give similar results to traditional analysis. In contrast, Lindley (1998) explores a decision-theoretic formulation that can give radically different conclusions.

Crossover trials. The Bayesian approach to crossover designs, in which each patient is given two or more treatments in an order selected at random, is fully reviewed by Grieve (1994a). More recent references concentrate on Gibbs sampling approaches (Forster, 1994) – see Section 6.13 for other relevant papers.

N-of-1 trials. *N*-of-1 studies can be thought of as repeated within-person crossover trials in which interest focuses on the response of an individual patient: such trials may be appropriate in chronic conditions in which short-term symptom relief is of interest. A natural approach to combining such studies is to assume patients are exchangeable (perhaps conditional on covariates), and adopt a hierarchical model – an example based on Zucker *et al*. (1997) is given in Example 6.10. This can be thought of as an extreme example of the subset procedure described in Section 6.8.1, in which the subsets have been reduced to individual patients.

Example 6.10 *N of 1: pooling individual response studies*

Reference: Zucker *et al*. (1997).

Intervention: Amitriptyline for treatment of fibromyalgia to be compared with placebo.

Aim of study: To estimate population treatment effects and evaluate individual patient responses.

Study design: Each individual had an *N*-of-1 study in which they were treated in a number of periods (3 to 6 per patient), and in each period both amitriptyline and placebo were administered in random order. All trials were carried out by a single physician at a single centre.

Outcome measure: Each measurement comprised a difference (amitriptyline minus placebo) in response to a symptom questionnaire in each paired crossover period. Higher scores indicated fewer negative symptoms, and so a positive difference indicated amitriptyline as the superior treatment.

Statistical model: If y_{kj} is the jth measurement on the kth individual, we assume

$$y_{kj} \sim \mathrm{N}[\theta_k, \sigma_k^2].$$

We then assume that both θ_ks and σ_k^2s are exchangeable, as it may not be reasonable to assume common between-period variability for all individuals. We make the specific distributional assumption that

$$\begin{aligned} \theta_k &\sim \mathrm{N}[\mu_\theta, \tau_\theta^2], \\ \log(\sigma_k^2) &\sim \mathrm{N}[\mu_\sigma, \tau_\sigma^2]. \end{aligned}$$

A normal distribution for the log-variances is equivalent to a log-normal distribution for the variances (Section 2.6.8).

Prospective analysis?: No.

Prior distribution:

Independence model. In order to reproduce the classical analysis, we may assume each θ_k has a uniform distribution, and each σ_k^{-2} has a Gamma[0.001,0.001] distribution. The latter is essentially equivalent to $\log(\sigma_k^2)$ having a uniform distribution and hence leads to the classical t distribution as a basis for testing for an effect in an individual (Sections 5.5.1 and 5.7.3).

Exchangeable model. We initially adopt uniform priors for μ_θ, τ_θ, μ_σ and τ_σ. Other prior distributions for the between-individual variation τ_θ are considered as part of a sensitivity analysis.

Loss function or demands: Zucker *et al*. (1997) suggest that a difference of 0.5 might be considered as important.

Computation/software: Markov chain Monte Carlo in WinBUGS software.

Evidence from study: The raw data are shown in Figure 6.15, ordered in terms of the observed sample mean. Seven out of 23 experienced

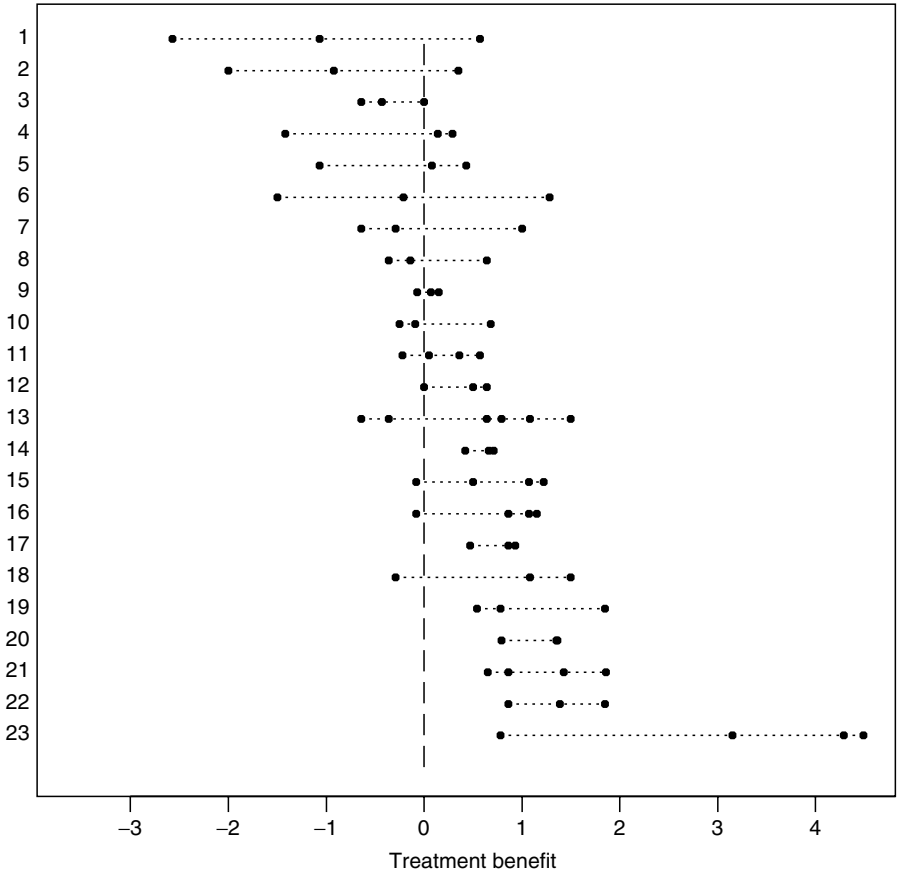

Figure 6.15 Raw data from *N*-of-1 clinical trials on 23 patients, ordered by their mean response. Each dot represents the difference in responses (amitriptyline minus placebo) in a single period in which both treatments have been tried in random order.

benefit from the new treatments in all their periods. There appears to be substantial variability both in the average response and within patients, justifying the statistical model adopted.

Bayesian interpretation: The independent and exchangeable estimates of the individual and overall treatment effects are shown in Figure 6.16. The independent estimates closely follow the raw data, exhibiting substantial uncertainty. In only six patients do the 95% intervals exclude 0, although Zucker *et al*. (1997) report that patients 11–23 were all advised to continue on the active treatment, while patients 1–10 were advised to stop active treatment.

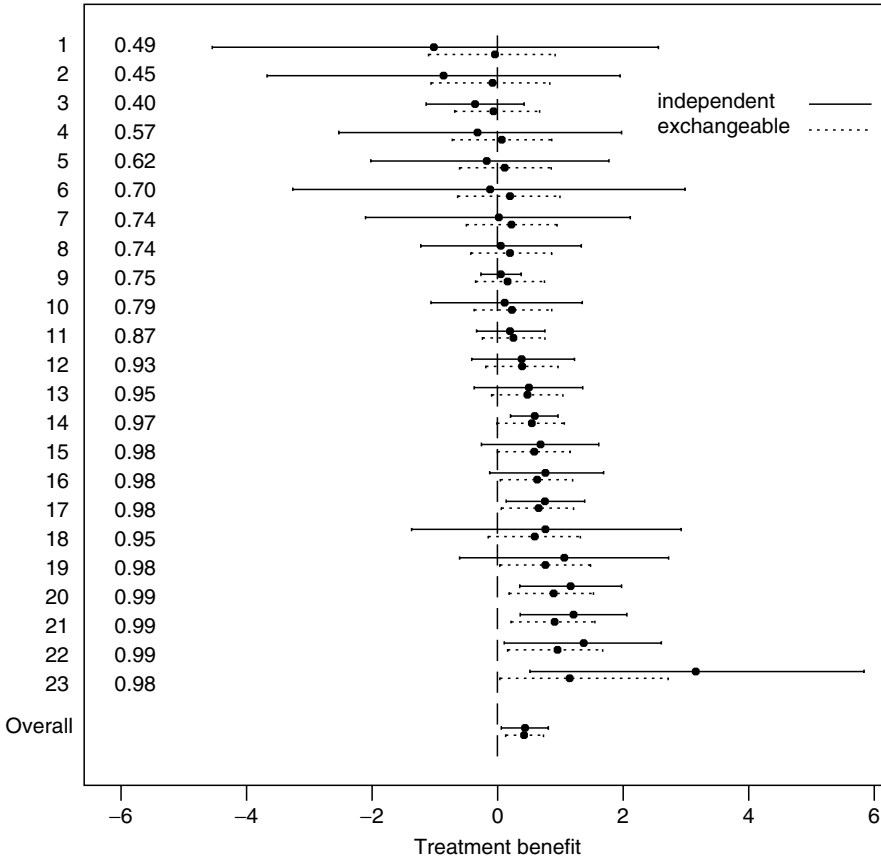

Figure 6.16 Estimates and 95% intervals for the response in each person, assuming both independent and exchangeable individuals. The vertical lines represent the null hypothesis of no treatment difference. $P(\theta_k > 0)$, the posterior probability that each individual's effect lies above 0, is given on the left.

Table 6.12 Summary of posterior distributions of parameters in exchangeable analysis.

Parameter		Median / estimate	95% interval
Overall mean	μ_θ	0.42	0.13 to 0.73
Prob. overall positive effect	$P(\mu_\theta > 0)$	0.997	
Prob. overall important effect	$P(\mu_\theta > 0.5)$	0.29	
Between-patient sd	τ_θ	0.50	0.20 to 0.92
Between-patient variability in log-variances	τ_σ	1.03	0.42 to 1.77
Mean within-patient variance	$\exp(\mu_\tau + \tau_\theta^2)/2$	0.94	0.49 to 3.05

The parameter estimates resulting from the exchangeable analysis are shown in Table 6.12. There is a clear overall positive effect in the population which is estimated to be 0.42, although the chance that it is an important effect (*i.e*. greater than 0.5) is only 29%. There is also strong evidence of patient heterogeneity in their response, with an estimated between-patient standard deviation of 0.50, suggesting that individual patient effects might vary between roughly -0.5 and 1.5.

There is also clear evidence of between-patient heterogeneity in their variability in responses, as shown by τ_σ being substantially away from 0. Transforming from a log-variance to a variance scale (Section 2.6.8) reveals a mean within-patient variance of 0.94.

From the individual estimates shown in Figure 6.16 it is clear that the exchangeable model brings about substantial shrinkage in the extreme patients, reflecting the limited information from each individual. For example, patient 23, with four positive measurements, three of which are extreme, has a posterior mean of 0.55, less than its minimum observation! It might be felt that the model is exercising undue influence in this situation, and some possible alternatives are discussed below. In spite of the shrinkage, the narrower intervals mean that the number of patients with 95% intervals excluding 0 rises to nine, compared to six with the independent analysis. We note one consequence of allowing exchangeable within-patient variances: patient 9, whose observations were remarkably close together and who hence has a very tight independent interval, obtains an exchangeable interval that is *wider* due to their within-patient variance being pulled towards the population mean of around 0.94.

Sensitivity analysis: Changing the prior distribution for τ_θ to the alternatives listed in Section 5.7.3 makes negligible difference to the conclusions, due to the considerable evidence available concerning τ_θ.

Comments: As pointed out by Zucker *et al*. (1997), it is straightforward to include patient-level covariates in such a model, and they illustrate this by including dose as a predictor. However, this can be shown to have minimal influence. It might be reasonable to carry out further analysis of sensitivity to the shape of both the sampling and the random-effects distribution: assuming *t* distributions (Section 2.6.9) for either may result, for example, in substantially less shrinkage for patient 23.

Factorial designs: Factorial trials, in which multiple treatments are given simultaneously to patients in a structured design, can be seen as another example of multiplicity and hence a candidate for hierarchical models. Simon and Freedman (1997) and Miller and Seaman (1998) suggest suitable prior assumptions that avoid the need to decide whether interactions do or do not exist.

6.12 OTHER ASPECTS OF DRUG DEVELOPMENT

Pharmacokinetics. The 'population' approach to pharmacokinetics, in which the parameters underlying each individual's drug clearance curve are viewed as being drawn from some population, is well established and is essentially an empirical Bayes procedure (Sheiner and Wakefield, 1999). Proper Bayesian analysis of this problem is described in Racine-Poon and Wakefield (1996) and Wakefield and Bennett (1996), emphasising MCMC methods for estimating both population and individual parameters, as well as individualising dose selection (Wakefield and Walker, 1997).

Phase I trials. Phase I trials are conducted to determine that dosage of a new treatment which produces a level of risk of a toxic response which is deemed to be acceptable. The primary Bayesian contribution to the development of methodology for phase I trials has been the continual reassessment method (CRM) originally proposed by O'Quigley *et al*. (1990). In CRM a parameter underlying a dose–toxicity curve is given a proper prior which is updated sequentially and used to find the current 'best' estimate of the dosage which would produce the acceptable risk of a toxic event if given to the next subject, as well as giving the probability of a toxic response at the recommended dose at the end of the trial (O'Quigley, 1992). High sensitivity of the posterior to the prior distribution (Gatsonis and Greenhouse, 1992) has been reported in a similar procedure. Numerous simulations and modifications of the method have been proposed (Section 6.13); Dougherty *et al*. (2000) report a practical application described in Example 6.11.

Example 6.11 *CRM: An application of the continual reassessment method*

Dougherty *et al*. (2000) provide the following application of the continual reassessment method, in which they wish to establish the maximum tolerated dose of the opioid antagonist nalmefene. Lack of tolerability is measured by reversal of anaesthesia. They are interested in establishing the maximum dose with probability p of reversal of anaesthesia nearest to 0.20. The available doses are 0.25, 0.50, 0.75 and 1.00, which are given labels 1 to 4. They adopt a one-parameter logistic response model in which, for dose i,

$$\text{logit}(p_i) = 3 + \alpha d_i, \tag{6.28}$$

where α is an unknown parameter with prior set as an exponential distribution with mean 1 (*i.e*. Gamma[1,1]), and the d_i are transformations of the dose to enable this logistic curve to fit the prior judgements of p_i, denoted p_i^0. Hence the d_i are calculated by setting α equal to its prior mean of 1, and inverting (6.28) to give $d_i = \text{logit}(p_i^0) - 3$.

Table 6.13 Summary of prior and posterior distributions of parameters in CRM experiment.

	Prior		Observed data		Posterior	
Dose	p_i^0: prior guess at p_i	d_i	No. patients	No. not tolerating	Mean	SD
1	0.10	−5.20	4	0	0.10	0.05
2	0.20	−4.39	18	3	0.19	0.08
3	0.40	−3.41	3	2	0.38	0.09
4	0.80	−1.61	0	0	0.79	0.03

Table 6.13 shows the prior judgements, the observed data and consequent posterior distributions. The analysis is straightforward to carry out in WinBUGS.

We can make a number of observations concerning this analysis. First, the posterior means for the p_i show strong agreement with the prior, perhaps suggesting undue influence. Second, the actual doses used do not enter into the model. Third, a tolerability for dose 4 is estimated with considerable accuracy, even though no one was ever given this dose. Finally, the implied prior distributions for the p_i are actually bimodal. These all suggest that the basic CRM procedure should be used with great caution.

Etzioni and Pepe (1994) suggest monitoring a phase I trial with two possible adverse outcomes via the joint posterior distribution of the probabilities of the two outcomes with frequentist inference at the end of the trial.

Phase II trials. Phase II clinical trials are carried out in order to discover whether a new treatment is promising enough (in terms of efficacy) to be submitted to a controlled phase III trial, and often a number of doses may be compared. Bayesian work has focused on monitoring, sample-size determination and adaptive design. Monitoring on the basis of posterior probability of exceeding a desired threshold response rate was first recommended by Mehta and Cain (1984), while Heitjan (1997), Cronin *et al.* (1999) and Weiss *et al.* (2001) adapt the proposed use of sceptical and enthusiastic priors (Section 6.6.2) in phase III studies.

With regard to design, Herson (1979) used predictive probability calculations to select among designs with high power in regions of high prior probability. Thall and co-workers have also developed stopping boundaries for sequential phase II studies based on posterior probabilities of clinically important events, but where the designs are selected from the frequentist properties derived from extensive simulation studies: see Section 6.13 for references. However Stallard

(1998) has criticised this approach as being demonstrably sub-optimal when evaluated using a full decision-theoretic model with a monetary loss function.

Finally, John Whitehead and colleagues have taken a full decision-theoretic approach to allocating subjects between phase II and phase III studies. For example, Brunier and Whitehead (1994) consider the case where a single treatment with a dichotomous outcome is being evaluated for a possible phase III trial, and use Bayesian decision theory to determine the number of subjects needed. They place a prior on the probability of success and calculate the expected cost of performing or not performing a phase III trial, using a cost function which includes consideration of the costs to future patients if the inferior treatment is eventually used, the power of the possible phase III trial (which they assume will be carried out by frequentist methods), and the costs of experimentation. They show how to determine, for given parameter values, the expected cost of performing a phase II trial of any particular size, and thus the optimal size for a trial.

When faced with selecting among a list of treatments and allocating patients, Pepple and Choi (1997) have considered two-stage designs, Yao *et al.* (1996) deal with screening multiple compounds and allocating patients within a programme, while Strauss and Simon (1995) use a prior distribution and horizon. The successful adaptive study of Berry *et al.* (2001a) discussed in Section 6.10 can also be considered as a phase II dose-finding study monitored using posterior tail areas.

Phase IV – safety monitoring. A considerable literature exists on Bayesian causality assessment in adverse drug reactions: see, for example, Lanctot and Naranjo (1995).

6.13 FURTHER READING

There is a huge literature on Bayesian appraches to trials, which is reviewed in Spiegelhalter *et al.* (2000). General discussion papers include tutorial introductions at a non-technical (Lewis and Wears, 1993) and slightly more technical level (Abrams *et al.*, 1994). Pocock and Hughes (1990) provide a non-mathematical discussion concentrating on estimation issues, while Armitage (1989) attempts a balanced view of the competing methodologies. A special issue of *Statistics in Medicine* has been devoted to 'Methodological and Ethical Issues in Clinical Trials', containing papers both for (Berry, 1993; Urbach, 1993; Spiegelhalter *et al.*, 1993) and against (Whitehead, 1993) the Bayesian perspective, and featuring incisive discussion by Armitage, Cox and others. Particular emphasis has been placed on the ability of Bayesian methods to take full advantage of the accumulating evidence provided by small trials (Lilford *et al.*, 1995; Matthews, 1995).

Somewhat more technical reviews are given by Spiegelhalter *et al.* (1993, 1994). Berry (1991, 1995) has long argued for a Bayesian decision-theoretic

basis for clinical trial design, and has described in detail methods for elicitation, monitoring, decision-making and using historical controls. Proponents of a decision-theoretic choice of sample size include Claxton and Posnett (1996), Hornberger and Eghtesady (1998) and Hornberger (2001).

Pocock (1992), O'Brien (1998) and Whitehead (1997b) provide good reviews on sequential trials, and applications of monitoring using posterior intervals include Berger and Berry (1988), Brophy and Joseph (1997), Carlin *et al.* (1993), DerSimonian (1996), George *et al.* (1994) and Rosner and Berry (1995). Papers investigating monitoring using predictions include Choi and Pepple (1989), Qian *et al.* (1996) and Spiegelhalter *et al.* (1986).

Empirical Bayes analyses of subsets are provided by Louis (1991) and Pocock and Hughes (1990), which give rise to traditional confidence intervals that are not given a Bayesian interpretation. Bayesian techniques for subsets are elaborated in Dixon and Simon (1991), Simon (1994b) and Simon *et al.* (1996). Hierarchical models for multicentre analysis have been considered by Gray (1994), Stangl (1996) and Stangl and Greenhouse (1998), while Matsuyama *et al.* (1998) allow a random centre effect on both baseline hazard and treatment, and examine the centres for outliers using a Student's t prior distribution for the random effects.

Examples of the Bayesian approach to equivalence trials have been reported by Selwyn *et al.* (1981), Fluehler *et al.* (1983), Selwyn and Hall (1984), Breslow (1990), Grieve (1991) and Baudoin and O'Quigley (1994). Bayesian approaches to crossover trials include Grieve (1985, 1995), Albert and Chib (1996) and Grieve and Senn (1998).

The continuous reassessment method for phase I studies has been developed by Goodman *et al.* (1995), Whitehead and Brunier (1995), and Gasparini and Eisele (2000). For phase II studies, Korn *et al.* (1993) consider a phase II study which was stopped after three out of four patients exhibited toxicity; Bring (1995) and Greenhouse and Wasserman (1995) re-examine their problem from a Bayesian perspective. See also Thall and Estey (1993), Thall *et al.* (1996), Thall and Russell (1998) and Whitehead (1986, 1997a).

6.14 KEY POINTS

Table 6.14 briefly summarises some major distinctions between the Bayesian and the frequentist approach to trial design and analysis.

1. The Bayesian approach provides a framework for considering the ethics of randomisation.
2. Prior information can be incorporated in power calculations, which should warn against conditioning on optimistic alternative hypothesis. 'Average' power may give a more realistic assessment of the chances of a trial reaching a positive conclusion.

Table 6.14 A brief comparison of Bayesian and frequentist methods in clinical trials.

Issue	Frequentist	Bayesian
Information other than that in the study being analysed	Informally used in design	Used formally by specifying a prior probability distribution
Interpretation of the parameter of interest	A fixed state of nature	An unknown quantity which can have a probability distribution
Basic question	How likely are the data given a particular value of the parameter?	How likely is a particular value of the parameter, given the data?
Presentation of results	Likelihood functions, *P*-values, confidence intervals	Plots of posterior distributions of the parameter, calculation of specific posterior probabilities of interest, and use of the posterior distribution in formal decision analysis
Interim analyses	*P*-values and estimates adjusted for the number of analyses	Inference not affected by the number or timing of interim analyses
Interim predictions	Conditional power analyses	Predictive probability of getting a firm conclusion
Dealing with subsets in trials	Adjusted *p*-values (*e.g.* Bonferroni)	Subset effects shrunk towards zero by a 'sceptical' prior

3. Monitoring trials with a sceptical and other priors may provide a unified approach to assessing whether a trial's results would be convincing to a wide range of reasonable opinion, and could provide a formal tool for data monitoring committees.
4. Predictions of the consequences of continuing a trial provide a useful adjunct to current posterior distributions, but should not be used as a formal monitoring tool.
5. Various sources of multiplicity can be dealt with in a unified and coherent way using hierarchical models.
6. A variety of models exist for incorporating historical controls, analogous to those for using historical data as a basis for a prior distribution.
7. Adaptive studies that change the randomisation ratio dependent on outcomes may be appropriate when a large proportion of available patients are taking part in the trial, or when many treatment arms are being simultaneously investigated.
8. It is generally unrealistic to formulate a phase III trial as a decision problem, except in circumstances where future treatments can be reasonably predicted. Earlier phase studies may be more amenable to this approach.

EXERCISES

6.1. Prove (6.4), (6.6) and (6.7).

6.2. In Example 6.2, calculate the expected power given that the treatment is effective. [Hint: There are two possible methods. You could generate the joint distribution of θ and the power, and only count those iterations for which $\theta > 0$. Alternatively, generate θ from its prior distribution constrained to be positive, using the I(0,) construct in WinBUGS.]

6.3. Consider the prior beliefs for the MRC neutron therapy RCT introduced in Exercise 5.2. The actual trial results at an interim analysis produced a hazard ratio of 0.66 (95% CI from 0.40 to 1.10) in favour of the control group. For each of the prior distributions in Exercise 5.2, update these priors in the light of the observed results.

6.4. Ben-Shlomo *et al.* (1998) report the results of the UK Parkinson's Disease Research Group RCT of the evaluation of levodopa, levodopa and selegiline, and bromocriptine in the treatment of early stage Parkinson's disease; we focus on the comparison of levodopa against levodopa and selegiline in terms of mortality. At a second interim analysis 44 deaths were observed out of 249 patients in the levodopa alone arm and 76 out of 271 patients in the levodopa and selegiline arm, producing a hazard ratio of 1.57 (95% CI from 1.09 to 2.03) for levodopa and selegiline vs. levodopa alone. At this point the trial was terminated, but follow-up continued and a subsequent analysis reported 73 and 103 deaths, producing a hazard ratio of 1.32 (95% CI from 0.98 to 1.79).
   (a) Use the credibility analysis of Section 3.11 to establish the degree of scepticism that would be required not to have found the interim results convincing of benefit.
   (b) In a trial in which $m = 120$ events were to be observed, what alternative log(hazard ratio) could be detected with 80% power?
   (c) What sceptical prior would express 5% belief that the effect would be as large as this alternative hypothesis?
   (d) Discuss whether, on the evidence provided, it was reasonable to stop the trial early.

6.5. Table 6.15, adapted from Wheatley and Clayton (2003), shows the accumulating data in a trial of five vs. four treament courses in the MRC Acute Myeloid Leukaemia trial. An unexpectedly large treatment effect in favour of five courses was observed early in the trial, which disappeared as the trial progressed.
   (a) Plot the likelihoods for the log(hazard ratio) at each timepoint, and calculate the two-sided P-values.
   (b) If the trial were planned to observe 300 events, what might a reasonable sceptical prior distribution be?
   (c) What would have been the effect had this prior been used to monitor the trial?

Table 6.15 Mortality in MRC Acute Myeloid Leukaemia RCT.

Timepoint	5 courses		4 courses		$O-E$	$V[O-E]$
	deaths	total	deaths	total		
1997	7	102	15	100	−4.6	5.5
1998(1)	23	171	42	169	−12.0	15.9
1998(2)	41	240	66	240	−16.0	26.7
1999	51	312	69	309	−11.9	30.0
2000	79	349	91	345	−9.5	42.4
2001	106	431	113	432	−6.2	53.7
2002	157	537	140	541	+6.7	74.0

6.6. Prove (6.17) and (6.18).

6.7. Consider the situation in which the Parkinson's disease trial was stopped in Exercise 6.4, and the predictions that could have been made concerning the status of the trial at its eventual publication when 176 events had occurred (an additional 56).
   (a) What would have been the expected power, given the data so far, of rejecting the hypothesis that the log(hazard ratio) was 0, *i.e.* the probability that the final 95% interval will lie wholly above 0, with and without the inclusion of the sceptical prior?
   (b) Was there evidence of conflict between the data in the first part of the trial and that collected in the second part, *i.e.* after the decision was made to stop? [Hint: One way to do this is to calculate the predictive distribution for the observed log(hazard ratio) arising in the second part and use Box's measure of conflict to compare it to that actually observed.]

6.8. (a) Derive the results given in the ECMO study in Example 6.9. (b) Reanalyse the ECMO study assuming the historical data are to be discounted using the 'power prior' model explored in Example 5.2, with prior weights 0, 10%, 50% and 100%.

6.9. Reanalyse the ECMO study in Example 6.9 with full binomial likelihoods instead of normal approximations and using WinBUGS for the analysis. You will need to select a prior distribution for the mortality rates in the control and ECMO groups ignoring both historical and trial data: compare the use of (a) independent uniform distributions in each group, (b) independent Beta[0.5,0.5] distributions, (c) a uniform distribution for the control group mortality and a sceptical prior for the treatment effect on the log(odds ratio) scale.

6.10. Consider Exercise 2.1, repeating the study with the *other* hand. Using a subjectively chosen sceptical prior distribution for the log(odds ratio) for the difference between hands, conduct the second 12 tosses, and update the prior beliefs in the light of the evidence that you have collected.

Table 6.16 Estimates of log(hazard ratio) and standard errors for disease-free survival comparing tamoxifen with control for women with breast cancer within subgroups defined by œstrogen receptor status, nodal status and postmenopausal status.

			No. patients				
Oestrogen receptor +ve	Node + ve	Postmenopausal	Total	Tamoxifen	Control	log (HR)	SE [log(HR)]
1	0	0	183	72	111	−0.520	0.207
1	1	0	57	27	30	−0.096	0.319
1	0	1	262	101	161	−0.551	0.190
1	1	1	92	44	48	+0.040	0.278
0	0	0	493	210	283	−0.061	0.152
0	1	0	128	52	76	−0.256	0.242
0	0	1	583	280	303	−0.287	0.131
0	1	1	161	72	89	−0.275	0.205

6.11. Table 6.16 displays estimates of log(hazard ratio) for disease-free survival comparing tamoxifen with control for women with breast cancer for eight mutually exclusive subgroups of women defined by three binary factors: œstrogen receptor status, nodal status and postmenopausal status. Assuming exchangeable subgroups, obtain the posterior estimates of the hazard ratio for each subgroup, and thus assess the evidence for specific subgroup–treatment interactions. [Hint: You could use the empirical Bayes methodology of Example 3.13, or the full Bayes approach using WinBUGS shown in Example 8.1.]. Do you think the exchangeability assumption is reasonable?