5

Prior Distributions

5.1 INTRODUCTION

There is no denying that quantifiable prior beliefs exist in medicine. For example, in the context of clinical trials, Peto and Baigent (1998) state that 'it is generally unrealistic to hope for large treatment effects' and that 'it might be reasonable to hope that a new treatment for acute stroke or acute myocardial infarction could reduce recurrent stroke or death rates in hospital from 10% to 9% or 8%, but not to hope that it could halve in-hospital mortality'. However, turning informally expressed opinions into a mathematical prior distribution is perhaps the most difficult aspect of Bayesian analysis. Five broad approaches are outlined below: elicitation of subjective opinion; summarising past evidence; default priors; 'robust' priors; and estimation of priors using hierarchical models. The discussion mainly focuses on priors for the primary treatment effects of interest, although we also consider the difficult issue of specifying a prior for the variance component in a hierarchical model. Finally, we consider the criticism of prior assessments, from both an empirical and a methodological perspective.

We should repeat the statements made in Section 3.9 concerning possible misconceptions about prior distributions: they are not necessarily prespecified, unique, known or important. Since there is no 'correct' prior, Bayesian analysis can be seen as a means of transforming prior into posterior opinions, rather than producing *the* posterior distribution. It is therefore vital to take into account the context and audience for the assessment (Section 3.1), and analysis of sensitivity to alternative assumptions should be considered essential. Kass and Greenhouse (1989) introduced the term 'community of priors' to describe the range of viewpoints that should be considered when interpreting evidence, and the suggestions in this chapter represent possible members of that community.

It is also important to keep in mind that, in certain circumstances, it may be quite reasonable for a prior to be elicited and used solely for design purposes, and excluded when publicly reporting a study. However, when wishing to convince an audience of the benefits of an intervention, it may be important

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

to elicit their priors and possibly their utilities (Kadane and Wolfson, 1996).

From a mathematical and computational perspective, we have seen in Section 3.6.2 that it can be convenient if the prior distribution is a member of a family of distributions that is conjugate to the form of the likelihood, in the sense that they 'fit together' to produce a posterior distribution that is in the same family as the prior distribution. We also saw in Section 2.4 that in many circumstances likelihoods for treatment effects can be assumed to have an approximately normal shape, and thus in these circumstances it will be convenient to use a normal prior (the conjugate family), provided it approximately summarises the appropriate external evidence. Modern computing power is, however, reducing the need for conjugacy, and in this chapter we shall largely concentrate on the source and use of the prior rather than its precise mathematical form.

5.2 ELICITATION OF OPINION: A BRIEF REVIEW

5.2.1 Background to elicitation

A true subjectivist Bayesian approach requires only a prior distribution that expresses the personal opinions of an individual but, if the health-care intervention is to be generally accepted by a wider community, it would appear to be essential that the prior distributions have some evidential or at least consensus support. In some circumstances there may, however, be little 'objective' evidence available and summaries of expert opinion may be indispensable. We shall use the generic term 'clinical prior' for such expert assessments.

There is an extensive literature concerning the elicitation of subjective probability distributions from experts, with some good early references on statistical (Savage, 1971) and psychological aspects (Tversky, 1974), as well as on methods for pooling distributions obtained from multiple experts (Genest and Zidek, 1986). The fact that people are generally not good probability assessors is well known, and the variety of biases they suffer are summarised by Kadane and Wolfson (1997):

1. *Availability*. Easily recalled events are given higher probability, and vice versa.
2. *Adjustment and anchoring*. Initial assessments tend to exert an inertia, so that further elicited quantities tend to be insufficiently adjusted. For example, if a 'best guess' is elicited first, then subsequent judgements about an interval may be too close to the first assessment.
3. *Overconfidence*. Distributions are too tight.
4. *Conjunction fallacy*. A higher probability can be given to an event which is a subset of an event with a lower probability.
5. *Hindsight bias*. If the prior is assessed after seeing the data, the expert may be biased.

Nevertheless it has been shown that training can improve experts' ability to provide judgements that are 'well calibrated', in the sense that if a series of events are given a probability of, say, 0.6, then around 60% of these events will occur: see, for example, Murphy and Winkler (1977) with regard to weather forecasting.

Chaloner (1996) provides a thorough review of methods for prior elicitation in clinical trials, including interviews with clinicians, postal questionnaires, and the use of an interactive computer program to draw a prior distribution. She concludes that fairly simple methods are adequate, using interactive feedback with a scripted interview, providing experts with a systematic literature review, basing elicitation on 2.5th and 97.5th percentiles, and using as many experts as possible. Both Kadane and Wolfson (1996) and Berry and Stangl (1996a) emphasise the potential benefits of two approaches: eliciting predictive distributions of future events from which an implicit prior distribution can be derived, and asking additional questions as a consistency check.

5.2.2 Elicitation techniques

Methods used in practice can be divided into four main categories of increasing formality, which are listed here with some experience of their use:

1. *Informal discussion*. Prominent individuals can be informally interviewed for their opinion, as illustrated in Example 3.6. In a trial of paclitaxel in metastatic breast cancer, the study's principal clinical investigator expected the overall success rate to be 25% and had 50% belief that the true success rate lay between 15% and 35% (Rosner and Berry, 1995). Example 7.1 features priors obtained from two doctors for the relative risk of venous thrombosis associated with the use of oral contraceptives (Lilford and Braunholtz, 1996). There are clear difficulties in using such individual opinions in any formal context.
2. *Structured interviewing and formal pooling of opinion*. Freedman and Spiegelhalter (1983) describe an interviewing technique in which a set of experts were individually interviewed and hand-drawn plots of their prior distributions elicited, while deliberate efforts were made to prevent the opinions being overconfident (too 'tight'). The distributions were converted to histograms and averaged to produce a composite prior. This technique was also used for trials of thiotepa in superficial bladder cancer (Spiegelhalter and Freedman 1986) and osteosarcoma (Spiegelhalter *et al.*, 1993). Gore (1987) introduced the concept of 'trial roulette', in which 20 gaming chips, each representing 5% belief, could be distributed amongst the bins of a histogram: in a trial of artificial surfactant in premature babies, 12 collaborators were interviewed using this technique to obtain their opinion on the possible benefits of the treatment (Ten Centre Study Group, 1987). Using an elec-

tronic tool so that individuals in a group could respond without attribution, Lilford (1994) presented collaborators in a trial with a series of imaginary patients in order to elicit their opinions on the benefit of early delivery. The appropriate means of pooling such opinions is discussed in Section 5.2.3.

3. *Structured questionnaires*. The 'trial roulette' scheme described above was administered by post by Hughes (1991) for a trial in treatment of oesophageal varices and by Abrams *et al.* (1994) for a trial of neutron therapy. Parmar *et al.* (1994) elicited prior distributions for the effect of a new radiotherapy regime (CHART), in which the possible treatment effect was discretised into 5% bands and the form was sent by post to each of nine clinicians. Each provided a distribution over these bands and an arithmetic mean was then taken: see Example 5.1 for details. Tan *et al.* (2003) adapted this questionnaire, while Fayers *et al.* (2000) provide a similar questionnaire and document the variability between the elicited responses.

   Chaloner and Rhame (2001) provide a copy of the questionnaire they used to elicit opinions from 58 practising HIV clinicians concerning the baseline event rates and the potential benefit of two prophylactic treatments. This asks the minimum information comprising a point estimate and an estimated 95% interval. They used both post and telephone to carry out the elicitations.
4. *Computer-based elicitation*. Chaloner *et al.* (1993) provide a detailed case study of the use of a rather complex computer program that interactively elicited distributions from five clinicians for a trial of prophylactic therapy in AIDS. Kadane (1996) reports the results of an hour-long telephone interview with each of five clinicians, using software to estimate prior parameters from the results of a series of questions eliciting predictive probability distributions for responses of various patient types. When a second round of elicitation became necessary, the proposal was met by 'little enthusiasm'. Kadane and Wolfson (1996) provide an edited transcript of a computerised elicitation session in a non-trial context.

We agree with Chaloner (1996) that extremely detailed elicitation methods have not yet been shown to have any advantage over simple methods. However, it is feasible that complex policy problems, which necessarily may require substantial subjective input, would justify a more sophisticated approach. In any case, Chaloner and Rhame (2001) 'recommend documenting prior beliefs irrespective of whether a Bayesian or frequentist approach is taken to data analysis and formal statistical monitoring'.

5.2.3 Elicitation from multiple experts

Faced with varying prior distributions elicited from multiple experts, we could adopt one of a number of alternative strategies.

- *Elicit a consensus*. If the aim is to produce a single assessment expressing the belief of the group as a whole, then a range of techniques exist for bringing diverse opinions into consensus, including both informal and more formal Delphi-like methods. Care must of course be taken to avoid influence of dominant individuals.
- *Calculate a 'pooled' prior*. The choice of a method for pooling K multiple opinions is not clear cut, and Genest and Zidek (1986) provide a detailed annotated review of the issues. *Arithmetic pooling* simply takes the average of the height of the prior distributions for each parameter value θ, so that $p(\theta) = \sum_k p_k(\theta)/K$. This has the property that pooled probabilities for any event, such as tail areas, are also averages of the individually assessed tail areas. An alternative is *logarithmic pooling*, which takes the average of the logarithms of the density, equivalent to using a geometric mean of the original densities, so $p(\theta) \propto [\prod_k p_k(\theta)]^{1/K}$. This has the apparently attractive property that the same pooled posterior distribution is achieved, whether the pooling is done before or after the common likelihood is taken into account. With both proposals there is an opportunity to apply unequal weights to experts, dependent on their experience or past predictive ability. A further development is that of the *supra-Bayesian*, which takes the expressed opinions as data to manipulate using a statistical model.
- *Retain the individual priors*. The diversity of opinion might be just as important as the 'average' opinion, in that we may be interested in whether current evidence is sufficient to convince a full range of observers as to the benefits of a treatment, and hence to bring them into consensus. The extremes of opinion can be thought of as marking out the boundaries of the 'community of priors' mentioned in Section 5.1.

Our preference is to take a simple supra-Bayesian view, and treat the expressed heights of the prior distributions as data. Then, if we wish to assess the view of an 'average, well-informed participating clinician', it seems reasonable to simply use arithmetic pooling as in Example 5.1. Of course, we should not necessarily assume we have a random sample of clinicians, and so our estimate may be inevitably 'biased'.

Example 5.1 *CHART: Eliciting subjective judgements before a trial*

References: Parmar *et al*. (1994, 2001) and Spiegelhalter *et al*. (1994).

Intervention: In 1986 a new radiotherapy technique known as continuous hyperfractionated accelerated radio therapy (CHART) was introduced. The idea behind it was to give radiotherapy continuously (no weekend breaks), in many small fractions (three a day) and accelerated (the course completed in 12 days). There are clearly considerable logistical problems in efficiently delivering CHART.

Aim of studies: Promising non-randomised and pilot studies led the UK Medical Research Council to instigate two large randomised trials to compare CHART with conventional radiotherapy in both non-small-cell lung and head-and-neck cancer, and in particular to assess whether CHART provides a clinically important difference in survival that compensates for any additional toxicity and problems of delivering the treatment.

Study design: The trials began in 1990, randomised in the proportion 60:40 in favour of CHART, with planned annual meetings of the data monitoring committee (DMC) to review efficacy and toxicity data. No formal stopping procedure was specified in the protocol.

Outcome measure: Full data were to become available on survival (lung) or disease-free survival (head-and-neck), with results presented in terms of estimates of the hazard ratio, h, defined as the ratio of the hazard under CHART to the hazard under standard treatment. Hence, hazard ratios less than one indicate superiority of CHART.

Planned sample sizes: Lung: 600 patients were to be entered, with 470 expected deaths, with 90% power to detect at the 5% level a 10% improvement (15% to 25% survival). Using the methods described in Section 2.4.2, this can be seen to be equivalent to an alternative hypothesis of $h_A = \log(0.25)/\log(0.15) = 0.73$. Head-and-neck: 500 patients were to be entered, with 220 expected recurrences, with 90% power to detect at the 5% level a 15% improvement (45% to 60% disease-free survival), equivalent to an alternative hypothesis of $h_A = \log(0.60)/\log(0.45) = 0.64$.

Statistical model: Proportional hazards model, providing an approximate normal likelihood (Section 2.4.2) for the log(hazard ratio), $\delta = \log(h)$,

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

where y_m is the estimated log(hazard ratio), $\sigma = 2$ and m is the 'equivalent number of events' in a trial balanced in recruitment and follow-up.

Prospective analysis?: Yes, the prior elicitations were conducted before the start of the trials, and the Bayesian results presented to the DMC at each of their meetings.

Prior distribution: Although the participating clinicians were enthusiastic about CHART, there was considerable scepticism expressed by oncologists who declined to participate in the trial. Eleven opinions were elicited for the lung cancer trial and nine for the head-and-neck. The questionnaire used is described in detail in Parmar *et al*. (1994) and summarised in Figure 5.1.

	CHART worse than standard by %			CHART worse than standard by %						
	10 –15	5 –10	0 – 5	0 – 5	5 – 10	10 – 15	15 – 20	20 – 25	25+	TOTAL
Lung Study Your Entry										100
Head & Neck Study Your Entry										100
Hypothetical example	0	20	20	20	0	0	20	20	0	100

Figure 5.1 Part of the questionnaire used to elicit clinical opinions before the CHART trials. Participants were invited to distribute 100 points between the bins, indicating their 'weight of belief' in the true benefit from CHART. They were reminded to ignore the role of sampling variability – the hypothetical example was deliberately chosen to be a 'rather eccentric' radiotherapist so as not to provide an example that might inappropriately 'anchor' their opinions.

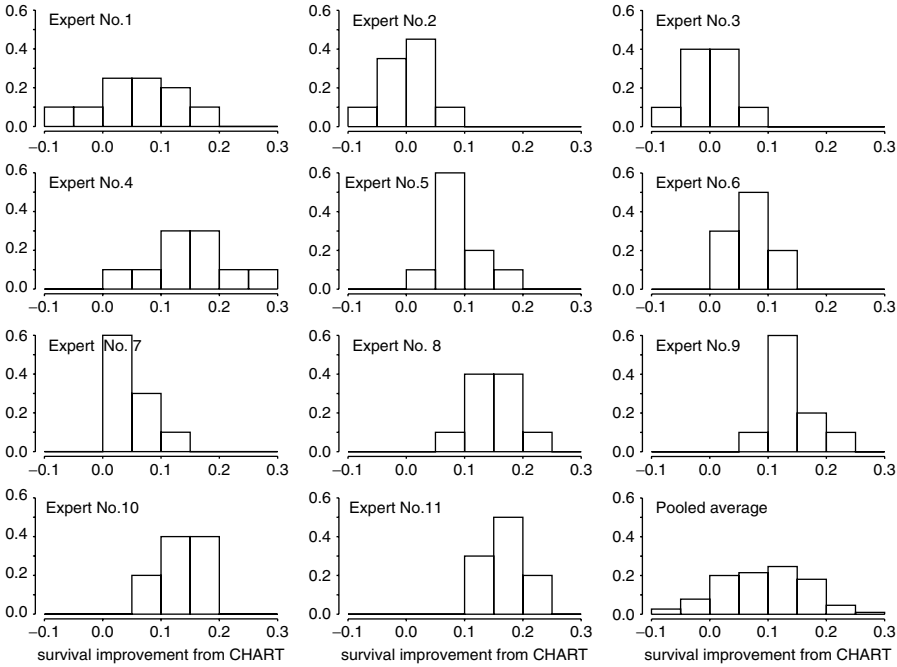

Figure 5.2 Prior opinions for lung cancer trial elicited from 11 clinical participants in the trial. The arithmetic average is used as the 'pooled' distribution.

Figure 5.2 shows the eleven lung cancer opinions as histograms. Note that subjects 7 and 11 have very different opinions and could be taken as extremes for a 'community' of priors. Here we use the arithmetic average of the distributions as a summary, since we wish to represent

an 'average' clinician. The prior distribution expressed a median anticipated 2-year survival benefit of 10%, and a 10% chance that CHART would offer no survival benefit at all. The histogram was then transformed to a log (hazard ratio) scale assuming a 15% baseline survival: for example, the 'bin' of the histogram with range 5% to 10% was transformed to one with upper limit $\log[\log(0.20)/\log(0.15)] = -0.16$ and lower limit $\log[\log(0.25)/\log(0.15)] = -0.31$. This subjective prior distribution had a mean of -0.28 and standard deviation of 0.232 (corresponding to an estimated hazard ratio of 0.76 with 95% interval from 0.48 to 1.19). A normal $N[\mu, \sigma^2/n_0]$ distribution with these characteristics was fitted, with $\mu = -0.28$, $\sigma = 2$, $\sigma/\sqrt{n_0} = 0.23$, which implies $n_0 = 74.3$. From Section 2.4.2, this prior could also be thought of as a posterior having observed a log-rank statistic ($L = O - E$) such that $4L/n_0 = -0.28$, and so $L = -5.5$. The expected E under the null hypothesis is $n_0/2 = 37.2$ and so the observed O under CHART is $37.2 - 5.5 = 31.7$. Thus the prior can be interpreted as being approximately equivalent to a balanced 'imaginary' trial in which 74 deaths had occurred (32 under CHART, 42 under standard).

For the head-and-neck trial, the fitted prior mean log(hazard ratio) is $\mu = -0.33$ with standard deviation 0.26, equivalent to $n_0 = 61.0$.

The clinical prior distributions are displayed in Figure 5.3, which shows the average transformed onto a log(hazard-ratio) scale for both lung and

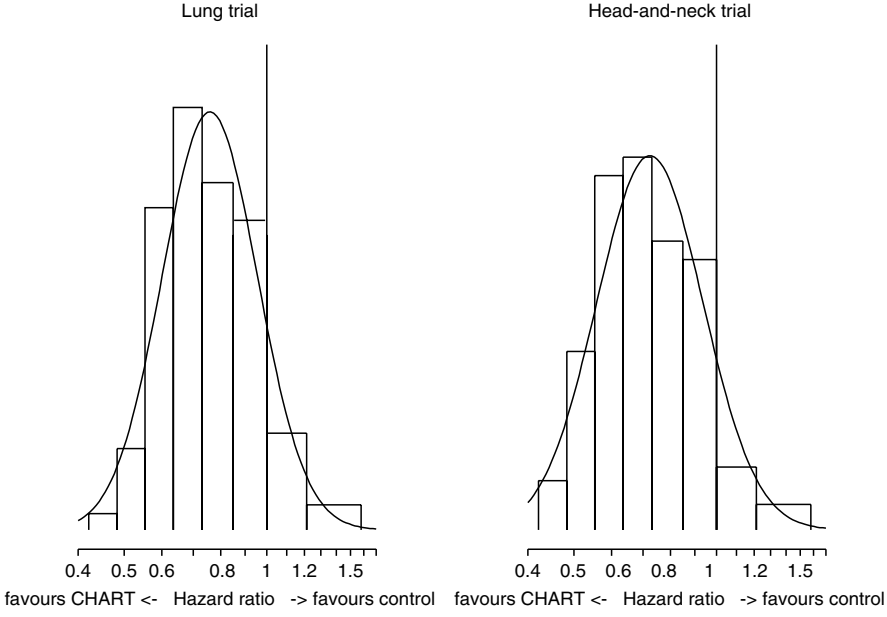

Figure 5.3 Average opinion for lung cancer and head-and-neck CHART trials with normal distributions fitted with matching mean and variance.

head-and-neck trials. The fit of the normal distribution is quite reasonable, and the similarity between the two sets of opinions is clear, each supporting around a 25% reduction in hazard, but associated with considerable uncertainty.

5.3 CRITIQUE OF PRIOR ELICITATION

There have been many criticisms of the process of eliciting subjective prior distributions in the context of health-care evaluation, and claims include the following:

1. *Subjects are biased in their opinions*. Gilbert *et al*. (1977) state that 'innovations brought to the stage of randomised trials are usually expected by the innovators to be sure winners', while the very fact that clinicians are participating in a trial is likely to suggest they expect the new therapy to be of benefit (Hughes, 1991) – we shall see that this appears to be borne out in the results to be shown in Table 5.3. Altman (1994) warns that investigators may even begin to exaggerate their prior beliefs in order to make their prospective trial appear more attractive (although we could claim this already happens both in public and industry-funded studies). Fisher (1996) believes the effort put into elicitation is misplaced, since the measured beliefs are likely to be based more on emotion than on scientific evidence.
2. *The choice of subject biases results*. The biases discussed in Section 5.2 mean that the choice of subject for elicitation is likely to influence the results. If we wish to know the distribution of opinions among well-informed clinicians, then trial investigators are not a random sample and may give biased conclusions. Fayers *et al*. (2000) provide a detailed case study in which there is clear over-optimism of investigators (see Example 6.4). Lewis (1994) says statisticians reviewing the literature may well provide much better prior distributions than clinicians, while Chalmers (1997) suggests even lay people are biased towards believing new therapies will be advances, and therefore we need empirical evidence on which to base the prior probability of superiority. Pocock (1994) states that the 'hardened sceptical trialist, the hopeful clinician and the optimistic pharmaceutical company will inevitably have grossly different priors'. An extreme view is that uncertainty as to whose prior to use militates against any use of Bayesian methods (Fisher, 1996).
3. *Timing of elicitation has an influence*. Senn (1997a) objects to any retrospective elicitation of priors as 'present remembrance of priors past is not the same as a true prior', while Hughes (1991) points out that opinions are likely to be biased by what evidence has recently been presented and by whom.

These concerns have led to a call for the evidential basis for priors to be made explicit, and for effort to go into identifying reasons for disagreement and attempting to resolve these (Fisher, 1996). Even advocates of Bayesian methods have suggested that the biases in clinical priors suggest more attention should be paid to empirical evidence from past trials, possibly represented as priors expressing a degree of scepticism concerning large effects: Fayers (1994) asks, given the long experience of negative trials, 'should we not be using priors strongly centred around 0, irrespective of initial opinions, beliefs and hopes of clinicians?'. Our view is similar: elicited priors from investigators show predictable positive bias and should be supplemented, if not replaced, by priors that are either based on evidence or reflect archetypal views of 'scepticism' or 'enthusiasm'. Taking context into account (Section 3.1) means that it is quite reasonable to allow for differing perspectives, and in many cases substantial effort in careful elicitation from representative clinicians may not be worthwhile.

5.4 SUMMARY OF EXTERNAL EVIDENCE*

If the results of previous similar studies are available, it is clear they may be used as the basis for a prior distribution. Suppose, for example, we have historical data $y_1, \ldots, y_H$ each assumed to have a normal likelihood

$$y_h \sim \mathrm{N}[\theta_h, \sigma_h^2],$$

where each of these estimates could itself be based on a pooled set of studies. Numerous options are available for specifying the relationship between $\theta_h, h = 1, \ldots, H$, and θ, the parameter of interest, and we shall expand on the list given in Section 3.16. Each option is represented graphically in Figure 5.4 using a similar convention to that in Section 3.19.3: these approaches for handling historical data are also considered when considering historical controls in randomised trials (Section 6.9), modelling the potential biases in observational studies (Section 7.3), and in pooling data from many sources in an evidence synthesis (Section 8.2).

(a) *Irrelevance*. Each θ_h is of no relevance to θ, and the prior will need to be formulated without reference to previous studies.
(b) *Exchangeable*. We might be willing to assume $\theta_h, h = 1, \ldots, H$, and θ are exchangeable so that, for example,

$$\theta_h, \theta \sim \mathrm{N}[\mu, \tau^2].$$

This leads to a direct use of a meta-analysis of many previous studies.

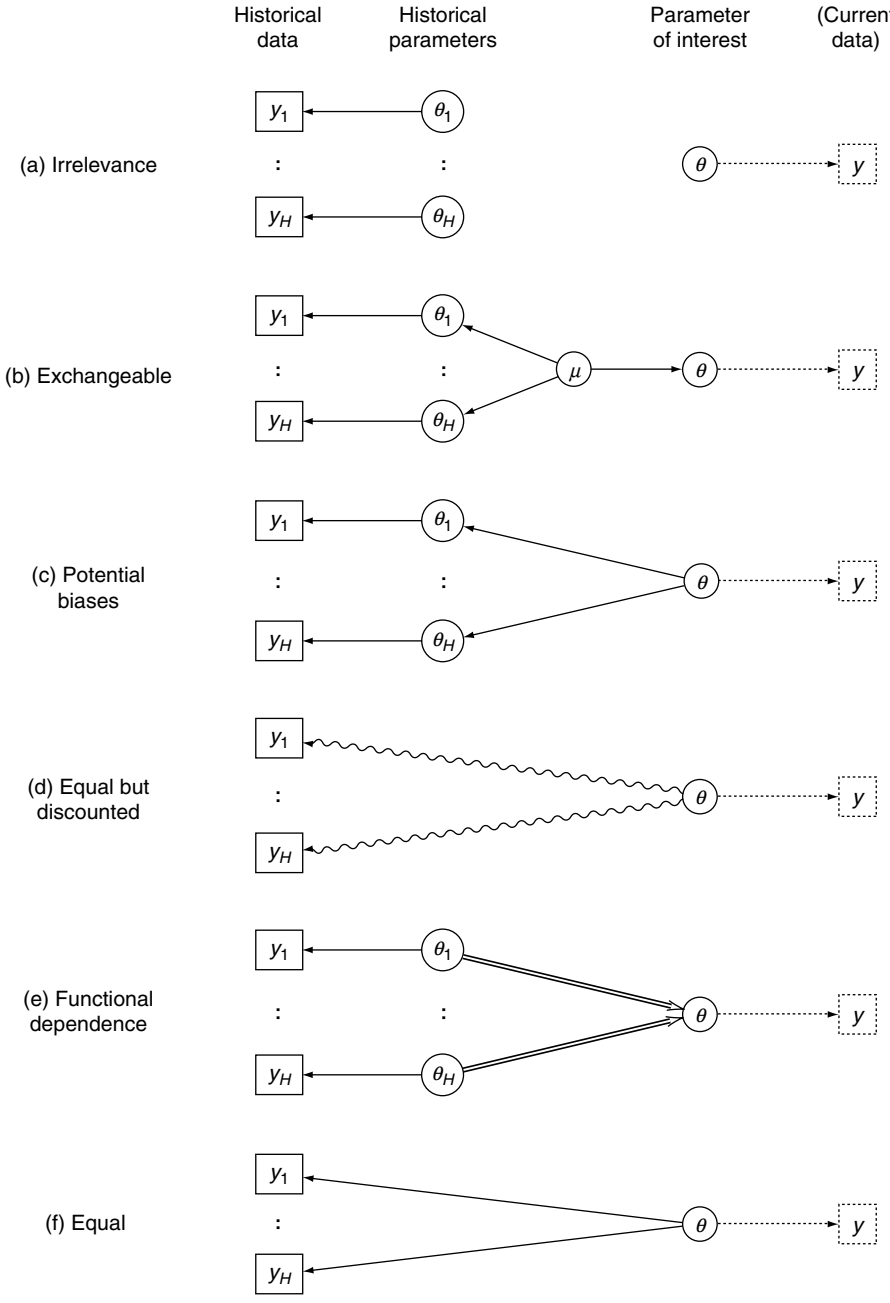

Figure 5.4 Different assumptions relating parameters underlying historical data to the parameter of current interest: single arrows represent a distribution, double arrows represent logical functions, and wavy arrows represent discounting.

It is important to note that the appropriate prior distribution for θ is the predictive distribution of the effect θ in a new study, and not the posterior distribution of the 'average' effect μ. In particular, assuming τ is known and adopting a uniform prior for μ before the historical studies, we have from Section 3.18.2 that the posterior distribution for μ given the historical studies is

$$\mu|\, y_1, \ldots, y_H \sim \mathrm{N}\left[\frac{\sum_h y_h w_h}{\sum_h w_h}, \frac{1}{\sum_h w_h}\right],$$

where $w_h = 1/(\sigma_h^2 + \tau^2)$. Hence the prior distribution for θ is

$$\theta|\, y_1, \ldots, y_H \sim \mathrm{N}\left[\frac{\sum_h y_h w_h}{\sum_h w_h}, \frac{1}{\sum_h w_h} + \tau^2\right].$$

If there is just a single historical study h, then

$$\theta|\, y_h \sim \mathrm{N}[y_h,\ 2\tau^2 + \sigma_h^2].$$

In general τ will be unknown and need to be estimated, although with few historical studies it will need to be assumed known or be given an informative prior distribution.

Exchangeability is quite a strong assumption, but if this is reasonable then it is possible to use databases to provide prior distributions (Gilbert *et al.*, 1977). Lau *et al.* (1995) point out that cumulative meta-analysis can be given a Bayesian interpretation in which the prior for each trial is obtained from the meta-analysis of preceding studies, while DerSimonian (1996) derives priors for a trial of the effectiveness of calcium supplementation in the prevention of pre-eclampsia in pregnant women by a meta-analysis of previous trials using both random-effects and fixed-effects models.

(c) *Potential biases*. We could assume that $\theta_h, h = 1, \ldots, H$, are functions of θ. A common choice is the existence of a bias δ_h so that $\theta_h = \theta + \delta_h$. Possibilities then include making the following assumptions:

1. δ_h is known.
2. δ_h has a known distribution with mean 0, say $\delta_h \sim \mathrm{N}(0, \sigma_{\delta h}^2)$, and so $\theta_h \sim \mathrm{N}(\theta, \sigma_{\delta h}^2)$. This is now almost identical to the exchangeability assumption, except that the previous study parameters are centred around the parameter of interest θ and not the population mean μ and the potential site of the bias may be study-specific. Adapting the results for the exchangeability case reveals that the posterior distribution for θ given the historical studies is

$$\theta|y_1, \ldots, y_H \sim \mathrm{N}\left[\frac{\sum_h y_h w_h^{'}}{\sum_h w_h^{'}}, \frac{1}{\sum_h w_h^{'}}\right],$$

where $w_h' = 1/(\sigma_h^2 + \sigma_{\delta h}^2)$, which follows by noting the predictive distribution $y_h \sim \mathrm{N}[\theta, \sigma_h^2 + \sigma_{\delta h}^2]$. If there is just a single historical study h, then

$$\theta | y_h \sim \mathrm{N}[y_h, \sigma_h^2 + \sigma_{\delta h}^2];$$

again, with only one historical study σ_δ^2 will need to be assumed known or have a strong prior distribution.

3. If we suspect systematic bias in one direction, we might take δ_h to have a known distribution with non-zero mean, say $\delta_h \sim \mathrm{N}[\mu_\delta, \sigma_{\delta h}^2]$. We then obtain a prior distribution, for a single historical study,

$$\theta \sim \mathrm{N}[y_h + \mu_\delta, \sigma_h^2 + \sigma_{\delta h}^2].$$

(d) *Equal but discounted*. Previous studies may not be directly related to the one in question, and we may wish to discount their influence: for example, in the context of control groups, Kass and Greenhouse (1989) state that 'we wish to use this information, but we do not wish to use it as if the historical controls were simply a previous sample from the same population as the experimental controls'. Ibrahim and Chen (2000) suggest the 'power' prior, in which we assume $\theta_h = \theta$, but discount the historical evidence by taking its likelihood $p(y_h|\theta_h)$ to a power α. For normal historical likelihoods this corresponds to adopting a prior distribution for θ, given the historical studies, of

$$\theta | y_1, \ldots, y_H \sim \mathrm{N}\left[\frac{\sum_h y_h w_h''}{\sum_h w_h''}, \frac{1}{\alpha \sum_h w_h''}\right]$$

where $w_h'' = 1/\sigma_h^2$; α varies between 0 (totally discount past evidence) to 1 (include past evidence in its totality and at 'face value'). If there is just a single historical study h, then

$$\theta | \, y_h \sim \mathrm{N}[y_h, \sigma_h^2/\alpha].$$

For example, Greenhouse and Wasserman (1995) downweight a previous trial with 176 subjects to be equivalent to only 10 subjects, and Tan *et al.* (2002) take $\alpha = 0.25$ in basing a prior on a previous phase III study; see Example 5.2 for a detailed illustration of using such a 'power' prior. We note, however, that Eddy *et al.* (1992) are very strong in their criticism of this method, claiming it has no operational interpretation and hence no means of assessing a suitable value for α.

(e) *Functional dependence*. It is possible that the parameter of interest may be logically expressed as a function of parameters from historical studies. For example, suppose θ_1 were the treatment effect in men derived from a

male-only study, and θ_2 were the treatment effect in women derived from a female-only study. Then the expected treatment effect in a study to be carried out in a population with proportion p males would be

$$\theta = p\theta_1 + (1-p)\theta_2,$$

and a prior for θ could be derived from evidence on θ_1 and θ_2.

(f) *Equal*. This assumes the past studies have all been measuring identical parameters: if θ is a property of a single patient group rather than a treatment effect, this assumption is essentially equivalent to direct pooling of the past data with those in the current study, and hence is based on the very strong assumption of exchangeability of individual patients. In our normal model we would assume $\theta_h = \theta$ and individuals are exchangeable, and so completely pool the data to obtain a prior

$$\theta|y_1, \ldots, y_H \sim \mathrm{N}\left[\frac{\sum_h y_h w''_h}{\sum_h w''_h}, \frac{1}{\sum_h w''_h}\right]$$

where $w''_h = 1/\sigma^2_h$. If there is just a single historical study h, then

$$\theta|y_h \sim \mathrm{N}[y_h, \sigma^2_h].$$

Such a strong assumption may be more acceptable if a prior is to be used in the design and not the analysis, and Brown *et al*. (1987) provide such an example using data from a pilot trial.

We note that, for the Normal model, exchangeability (b), bias (c) and discounting (d) could under certain circumstances all lead to the same prior distribution for θ, provided there is only one historical study. If there are multiple studies then these three approaches will generally all lead to different priors for θ.

Various combinations of these techniques are possible. For example, Berry and Stangl (1996a) assume a fixed probability p that each historical patient is exchangeable with those in the current study, *i.e*. either option (f) (complete pooling) with probability p, or option (a) (complete irrelevance) with probability $1-p$. Example 9.3 illustrates the combination of an exchangeable and a bias model: a past parameter θ_h is assumed to have distribution $\theta_h \sim \mathrm{N}[\mu + \delta_h, \tau^2]$, where the additional bias term has distribution $\delta_h \sim \mathrm{N}(0, \sigma^2_{\delta h})$. Hence the overall likelihood contribution from the past study is $\theta_h \sim \mathrm{N}[\mu, \tau^2 + \sigma^2_{\delta h}]$; the variance can also be expressed as τ^2/q_h, where $q_h = \tau^2/(\tau^2 + \sigma^2_{\delta h})$ can be considered as a 'quality weight' of the past study. Values of q_h near 1 mean little bias, near 0 mean substantial bias. This model formally justifies the use of 'quality-weights' in random-effects meta-analysis.

Example 5.2 *GUSTO: Using previous results as a basis for prior opinion*

References: Brophy and Joseph (1995), Fryback *et al*. (2001b), Harrell and Shih (2001), Brophy and Joseph (2000) and Ibrahim and Chen (2000).

Intervention: Streptokinase (SK) compared to tissue plasminogen activator (tPA) to dissolve clots in occluded coronary arteries following a myocardial infarction. tPA is considerably more expensive than SK.

Aim of study: Two previous trials of SK versus tPA (GISSI-2 and ISIS-3) showed minimal difference, although the stroke rate was consistently higher under tPA.

Study design: Parallel-group unblinded RCT, with two SK arms with different administrations of heparin (later pooled), tPA arm and an arm with both SK and tPA (ignored in this analysis).

Outcome measure: Odds ratio (OR) of stroke and/or death, with OR < 1 favouring tPA.

Planned sample size: The sample size of the GUSTO trial was calculated on the basis of having 80% power to detect a 15% relative reduction in the risk of death or a 1% absolute decrease at the 5% significance level.

Statistical model: A normal likelihood was assumed based on the estimated log(odds ratio) (Section 2.4.1); σ has been taken as 2.

Prospective analysis?: No.

Prior distribution: It is natural to base, to some extent, a prior distribution on the two preceding trials, whose results are shown in Table 5.1, using data presented by Brophy and Joseph (1995). Taking the previous trials at full weight, the pooled previous trials give rise to a prior for GUSTO with mean 0.0002 and standard deviation $\sigma/\sqrt{4604} = 0.03$: a very sceptical prior indeed, with a 95% interval for the OR from 0.94 to 1.06.

Table 5.1 Historical and observed data for GUSTO study. The ms are the 'effective number of events' in a balanced trial, obtained from setting the estimated variances of the log(odds ratios) to σ^2/m: the ms do not exactly match the actual number of events, particularly in GUSTO, due to imbalance in allocation. The 'pooled' results are obtained by adding the ms and weighting the log(odds ratios) by their respective ms: this pooled m can be relabelled n_0 if it is used as the basis for a prior distribution for GUSTO.

Trial	SK events/cases	%	tPA events/cases	%	OR	log(OR)	m (when $\sigma = 2$)
GISSI-2	985/10 396	9.5%	1067/10 372	10.3%	1.09	0.09	1847
ISIS-3	1596/13 780	11.6%	1513/13 746	11.0%	0.94	−0.06	2757
Pooled						0.0002	$n_0 = 4604$
GUSTO	1574/20 173	7.8%	714/10 343	6.9%	0.88	−0.13	1825

However, Brophy and Joseph (2000) emphasise important differences between the studies: the GUSTO study featured an 'accelerated' tPA protocol, more aggressive use of intravenous heparin, increased revascularisation in the tPA arm, and possible increased tPA benefit in US patients. This suggests downweighting the prior evidence in some way, and different authors have subsequently used almost all the approaches outlined in Section 5.4. We shall focus on simple discounting (method (d)), but other methods are mentioned under 'Comments'. Brophy and Joseph (1995) 'discounted' the previous trials, essentially implementing the power prior distributions of Ibrahim and Chen (2000), which is equivalent to adjusting the prior 'number of events' from n_0 to αn_0. They considered α to be 0, 0.1, 0.5 and 1.0, equivalent to taking the prior 'number of events' to be 0, 460.4, 2302 and 4604. Taking $\alpha = 0$ is equivalent to treating the previous trials as irrelevant (option (a)) and hence selecting a uniform prior on the log(odds ratio), while taking $\alpha = 1$ is equivalent to assuming the trials are measuring equal parameters (option (f)) – note that this is not equivalent to pooling the patients on each arm, but is equivalent to pooling the estimated treatment effects.

Loss function or demands: The GUSTO trial was designed around a 15% reduction in mortality, so we might take an odds ratio of 0.85 to reflect a clinically important difference.

Computation/software: Conjugate normal model.

Evidence from study: This is provided in Table 5.1. The standardised test statistic based on the data alone is $z_m = y_m\sqrt{m}/\sigma = -0.13\sqrt{1825}/2 = -2.78$, providing a two-sided *P*-value of 0.005.

Bayesian interpretation: Figure 5.5 shows plots of prior, likelihood and posterior under different assumptions concerning α, superimposed on a clinically important difference of 0.85. The probability that tPA is inferior to SK is very low unless the prior trials are considered at almost full weight. However, it is clear that although GUSTO may show 'statistical significance' in that the posterior probability that OR < 1 is high, there is not strong evidence of 'practical significance', in that the posterior probability that OR < 0.85 is moderate even when the prior evidence is totally ignored.

Sensitivity analysis: Figure 5.6 shows changing conclusions as α ranges from 0 (ignore historical evidence) to 1 (completely pool with historical evidence). This clearly shows evidence for benefit unless the past data are quite strongly weighted, but even slight inclusion of past data serves to exclude a clinically important difference of 15%.

Comments: We can fit previous approaches to this problem within the structure outlined in Section 5.4.

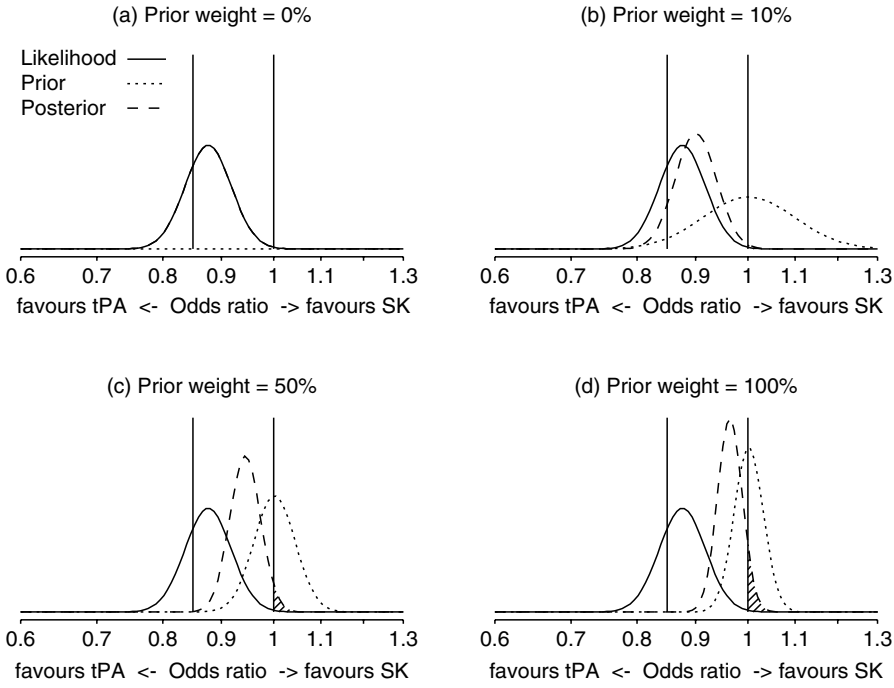

Figure 5.5 Posterior estimate of the odds ratio for the GUSTO trial under different prior assumptions: weighting the previous trial results by a factor (a) 0% (*i.e*. the reference prior in which the posterior is proportional to the likelihood), (b) 10%, (c) 50% and (d) 100% (*i.e*. full pooling with the past data). The shaded area represents the posterior probability that OR > 1 and hence favours SK, and is very low unless very high weight is given to the previous trials. However, the chance of an odds ratio less than 0.85 is only moderate even when using the trial data alone, and drops severely for even 10% weighting of the past trial data.

(a) *Irrelevance*. Harrell and Shih (2001) consider that the previous trials are entirely irrelevant to GUSTO due to the revised tPA protocol, and so only consider a 'reference' and 'sceptical' prior (Section 5.5): the reference prior is uniform on the log(OR) scale and hence the posterior distribution is the same shape as the likelihood, while the sceptical prior was centred on the null hypothesis of OR $= 1$, and expressed 95% belief that the true OR lay within the bounds 0.75–1.33, *i.e*. it is unlikely that there is more than a 25% relative change between the treatments: this prior is even more diffuse than that shown in Figure 5.5(b).

(b) *Exchangeable*. One of the models considered by Brophy and Joseph (2000) assumes the treatment effects in the three trials are exchangeable, and places a normal population distribution on the three log(odds ratios) – they use 'diffuse' priors on the parameters of mean and variance of the normal population. However, both the exchangeability

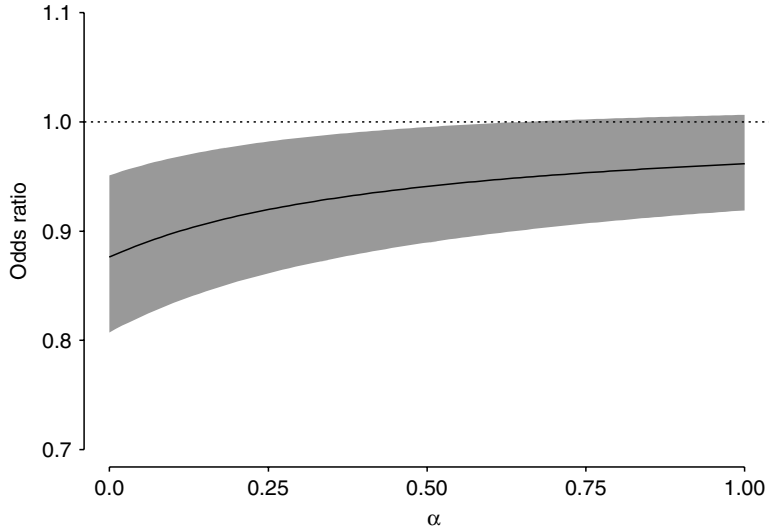

Figure 5.6 Posterior estimate of the odds ratio for the GUSTO trial downweighting previous trial results by varying amounts ($\alpha = 0$ implies total discounting, whilst $\alpha = 1$ implies acceptance of previous evidence at 'face-value').

assumption, and the attempt to estimate population parameters from just three trials (regardless of their size), make this prior formulation somewhat doubtful.

(c) *Potential biases*. Acknowledging the possible systematic differences between the trials, Brophy and Joseph (2000) also consider two possible sources of bias: differences in revascularisation rates in GUSTO, and differences in tPA administration between GUSTO and the previous trials. These are applied to the hierarchical model described under (b).

(d) *Equal but discounted*. In a different application of the discounting approach, Fryback *et al*. (2001b) suggests the SK arm in GUSTO is reasonably compatible with the SK arm in previous trials, and so adopt $\alpha_C = 1/3$ for SK. However, they severely discount the tPA arm from a sample size of around 24 000 to one of 50, so that $\alpha_T \approx 1/500$ for tPA.

Now $\mathrm{V}(\log(\mathrm{OR})) = \mathrm{V}(\log O_C) + \mathrm{V}(\log O_T)$, where O_C, O_T are the odds on death under SK and tPA, respectively. With no discounting, $\mathrm{V}(\log O_C) \approx \mathrm{V}(\log O_T) = V$. With differential discounting,

$$\mathrm{V}(\log(\mathrm{OR})) = \frac{\mathrm{V}(\log O_C)}{\alpha_C} + \frac{\mathrm{V}(\log O_T)}{\alpha_T} \approx \mathrm{V}\left(\frac{1}{\alpha_C} + \frac{1}{\alpha_T}\right).$$

Thus the overall discount factor, relative to the undiscounted variance of $2V$, is $\alpha = 2/(\alpha_C^{-1} + \alpha_T^{-1})$ which is the 'harmonic mean' of the individual discounts. Fryback *et al*.'s assumptions therefore lead to an overall discount factor of $2/(3 + 500) \approx 1/250$, which means the prior will have little impact on the likelihood.

(f) *Equal*. As an extreme of the discounting procedure, if we assume $\alpha = 1$ we are led to completely pool the results of the three trials.

5.5 DEFAULT PRIORS

It would clearly be attractive to have prior distributions that could be taken 'off the shelf', rather than having to consider all available evidence external to the study in their construction: such priors can, at a minimum, be considered as 'baselines' against which to measure the impact of past evidence or subjective opinion. Four main suggestions can be identified.

5.5.1 'Non-informative' or 'reference' priors

There has been a huge volume of research into so-called *non-informative* or *reference* priors, that are intended to provide a kind of default or 'objective' Bayesian analysis free from subjectivity. Kass and Wasserman (1996) review the literature, but emphasise the continuing difficulties in defining what is meant by 'non-informative', and the lack of agreed reference priors in all but simple situations.

In many situations we might adopt a uniform distribution over the range of interest, possibly on a suitably transformed scale of the parameter (Box and Tiao, 1973). Formally, a uniform distribution means the posterior distribution has the same shape as the likelihood function, which in turn means that the resulting Bayesian intervals and estimates will essentially match the traditional results. Results with reference priors are generally quoted as one part of a Bayesian analysis, and may even form the main basis for inferences. For example, Burton (1994) suggests that most doctors interpret frequentist confidence intervals as credible intervals, and also that information external to a study tends to be vague, and that therefore results from a study should be presented by performing a Bayesian analysis with a non-informative prior and quoting posterior probabilities for the parameter of interest being in various regions. The fact that a reference prior may produce essentially identical conclusions to a classical analysis, and yet allow more flexible and intuitive presentations, has led to the use of what are essentially Bayesian methods but under names such as 'confidence levels' (Shakespeare *et al*., 2001).

Invariance arguments may be used as a basis for reference priors (Jeffreys, 1961): for example, if we feel a reference prior on an odds ratio OR should be the same whichever treatment is taken in the numerator of the odds ratio, then it means that the same prior should hold for OR and 1/OR, which means that we must be uniform on the log(OR) scale. Similar arguments can be used to justify a uniform prior on $\log(\sigma^2)$ for a sampling variance σ^2, since this prior is also equivalent to a uniform prior on $\log(\sigma)$ (or indeed any power of σ), and hence is invariant to whether one is working on the standard deviation or variance scale. This prior is equivalent to assuming $p(\sigma^2) \propto \sigma^{-2}$, or $p(\sigma) \propto \sigma^{-1}$. A standard result (DeGroot, 1970; Lee, 1997) is that, for normal likelihoods, this prior, combined with an independent uniform prior on the mean, gives rise to the familiar classical tail areas based on a t distribution.

The real problem with 'uniform' priors is that they are no longer uniform if the parameter is transformed, which is well illustrated by the problem of assigning a reference prior to the probability θ of an event. The classic solution, dating back to Bayes and Laplace in the eighteenth century, is to give a uniform prior for θ, equivalent to a Beta[1,1]. From the beta-binomial distribution (Section 3.13.2) we can show this leads to a uniform distribution over the number $0, 1, \ldots, n$ of occurrences in n Bernoulli trials, which might seem a reasonable justification for its claim to be 'non-informative'. However in many of our examples we place a uniform distribution over a log(odds) scale, *i.e.* $\log[p/(1-p)]$ has a uniform distribution. It can be shown that this is equivalent to a Beta[0,0] distribution for p – an improper distribution that strongly favours values of p near 0 or 1. As an intermediate suggestion, invariance arguments (Box and Tiao, 1973) have led to the use of a Beta[0.5,0.5] prior, which is proper but still favours extreme values of p (Section 2.6.3). Of course, all these priors will give essentially the same result with a large enough set of data, but could have some influence with rare events. Even when one has chosen a suitable scale for a uniform prior, it may be inappropriate to term it 'non-informative': Fisher (1996) points out that 'there is no such thing as a ''noninformative'' prior. Even improper priors give information: all possible values are equally likely'. There is a particular difficulty in assigning such a 'reference' prior to random-effect variances in hierarchical models, and we shall consider this issue in Section 5.7.

5.5.2 'Sceptical' priors

Informative priors that express scepticism about large treatment effects have been put forward both as a reasonable expression of doubt, and as a way of controlling early stopping of trials on the basis of fortuitously positive results (Section 6.6.2). Kass and Greenhouse (1989) suggest that a 'cautious reasonable sceptic will recommend action only on the basis of fairly firm knowledge', but that these sceptical 'beliefs we specify need not be our own, nor need they be

the beliefs of any actual person we happen to know, nor derived in some way from any group of ''experts'' '.

Mathematically speaking, a sceptical prior about a treatment effect will have a mean of zero and a shape chosen to include plausible treatment differences which determine the degree of scepticism. Spiegelhalter *et al.* (1994) argue that a reasonable degree of scepticism may be feeling that the trial has been designed around an alternative hypothesis that is *optimistic*, formalised by a prior with only a small probability γ (say, 5%) that the treatment effect is as large as the alternative hypothesis θ_A (see Figure 5.7).

Assuming a prior distribution $\theta \sim \mathrm{N}[0,\ \sigma^2/n_0]$ and such that $p(\theta > \theta_A)$ is a small value γ implies $\gamma = 1 - \Phi(\theta_A\sqrt{n_0}/\sigma)$ and so

$$-\sigma\frac{z_\gamma}{\sqrt{n_0}} = \theta_A, \tag{5.1}$$

where $\Phi(z_\gamma) = \gamma$. Now suppose the trial has been designed with size α and power $1 - \beta$ to detect an alternative hypothesis θ_A. Then we have the standard relation (2.38)

$$\sigma^2\frac{(z_{\alpha/2} + z_\beta)^2}{\theta_A^2} = n \tag{5.2}$$

between the proposed sample size n and θ_A. Equating θ_A in (5.1) and (5.2) gives

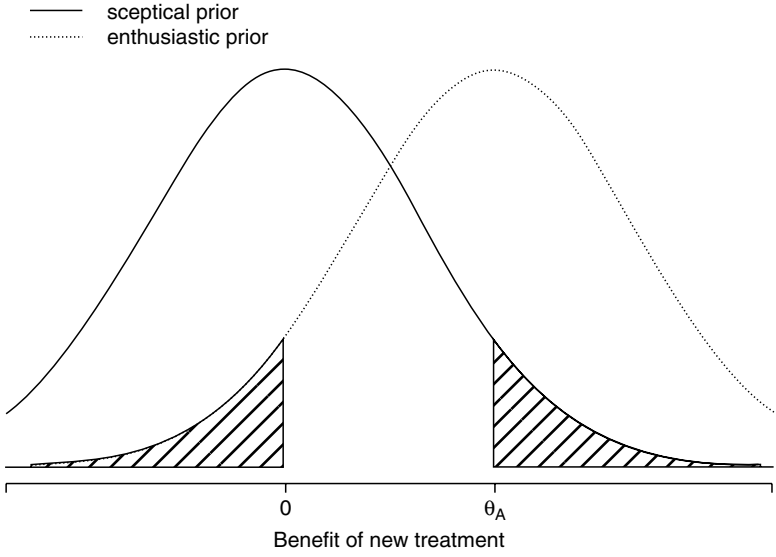

Figure 5.7 Sceptical and enthusiastic priors for a trial with alternative hypothesis θ_A. The sceptics' probability that the true difference is greater than θ_A is γ (shown shaded). This value has also been chosen for the enthusiasts' probability that the true difference is less than 0.

$$\frac{n_0}{n} = \left[\frac{z_\gamma}{z_{\alpha/2} + z_\beta}\right]^2.$$

Reasonable values might be $\alpha = 0.05$, $\beta = 0.1$ and $\gamma = 0.05$, which gives $n_0/n = 0.257$.

Thus in a trial designed with 5% size and 90% power, such a sceptical prior corresponds to adding a 'handicap' equivalent to already having run a 'pseudo-trial' with no observed treatment difference, and which contains around 26% of the proposed sample size.

This approach has been used in a number of case studies (Freedman *et al.*, 1994; Parmar *et al.*, 1994) and has been suggested as a basis for monitoring trials (Section 6.6) and when considering whether or not a confirmatory study is justified (Section 6.7). Other applications of sceptical priors include Fletcher *et al.* (1993), DerSimonian (1996), and Heitjan (1997) in the context of phase II studies, while a senior FDA biostatistician (O'Neill, 1994) has stated that he 'would like to see [sceptical priors] applied in more routine fashion to provide insight into our decision making'.

Example 5.3 *CHART (continued): Sceptical priors*

References: Parmar *et al*. (1994, 2001) and Spiegelhalter *et al*. (1994).

Prior distribution: A *sceptical prior* was derived using the ideas in Section 5.5.2: the prior mean is 0 and the precision is such that the prior probability that the true benefit exceeds the alternative hypothesis is low (5% in this case). Thus a prior with mean 0 and standard deviation $\sigma/\sqrt{n_0}$ will show a 5% chance of being less than δ_A if $n_0 = (1.65\sigma/\theta_A)^2$ by (5.1). For the lung trial, the alternative hypothesis on the log(hazard ratio) scale is $\theta_A = \log(0.73) = -0.31$. Assuming $\sigma = 2$ gives $n_0 = 110$. For the head-and-neck trial, the alternative hypothesis is $\theta_A = \log(0.64) = -0.45$, which gives a sceptical prior with $n_0 = 54$.

The sceptical prior distributions are displayed in Figure 5.8, with the clinical priors derived in Example 5.1.

5.5.3 'Enthusiastic' priors

As a counterbalance to the pessimism expressed by the sceptical prior, Spiegelhalter *et al.* (1994) suggest an 'enthusiastic' prior centred on the alternative hypothesis and with a low chance (say, 5%) that the true treatment benefit is negative. Use of such a prior has been reported in case studies (Freedman *et al.*, 1994; Heitjan, 1997; Vail *et al.*, 2001; Tan *et al.*, 2002) and as a basis for conservatism in the face of early negative results (Fayers *et al.*, 1997); see

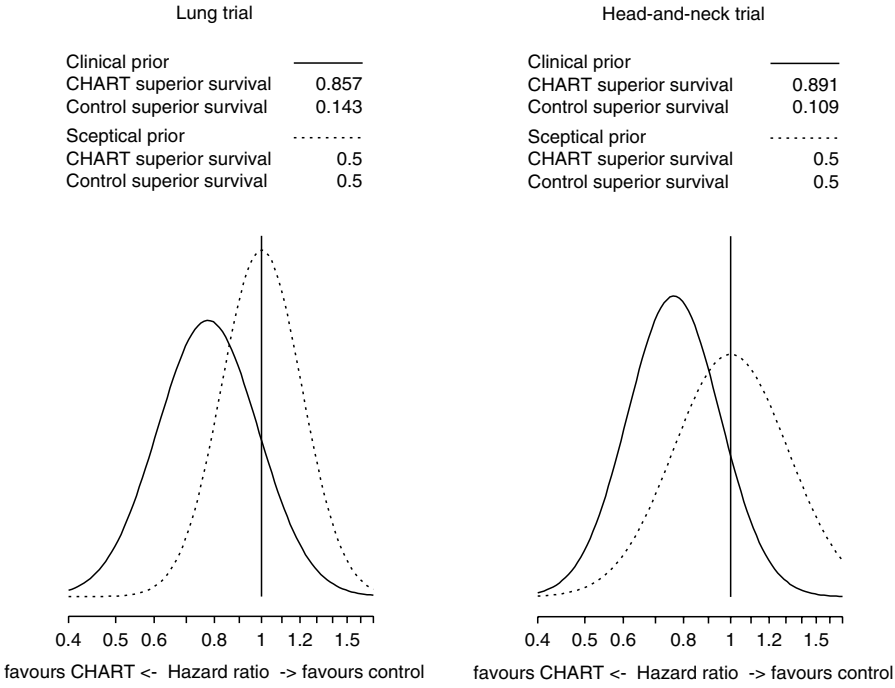

Figure 5.8 Sceptical and clinical priors for both lung and head-and-neck CHART trials, showing prior probabilities that CHART has superior survival. The sceptical priors express a 5% prior probability that the true benefit will be more extreme than the alternative hypotheses of HR = 0.73 for the lung trial and HR = 0.64 for the head-and-neck trial.

Section 6.6.2. Dignam *et al.* (1998) provide an example of such a prior but call it 'optimistic' (Example 6.7). Such a prior is intended to represent the opinion of an archetypal enthusiast and does not represent the opinion of an identifiable individual.

Other options for default priors are possible: for example, Cronin *et al.* (1999) adopt an 'indifference' prior that lies half-way between 'sceptical' and 'enthusiastic'.

5.5.4 Priors with a point mass at the null hypothesis ('lump-and-smear' priors)*

The traditional statistical approach expresses a qualitative distinction between the role of a null hypothesis, generally of no treatment effect, and alternative hypotheses. A prior distribution that retains this distinction would place a 'lump' of probability on the null hypothesis, and 'smear' the remaining probability over the whole range of alternatives; for example Cornfield (1969) uses a

normal distribution centred on the null hypothesis, while Hughes (1993) uses a uniform prior over a suitably restricted range. The resulting posterior distribution retains this structure, giving rise to a posterior probability of the truth of the null hypothesis; this is apparently analogous to a *P*-value but is neither numerically nor conceptually equivalent.

A specific assumption used in our examples is the following:

$$H_0 : \theta = \theta_0 \quad \text{with probability } p,$$
$$H_A : \theta \sim \mathrm{N}\left[\theta_0, \frac{\sigma^2}{n_0}\right] \text{ with probability } 1-p,$$

where we label the 'lump' and the 'smear' as null and alternative hypotheses, respectively.

Cornfield repeatedly argued for this approach, which naturally gives rise to the 'relative betting odds' or Bayes factor (Section 3.3) as a sequential monitoring tool, defined as the ratio of the likelihood of the data under the null hypothesis to the average likelihood (with respect to the prior) under the alternative. If we assume a normal likelihood $y_m \sim \mathrm{N}[\theta, \sigma^2/m]$, then we have shown in Section 4.4.3 that the Bayes factor is

$$\mathrm{BF} = \frac{p(y_m|H_0)}{p(y_m|H_A)} = \sqrt{1+\frac{m}{n_0}}\ \exp\left[\frac{-z_m^2}{2(1+n_0/m)}\right]. \tag{5.3}$$

Since

$$\frac{p(H_0|y_m)}{p(H_A|y_m)} = \mathrm{BF}\,\frac{p}{1-p},$$

we can obtain the posterior probability $p(H_0|y_m)$.

The relative betting odds are independent of the 'lump' of prior probability placed on the null (while depending on the shape of the 'smear' over the alternatives), and do not suffer from the problem of 'sampling to a foregone conclusion' (Section 6.6.5). Cornfield suggests a 'default' prior under the alternative as a normal distribution centred on the null hypothesis and with expectation (conditional on the effect being positive) equal to the alternative hypothesis θ_A. Then from the properties of the half-normal distribution (Section 2.6.7) it follows that

$$\mathrm{E}(\theta|\theta>0) = \sqrt{\frac{2\sigma^2}{\pi n_0}}. \tag{5.4}$$

Equating this to θ_A leads to assuming a prior standard deviation under the alternative hypothesis of $\sqrt{\pi/2}\theta_A$. This is similar to the formulation of a

sceptical prior described in Section 5.5.2, but with probability of exceeding the alternative hypothesis of $\gamma = \Phi(-\sqrt{2/\pi}) = 0.21$ – this is larger than the value of 5% often used for sceptical priors, but the lump of probability on the null hypothesis is already expressing considerable scepticism. Values for these prior distributions for 11 outcome measures are reported for the Urokinase Pulmonary Embolism Trial (Sasahara *et al*., 1973, p. 27), and Example 5.4 considers one of these outcomes. This method was used in a number of major studies alongside more standard approaches (Coronary Drug Project Research Group, 1970; University Group Diabetes Program, 1970), although relative betting odds were later dropped from the analysis (Coronary Drug Project Research Group, 1975). A mass of probability on the null hypothesis has also been used in a cancer trial (Freedman and Spiegelhalter, 1992) and for sensitivity analysis in trial reporting (Hughes, 1993).

Although such an analysis provides an explicit probability that the null hypothesis is true, and so appears to answer a question of interest, the prior might be somewhat more realistic were the lump to be placed on a small range of values representing the more plausible null hypothesis of 'no clinically effective difference'. Lachin (1981) has extended the approach to this situation where the null hypothesis forms an interval, although Cornfield (1969) points out that the 'lump' is in any case just a mathematical approximation to such a prior.

Example 5.4 *Urokinase: 'lump and smear' prior distributions*

Reference: Sasahara *et al*. (1973).

Intervention: Urokinase treatment for pulmonary embolism.

Aim of study: To compare thrombolytic capability in urokinase (new) with heparin (standard).

Study design: RCT entering 160 patients between 1968 and 1970. There was no prespecified sample size or stopping rule, although data were examined four times yearly by an advisory committee but not released to the investigators.

Outcome measure: Eleven endpoints based on continuous measures from angiograms, lung scans and haemodynamics.

Statistical model: Normal likelihoods assumed for an estimate y_m of treatment effect θ based on m pairs of randomised patients.

Prospective analysis?: Yes, the prior elicitations were conducted before the start of the trials, and the Bayesian results presented to the advisory committee at each of their meetings.

Prior distribution: A 'lump-and-smear' prior was assessed for each outcome (Section 5.5.4). To select n_0, Cornfield (1969) suggests setting the expectation, given there is a positive effect, to the alternative hypothesis, so from (5.4) the prior standard deviation $\sigma/\sqrt{n_0}$ is $\sqrt{\pi/2}\theta_A$, and hence $n_0 = 2\sigma^2/(\pi\theta_A^2)$. Alternative hypotheses were assessed by members of the advisory committee 'based on what appeared reasonable from previous experience with thrombolytics'.

For the outcome 'Absolute improvement in resolution on lung scan', we take σ to be the value observed in the study, 9.35 (see below). The alternative hypothesis was selected to be $\theta = 8$, slightly less than a 1 standard deviation effect, giving rise to $n_0 = 0.87$. Thus the prior under the alternative hypothesis is approximately equivalent to having observed a single pair of patients, each with the same response. This is a weak prior, but remarkably corresponds almost precisely to that recommended in recent theoretical work on Bayes factors (Kass and Wasserman, 1995); see Section 4.4.3.

Loss function or demands: None specified.

Computation/software: Conjugate normal analysis.

Evidence from study: For 'Absolute improvement in resolution on 24-hour lung scan', outcomes were available on 72 patients treated with urokinase and 70 with heparin. The difference in mean responses was $y_m = 3.61$, with standard error 1.11. Assuming $m = 71$ pairs, we have $\sigma = 1.11\sqrt{m} = 9.35$, as mentioned above. Using (5.3) the 'relative betting odds' (Bayes factor) can be calculated to be 0.052 – from Table 3.2 this corresponds to 'strong' evidence against the null hypothesis. Setting $p = 0.5$ to represent equal prior belief in the null and alternative hypotheses, this leads to a probability $0.052/(1 + 0.052) = 0.049$ that the null hypothesis is true.

Bayesian interpretation: Figure 5.9 shows the size of the 'lump' dropping dramatically from its prior level. The result is highly significant classically: $z = 3.61/1.11 = 3.25$, with a two-sided *P*-value of 0.001; Sasahara *et al*. (1973) report that due to many outcome measures and sequential analysis, only $z > 3$ would be taken as 'significant'. Note that the Bayesian posterior on the null is only 0.047, and so is not as extreme as the *P*-value (Section 4.4.3).

Comments: In this application, $m/n_0 = 71/0.87 = 82$; Figure 4.2 shows that for such results with a classical two-sided *P*-value of 0.001, the Bayes factor only provides 'strong' evidence against the null hypothesis. The prior drawn in Figure 5.9(a) provides a clue as to the difference between the two approaches: although the data observed are unlikely under the null hypothesis, the prior under the alternative is so diffuse

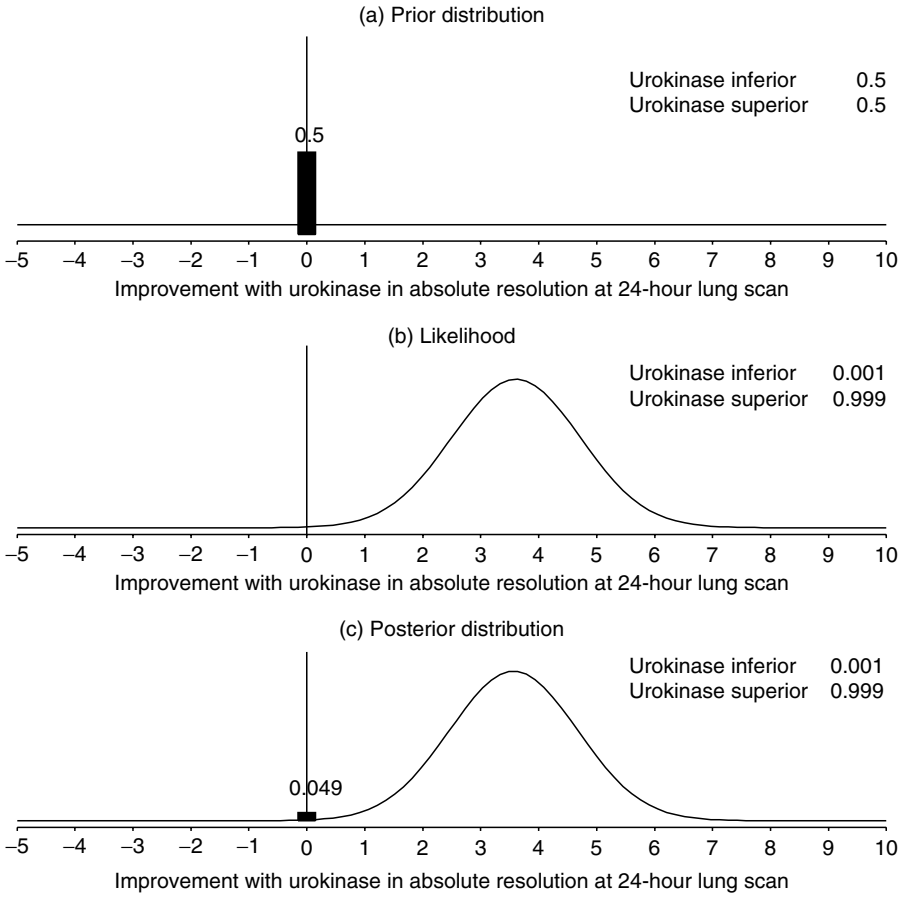

Figure 5.9 Results from the Urokinase trial analysed by Cornfield using 'relative betting odds' (Bayes factors). Data which are classically 'highly significant' ($z = 3.25$, two-sided P-value 0.001) only provide 'strong' evidence against the null hypothesis (Bayes factor $\approx 1/20$).

that it gives little weight to the parameter values suggested by the data. Hence the data are not strongly supported by either hypothesis, although the alternative receives the benefit of the doubt.

5.6 SENSITIVITY ANALYSIS AND 'ROBUST' PRIORS

An integral part of any good statistical report is a sensitivity analysis of assumptions concerning the form of the model (the likelihood). Bayesian approaches

have the additional concern of sensitivity to the prior distribution, both in view of its controversial nature and because it is by definition a subjective assumption that is open to valid disagreement. We reiterate that this fits naturally into the idea of a 'community of priors' (Kass and Greenhouse, 1989).

A natural development when carrying out a Bayesian post-hoc analysis, rather than a full Bayesian pre-study design, is to avoid all prespecification of priors and simply report the impact of the data on a suitable range of opinion: O'Rourke (1996) emphasises that posterior probabilities 'should be clearly and primarily stressed as being a ''function'' of the prior probabilities and not *the* probability of treatment effects'. We can therefore take the following steps after having observed the data:

1. Select a suitably flexible class of priors.
2. Examine how the conclusions depend on the choice of prior.
3. Identify the subsets of priors that, if seriously held, would lead to posterior conclusions of specific interest (say, the clinical superiority of an intervention).
4. Report the results and hence allow the audience to judge whether their own prior lies in the identified 'critical' subsets.

This is known as the 'robust' approach, and is also known as 'prior partitioning' (Carlin and Sargent, 1996; Sargent and Carlin, 1996). See Section 6.6.2 for further discussion of this approach to monitoring clinical trials.

Three increasingly complex 'communities' of priors have been considered:

1. *Discrete set*. Many case studies carry out analysis of sensitivity to a limited list of possible priors, possibly embodying scepticism, enthusiasm, clinical opinion and 'ignorance'; see, for example, Examples 6.6 and 6.7. It is also possible to consider sensitivity to the opinions of multiple experts, perhaps summarised by their extremes of opinion (Section 5.2.3).
2. *Parametric family*. If the community of priors can be described by one varying parameter, then it is possible to graphically display the dependence of the main conclusion on that parameter. Hughes (1991) suggested examining sensitivity of conclusions to priors based on previous trial results and that reflecting investigators' opinions, and later Hughes (1993) gives an example which features a point-mass prior on zero, and an explicit plot of the posterior probability against the prior probability of this null hypothesis. Example 5.2 carries out a similar analysis in which the 'discount' parameter is continuously varied, and the 'credibility' analysis described in Section 3.11 provides such a tool for the class of normal sceptical priors.
3. *Non-parametric family*. The 'robust' Bayesian approach has been further explored by allowing the community of priors to be a non-parametric family in the neighbourhood of an initial prior. For example, Gustafson (1989), considers the ECMO study (Example 6.9) with a community centred around a 'non-informative' prior but 20% 'contaminated' with a prior with minimal

restrictions, such as being unimodal. The maximum and minimum posterior probability of the treatment's superiority within such a class can be plotted, providing a sensitivity analysis. A similar approach has also been taken by Greenhouse and Wasserman (1995) and Carlin and Sargent (1996).

One should, however, beware of carrying out too restricted a sensitivity analysis. Stangl and Berry (1998) emphasise the need for a fairly broad community, taking into account not just the spread of the prior but also its location. They also stress that sensitivity to exchangeability and independence assumptions should be examined and that, while sensitivity analysis is important, it should not serve as a substitute for careful thought about the form of the prior distribution.

There is limited experience of reporting such analyses in the medical literature, and it has been suggested (Koch, 1991; Hughes, 1991; Spiegelhalter *et al.*, 1994) that a separate 'interpretation' section is required to display how the data in a study would add to a range of currently held opinions (Section 3.21). It would be attractive for people to be able to carry out their own sensitivity analysis of their own prior opinion; Lehmann and Goodman (2000) describe a computing architecture for this, and available software and web pages are described in Section A.2.

5.7 HIERARCHICAL PRIORS

The essence of hierarchical models was summarised in Section 3.17: by assuming that multiple parameters of interest are drawn from some common prior distribution, *i.e.* they are exchangeable, we can 'borrow strength' between multiple substudies and improve the precision for each parameter. These models form an essential component of much of Bayesian analysis, but their added power does not come without cost. The three essential assumptions are: exchangeability of parameters θ_k, a form for the random-effects distribution of the θ_k, and a 'hyperprior' distribution for the parameters of the random-effects distribution of the θ_k. All these assumptions can be important, and none can be made lightly.

5.7.1 The judgement of exchangeability

An assumption of exchangeability underlies any random-effects analysis, whether Bayesian or classical. Nevertheless, Tukey (1977) says that 'to treat the true improvements for the classes concerned as a sample from a nicely behaved population . . . does not seem to me to be near enough the real world to be a satisfactory and trustworthy basis for the careful assessment of strength of evidence'. But, as noted in Section 3.4, there does not need to be any actual

population from which units are sampled, and the very fact that we are carrying out simultaneous analysis on a number of units suggests some relationship between them. In addition, if there are known reasons to suspect that specific units are systematically different, then those reasons might be modelled by including relevant covariates and then the residual variability more plausibly reflects exchangeability; for example, Dixon and Simon (1991) discuss the reasonableness of exchangeability assumptions in the context of subset analysis (Section 6.8.1), and observe that any subsets of prior interest should be considered separately.

5.7.2 The form for the random-effects distribution

This is generally taken to be normal until evidence shows otherwise: if there is no reason to suspect systematic difference between units, a central limit theorem argument could be used to justify normality as arising from the sum of many small unobserved differences between units. Normality is computationally helpful, although with the advent of MCMC methods it has less importance, and 'heavier-tailed' distributions such as the Student's t can be adopted (Smith *et al.*, 1995).

Unlike other prior assumptions, the form of the random-effects distribution can be empirically checked from the data, although strategies for this are outside the scope of this book; see, for example, Lange and Ryan (1989), Christiansen and Morris (1996) and Hardy and Thompson (1998).

5.7.3 The prior for the standard deviation of the random effects*

In a hierarchical model $\theta \sim \mathrm{N}[\mu,\tau^2]$, the random-effects standard deviation τ plays an important role, and its value can be very influential in assessing the uncertainty concerning μ or in predicting future θs. However, there may be limited information in the data to provide a precise estimate of τ due either to there being few units, or to each unit providing little information, or both. This can make the prior for τ particularly important, and yet neither is there any generally accepted reference prior for τ, nor are there formally established techniques for assessing a subjective prior distribution.

Three strategies have been adopted which broadly follow the ideas for parameters of primary interest described earlier: elicitation (Section 5.2), summary of evidence (Section 5.4), and reference priors (Section 5.5).

Elicitation of opinion. In order to be able to make judgements about their relative plausibility, we need to have a clear interpretation of what different values of τ signify. We can first note that 95% of values of θ will lie in the

interval $\mu \pm 1.96\tau$, and hence the 97.5% and 2.5% values of θ are $2 \times 1.96 \times \tau$ apart. θ will often be measured on a logarithmic scale, for example as a log(odds ratio), and hence the ratio of the 97.5% odds ratio to the 2.5% odds ratio is $\exp(3.92\tau)$, roughly representing the 'range' of odds ratios. For example, in the context of meta-analysis, Smith *et al.* (1995) thought that it was unlikely that the between-study odds ratios would vary by more than an order of magnitude, and hence considered $\exp(3.92\tau) = 10$, or $\tau = \log(10)/3.92 = 0.59$ to represent a 'high' value of the standard deviation τ.

An alternative approach is to imagine two randomly chosen θs drawn from the random-effects distribution, whose difference will have distribution $\theta_1 - \theta_2 \sim \text{N}[0,\ 2\tau^2]$ by (2.26). Their absolute difference $|\theta_1 - \theta_2|$ therefore has a normal distribution constrained to be greater than 0, which is a half-normal distribution $\text{HN}[2\tau^2]$ (Section 2.6.7). This distribution has median $\Phi^{-1}(0.75) \times \sqrt{2}\tau = 1.09\tau$, which is therefore the median difference between the maximum and minimum of a random pair of θs (Larsen *et al.*, 2000). If θ is, for example, a log(odds ratio), then $\exp(1.09\tau)$ is the median ratio of the maximum to the minimum of any random pair of odds ratios drawn from the distribution.

Table 5.2 illustrates these two interpretations for a range of values of τ when θ represents a log(odds ratio). It is apparent that $\tau = 1$ corresponds to a substantial heterogeneity, with a random pair having a median ratio of 3, for example one trial showing no effect and another showing an odds ratio of 3. $\tau = 2$ means the trials are effectively independent.

Table 5.2 Possible interpretations of τ, the standard deviation of the log(odds ratio) in a hierarchical model $\theta \sim \text{N}[\mu, \tau^2]$. The 'range' $\exp(3.92\tau)$ is actually the ratio of the 97.5% to the 2.5% point of the distribution of odds ratios, while $\exp(1.09\tau)$ is the median ratio of the maximum to minimum odds ratio in a random pair of θs drawn from the distribution.

τ	$\exp(3.92\tau)$: 'range' of odds ratios	$\exp(1.09\tau)$: median ratio of random pair
0.0	1.00	1.00
0.1	1.48	1.11
0.2	2.19	1.24
0.3	3.24	1.39
0.4	4.80	1.55
0.5	7.10	1.72
0.6	10.51	1.92
0.7	15.55	2.14
0.8	23.01	2.39
0.9	34.06	2.67
1.0	50.40	2.97
1.5	357.81	5.13
2.0	2540.20	8.84

In conclusion, values of τ from 0.1 to 0.5 may appear reasonable in many contexts, from 0.5 to 1.0 might be considered as fairly high, and above 1.0 would represent fairly extreme heterogeneity.

When assessing a subjective prior distribution for τ, we first need to consider whether $\tau = 0$ is a plausible value, representing no variability between θs. At the other extreme, we should think of an 'upper' value for τ which we shall label τ_u; Table 5.2 may be useful for this. A possible prior distribution is then a half-normal distribution HN$[(\tau_u/1.96)^2]$ (Pauler and Wakefield, 2000). This will have its mode at 0 and be steadily declining in τ, with an upper 95% point at τ_u. Its median will be $\Phi^{-1}(0.75) \times \tau_u/1.96 = 0.39\tau_u$. This is illustrated in Figure 5.10(a) for $\tau_u = 1$, which may be a reasonable prior in many situations; see Example 8.5.

Summary of evidence. It is natural to construct a prior distribution for τ from an analysis of past hierarchical models in the context being considered, in order to determine reasonable values of τ experienced in practice. Thus we could, for example, study the typical variability between subgroups, between institutions in their clinical performance, or between centres in multi-centre clinical trials. In the field of meta-analysis, Higgins and Whitehead (1996) and Smith *et al.* (1996) both consider empirical distributions of past τs: essentially they are carrying out a meta-analysis of meta-analyses. Higgins and Whitehead (1996) go on to formally construct an additional level in the hierarchical model in which τ is a random effect with a distribution. They restrict attention to gamma distributions for τ^{-2}, and estimate that a τ^{-2} for a new meta-analysis has a Gamma[1.0, 0.35] distribution. Transforming this onto the τ scale using standard theory for probability distributions yields a root-inverse-gamma distribution RIG[1, 0.35] (Section 2.6.6). This has its mode at $\tau = 0.48$, mean $\sqrt{0.35\pi} = 1.05$ and a standard deviation of ∞. Figure 5.10(b) reveals it to rule out low values of τ.

Default 'non-informative' priors. A number of suggestions have been made for placing a 'default' prior distribution on τ or, equivalently, τ^2. The standard reference prior for a sampling variance, $p(\sigma^2) \propto \sigma^{-2}$ (Section 5.5.1), is inappropriate at the random-effects level as it gives an *improper* posterior distribution (Berger, 1985). Five of the main contenders are listed below.

(a) *A 'just proper' prior*. An inverse gamma distribution such as

$$\tau^{-2} \sim \text{Gamma}[0.001, 0.001]$$

is proper and close to being uniform on $\log(\tau)$. Figure 5.10(c) shows that it gives a high weight near $\tau = 0$ and so, if the likelihood supports low values of τ, it could show a preference for a low variance. This may be reasonable behaviour but should be acknowledged.

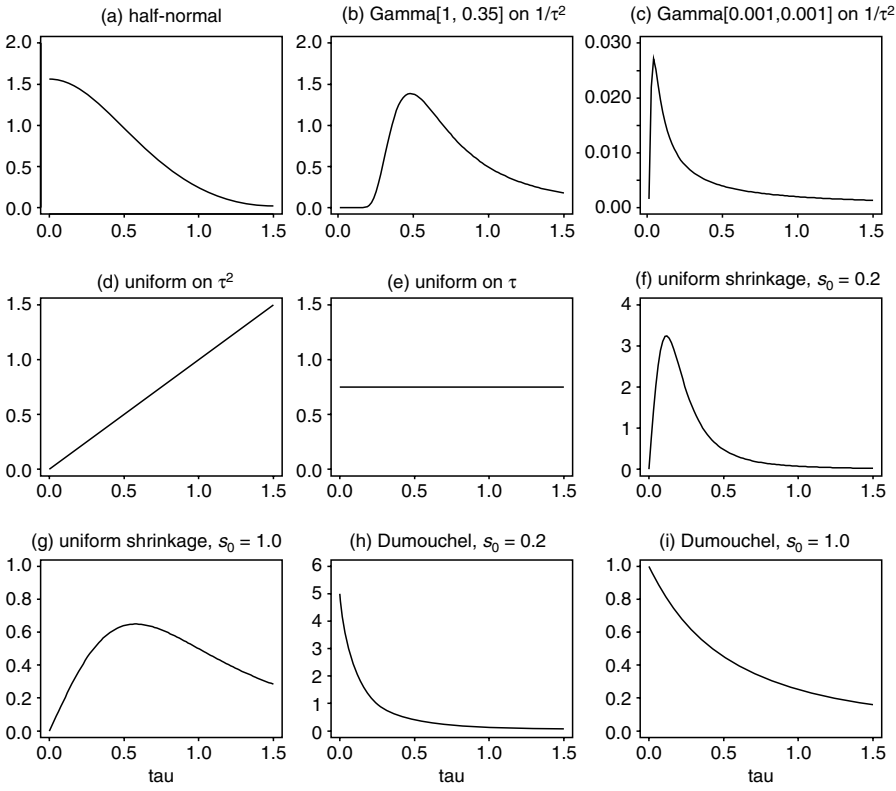

Figure 5.10 Alternative prior distributions on the between-unit standard deviation τ: see the text for discussion of each possible choice. (a) supports equality between units ($\tau = 0$) and discounts substantial heterogeneity ($\tau = 1$); (b) is based on an empirical summary of past meta-analyses and forces heterogeneity; (c) is an 'almost' improper prior that has been widely used but gives strong preference for small τ, (f) to (i) depend on the amount of evidence in the data, with $s_0 = 1$ representing weak evidence, and $s_0 = 0.2$ strong evidence.

(b) *Uniform on* τ^2. The uniform prior

$$p(\tau^2) \propto \text{ constant}$$

is recommended by Gelman *et al.* (1995) and can be restricted to a suitable range to make it a proper distribution. Figure 5.10(d) shows its preference for high values of τ, which does not appear attractive.

(c) *Uniform on* τ. The uniform prior

$$p(\tau) \propto \text{ constant}$$

is a natural contender and is shown in Figure 5.10(e). Nevertheless, it would be inappropriate to term this 'non-informative', as it is a fairly

strong statement to declare that small values of τ are as likely as large values.

(d) *Uniform shrinkage priors*. Following Section 3.17, we assume an approximate normal likelihood with $y_k \sim \text{N}[\theta_k, s_k^2]$. A number of authors (Christiansen and Morris, 1997b; Natarajan and Kass, 2000; Daniels, 1999; Spiegelhalter, 2001) have investigated a prior on τ^2 that is equivalent to a uniform prior on the 'average' shrinkage

$$B_0 = s_0^2/(s_0^2 + \tau^2)$$

where s_0^2 is the harmonic mean of the s_k^2, *i.e.*

$$\frac{1}{s_0^2} = \frac{1}{K}\sum_k \frac{1}{s_k^2}.$$

Placing a uniform distribution on B_0 is equivalent to $1 - B_0 = \tau^2/(s_0^2 + \tau^2)$ having a uniform distribution. This leads to

$$p(\tau^2) = \frac{s_0^2}{(s_0^2 + \tau^2)^2},$$
$$p(\tau) = \frac{2\tau s_0^2}{(s_0^2 + \tau^2)^2}.$$

The uniform shrinkage prior distributions have the following properties:

	τ^2	τ
Mode	0	$s_0/3 = 0.57s_0$
First quartile	$s_0^2/\sqrt{3}$	$s_0/\sqrt{3} = 0.57s_0$
Median	s_0^2	s_0
Mean	–	$\pi s_0/2 = 1.57s_0$
Third quartile	$3s_0^2$	$\sqrt{3}s_0 = 1.73s_0$
Variance	–	–

The prior on τ^2 has an asymptote at 0, but the implied prior on τ returns to 0 at the origin.

Suppose $s_k^2 = \sigma_k^2/n_k$, so that

$$y_k \sim \text{N}[\theta_k,\ \sigma_k^2/n_k].$$

Three situations can be distinguished:

(i) $\sigma_k^2 = \sigma^2$, which is assumed known, such as the frequent adoption of $\sigma^2 = 4$. Then $s_0^2 = \sigma^2/\overline{n}$.

(ii) $\sigma_k^2 = \sigma^2$, which is unknown. σ^2 could then be given a standard Jeffreys prior $p(\sigma^2) \propto \sigma^{-2}$ – this induces an appropriate dependency between τ^2 and σ^2.

(iii) Each σ_k^2 is unknown. The σ_k^2 could then be assumed either exchangeable or independent. Within-unit empirical estimates $\hat{\sigma}_k^2$ can be used to estimate s_0^{-2} by

$$\frac{1}{s_0^2} = \frac{1}{K}\sum_k \frac{n_k}{\hat{\sigma}_k^2}.$$

Essentially, fixed effects are fitted first and then the average precision is used as an estimate of s_0^{-2}. This approach is illustrated in Examples 6.10 and 8.1.

In studies based on events we might equate s_0^2 to $4/n_0$, where n_0 represents the mean number of events in each study. Hence $s_0 = 0.2$ corresponds to large studies with an average of 100 events each, while $s_0 = 1.0$ corresponds to very small studies with an average of 4 events each. These priors are shown in Figures 5.10(f) and 5.10(g), showing that large studies lead to strong prior weight on low values of τ and hence an expectation of the studies showing 'similar' results.

(e) *DuMouchel priors*. DuMouchel (DuMouchel and Normand, 2000) has suggested a similar form to the uniform shrinkage prior but assuming a uniform prior for $s_0/(s_0 + \tau)$, which implies

$$p(\tau) = \frac{s_0}{(s_0 + \tau)^2},$$
$$p(\tau^2) = \frac{s_0}{2\tau(s_0 + \tau)^2}.$$

The distributions have the following properties:

	τ^2	τ
Mode	0	0
First quartile	$s_0^2/9$	$s_0/3$
Median	s_0^2	s_0
Mean	–	–
Third quartile	$9s_0^2$	$3s_0$
Variance	–	–

Note that the quartiles are at $B_0 = 0.1,\ 0.5,\ 0.9$, showing the DuMouchel prior gives preference to either strong or weak shrinkage. Figures 5.10(h) and 5.10(i) show the DuMouchel priors for $s_0 = 0.2$ and $s_0 = 1.0$, revealing the preference of these priors for both low and high values of τ.

In general our preference will be to use a uniform prior on τ as a baseline when there is reasonable information from the data. When prior information is

strong or important a suitably informative prior can be chosen: the half-normal appears particularly attractive.

These points serve to underline the importance of carefully choosing and justifying the prior distributions used within a hierarchical setting, and subjecting those used to the type of sensitivity analysis adopted in Examples 6.10, 7.2, 8.1, 8.3 and 8.5.

5.8 EMPIRICAL CRITICISM OF PRIORS

The ability of subjective prior distributions to predict the true benefits of interventions is clearly of great interest, and Box (1980) suggested a methodology for comparing priors with subsequent data. The prior is used to derive a predictive distribution for future observations, and thus to calculate the chance of a result with lower predictive ordinate than that actually observed: when the predictive distribution is symmetric and unimodal, this is analagous to a traditional two-sided P-value in measuring the predictive probability of getting a result at least as extreme as that observed. With normal assumptions we can use (3.23) but substituting m for n, to give a pre-trial predictive distribution

$$Y_m \sim \mathrm{N}\left[\mu,\ \sigma^2\left(\frac{1}{n_0}+\frac{1}{m}\right)\right]. \tag{5.5}$$

Given observed y_m, the predictive probability of observing a Y_m less than that observed is

$$P(Y_m < y_m) = \Phi\left(\frac{y_m-\mu}{\sigma\sqrt{\dfrac{1}{n_0}+\dfrac{1}{m}}}\right), \tag{5.6}$$

and hence Box's generalised significance test is given by

$$2\min\left[P(Y_m < y_m),\ 1-P(Y_m < y_m)\right].$$

Another way of obtaining (5.6) is as the tail area associated with a standardised test statistic contrasting the prior and the likelihood, *i.e.*

$$z_m = \frac{y_m-\mu}{\sigma\sqrt{\dfrac{1}{n_0}+\dfrac{1}{m}}},$$

showing that Box's statistic explicitly acts as a measure of *conflict* between prior and data.

Example 5.5 *GREAT (continued): Criticism of the prior*

In Example 3.6, $\mu = -0.26$, $n_0 = 236.7$, $m = 30.5$, $\sigma = 2$ and hence the predictive distribution for the observed log(OR) has mean -0.26 and standard deviation 0.39. This is shown in Figure 5.11 with the observed OR $= 0.48$ ($y_m = \log(\text{OR}) = -0.74$) marked. Box's measure is twice the shaded area, which is $2\Phi((-0.74 + 0.26)/0.39) = 0.21$. We may also obtain this result as the standardised test statistic between prior and likelihood $z = -1.25$, with a two-sided *P*-value of 0.21. Thus there is no strong evidence for conflict between prior and data in the GREAT example.

There have been a number of prospective elicitation exercises for clinical trials, and many of these trials have now reported their results. Table 5.3 shows a selection of results, including the intervals for the prior distributions for treatment effects, the evidence from the likelihood, and Box's *P*-value summarising the conflict between the prior and the likelihood. The references for the prior assessments and the data are provided at the end of the section.

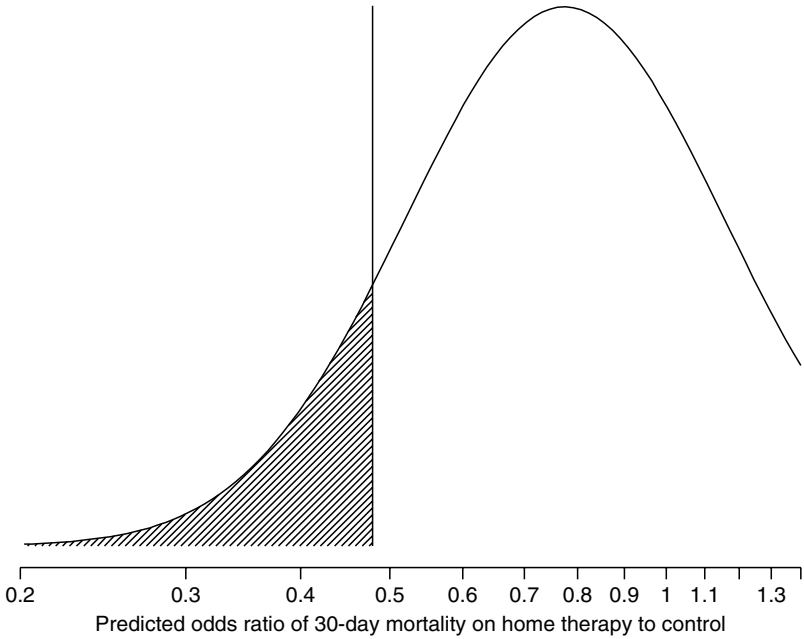

Figure 5.11 Predictive distribution for observed OR in the GREAT trial with observed OR $= 0.48$ (log(OR) $= -0.74$) marked. Box's measure of conflict between prior and data is twice the shaded area $= 0.21$.

Table 5.3 A comparison of some elicited subjective prior distributions and the consequent results of the clinical trials. In each case a pooled prior was provided, assumed normal on a log(hazard ratio) scale – Box's *P*-value is calculated on this scale. This is transformed to a hazard ratio (HR) scale where HR < 1 corresponds to benefit of the new treatment: median and 95% intervals are given (note the gastric cancer results are reported with the inverse hazard ratio in Example 6.4).

Study	Prior		Likelihood		*Z*	*P*
	HR	95% interval	HR	95% interval		
CHART (Lung)[1]	0.76	(0.48, 1.19)	0.76	(0.63, 0.90)	0.00	1.00
CHART (HN)[1]	0.72	(0.44, 1.20)	0.95	(0.79, 1.14)	1.02	0.31
Thiotepa X1[2]	0.61	(0.37, 1.01)	1.11	(0.78, 1.59)	1.91	0.06
Osteosarcoma[3]	0.90	(0.55, 1.50)	1.07	(0.79, 1.45)	0.58	0.56
Gastric cancer[4]	0.88	(0.61, 1.28)	1.10	(0.87, 1.39)	1.00	0.32

Sources: [1]Example 6.6. [2]Spiegelhalter and Freedman (1986) and Richards *et al.* (1994). [3]Spiegelhalter *et al.* (1993) and Souhami *et al.* (1994). [4]Example 6.4.

Table 5.3 shows the generally poor experience obtained from prior elicitation. The clinicians are universally optimistic about the new treatments (median of prior hazard ratios less than 1), whereas only two of the trials – the CHART trials – eventually showed any evidence of benefit from the new treatment (likelihood hazard ratio less than 1), and only the CHART lung trial showed 'significant' benefit. The thiotepa trial shows particularly high conflict between data and prior, with the clinicians expecting a substantial benefit from thiotepa which failed to materialise. This also reflects the experience of Carlin *et al.* (1993) in their elicitation exercise.

Far from invalidating the Bayesian approach, such a conflict between prior and data only serves to emphasise the importance of pre-trial elicitation of belief; having these opinions explicitly recorded will help a data monitoring committee to focus on the difference between anticipated and actual results. Of course, the precise action to be taken in the face of considerable conflict will depend on the circumstances.

5.9 KEY POINTS

1. The use of a prior is based on judgement and hence a degree of subjectivity cannot be avoided.
2. The prior may be important and is not unique, and so a range of options should be examined in a sensitivity analysis.
3. The quality of subjective priors (as assessed by predictions) show predictable biases in terms of enthusiasm.
4. For a prior to be taken seriously by an external audience, its basis must be explicitly given. A variety of models exist for using historical data as a basis for prior distributions.

5. Archetypal priors, expressing both scepticism and enthusiasm, may be useful for identifying a reasonable range of prior opinion.
6. Great care is required in using default priors intended to be minimally informative.
7. Exchangeability assumptions lead to hierarchical models that are valuable in many situations, but such judgements should not be made casually.
8. Sensitivity analysis plays a crucial role in assessing the impact of particular prior distributions, whether elicited, derived from evidence, or reference, on the conclusions of an analysis.

EXERCISES

5.1. Consider tossing a drawing-pin (thumbtack) onto a flat surface.
   (a) Assess *your* beliefs about the true proportion of times that it will fall point-up, in terms of a best estimate, and low and high assessments.
   (b) Derive a beta prior distribution for this proportion based on these beliefs.
   (c) Use the conjugate beta-binomial model of Section 3.6.2 to update these beliefs after 12 tosses using the *same* hand.

5.2. Prior to the publication of the UK Medical Research Council RCT evaluating the use of high-energy neutrons for treatment of patients with tumours of the pelvic region (bladder, cervix, prostate and rectum) in 1991 a number of RCTs evaluating low-energy neutrons had been reported (Errington *et al.*, 1991). The results of these RCTs are summarised in Table 5.4. (a) Assuming balanced trials, approximate the log(hazard ratio) and its variance for each of these studies. (b) Use the 'method of moments' (3.37) to estimate the between-study variance τ^2. Use this historical evidence to establish a prior distribution for the MRC trial, assuming (c) the new trial is estimating the

Table 5.4 Summary of RCT evidence in terms of survival at 12 months for low-energy neutron therapy compared to conventional radiotherapy for tumours of the pelvic region.

Study	Year of publication	Site	Neutrons		
			Deaths(O)	Expected(E)	V[O-E]
Batterman	1982	Bladder and Rectum	34	32.6	5.3
Pointon	1985	Bladder	16	13.7	5.1
Duncan	1987	Bladder	26	20.1	6.7
Duncan	1987	Rectum (inoperable)	17	12.8	2.1
Duncan	1987	Rectum (recurrent)	10	7.3	2.0
Duncan	1987	Bladder	4	4.2	0.6

mean treatment effect of the previous trials, and (d) the new trial is exchangeable with the previous trials. The fact that the previous trials were low-energy, and the new trial high-energy, might lead one to doubt the exchangeability model.

(e) What model for systematic bias might be reasonable?

5.3. In Exercise 5.2, on average the oncologists claimed that they required the survival rate for neutron therapy to be 61.5%, relative to a 1-year survival rate of 50% in the control group, before considering it for routine treatment. The range of equivalence was therefore taken to be from 50% to 61.5%. For each of the situations modelled, obtain the prior probabilities of no benefit of neutrons relative to conventional therapy, the range of equivalence, and clinical benefit in favour of neutron therapy.

5.4. In addition to the meta-analysis in Exercise 5.2, the beliefs and clinical demands of ten oncologists were elicited before the final analysis of the high-energy trial data. Table 5.5 summarises the elicited prior distributions for all ten oncologists for the 1-year survival rate on neutron therapy compared to a 50% survival rate with conventional therapy.

(a) Calculate an average histogram.

(b) Transform this to a histogram on the log(hazard ratio) scale using the techniques in Example 5.1.

(c) Fit a normal distribution to this distribution by matching the mean and variance or by some other method.

(d) Given the disagreement between the oncologists, do you think it reasonable to create such a pooled distribution?

5.5. Prior to the publication of the HAI RCT considered in Exercise 2.7, results from five previous RCTs had been published, and these are summarised in terms of overall survival in Table 5.6. (a) For each trial, estimate the

Table 5.5 Elicited prior beliefs in terms of percentage survival at 12 months for high-energy neutron therapy compared to a 50% survival rate for conventional radiotherapy for tumours of the pelvic region.

ID	Neutron 1-year survival rate (%)																	
	15–	20–	25–	30–	35–	40–	45–	50–	55–	60–	65–	70–	75–	80–	85–	90–	95–	Total
1	0	0	0	10	20	25	15	15	10	5	0	0	0	0	0	0	0	100
2	0	0	5	10	20	35	30	0	0	0	0	0	0	0	0	0	0	100
3	0	0	0	0	0	20	60	20	0	0	0	0	0	0	0	0	0	100
4	0	0	0	0	0	30	30	30	10	0	0	0	0	0	0	0	0	100
5	0	0	5	10	25	20	15	10	10	5	0	0	0	0	0	0	0	100
6	0	0	0	0	0	0	5	0	0	0	15	20	20	15	10	10	5	100
7	5	5	10	25	25	15	5	2.5	2.5	2.5	2.5	0	0	0	0	0	0	100
8	0	0	0	5	5	10	25	25	15	5	2.5	2.5	2.5	2.5	0	0	0	100
9	0	0	0	5	5	10	25	25	15	5	2.5	2.5	2.5	2.5	0	0	0	100
10	0	0	0	15	15	25	20	20	5	0	0	0	0	0	0	0	0	100

log(hazard ratio) and the effective number of events assuming $\sigma = 2$. Obtain a prior distribution for the log(hazard ratio) for overall survival of HAI compared to control patients, assuming (b) a common effect in all trials, (c) that the past trials are exchangeable with the current trial.

5.6. Sutton *et al.* (2000, p. 261) consider 17 single-arm studies of either radiotherapy alone (RTx) following surgery for childhood medulloblastoma, or radiotherapy together with adjuvant chemotherapy (RTx + Chm) following surgery. Table 5.7 displays the 5-year survival rates together with standard errors for all 17 studies.

Table 5.6 Summary of RCT evidence in terms of overall survival, prior to 1994, for HAI compared to control for the treatment of non-resectable liver metastases associated with primary colorectal cancer.

Study	Year publication	HAI		Control		$O-E$	V[O-E]
		Deaths	Total	Deaths	Total		
MSKCC	1987	43	45	48	48	−5.8	21.9
NCCTG	1990	39	39	35	35	−1.0	17.9
NCI	1987	25	32	26	32	−2.7	12.5
City of Hope	1986	9	9	6	6	−2.3	3.3
France	1992	72	81	78	82	−14.2	36.4

Table 5.7 Five-year survival rates and standard errors for single-arm studies considering either radiotherapy alone (RTx) or radiotherapy together with adjuvant chemotherapy (RTx + Chm) following surgery for childhood medulloblastoma.

Study	RTx + Chm		RTx	
	S_5	SE(S_5)	S_5	SE(S_5)
1	0.83	0.030	–	–
2	0.82	0.120	–	–
3	0.96	0.039	–	–
4	0.82	0.384	–	–
5	0.55	0.188	–	–
6	0.64	0.170	–	–
7	0.26	0.196	–	–
8	0.60	0.097	–	–
9	0.36	0.170	–	–
10	0.93	0.120	–	–
11	–	–	0.71	0.184
12	–	–	0.48	0.223
13	–	–	0.41	0.087
14	–	–	0.32	0.057
15	–	–	0.34	0.080
16	–	–	0.71	0.068
17	–	–	0.33	0.071

(a) Looking at the data, do you think a pooled effect is a reasonable assumption?
(b) Estimate the between-study variance for each treatment using (3.37).
(c) Assuming a normal random-effects model, estimate a prior distribution for the 5-year survival in a new study, assuming exchangeability with the previous studies.
(d) Combine these two prior distributions into a prior for the difference in the 5-year survival rate, *i.e.* RTx + Chm − RTx, in a proposed clinical trial.
(e) Is normality a reasonable assumption for the random-effects distribution?

5.7. The trial discussed in Exercise 5.2 ended by yielding an estimated hazard ratio of 1.52 (95% CI from 0.91 to 2.50), *i.e.* in favour of the control group (Errington *et al.*, 1991).
(a) For the data-based prior using all six previous studies, assess the conflict of these prior distributions, using the methods of Section 5.8.
(b) Repeat this for oncologists 6 and 7.

5.8. Verify for a normal model in Section 5.4, when there is a single historical study, the assumptions under which exchangeability, bias and discounting can lead to the same prior distribution. Does this hold for multiple studies?

5.9. Plot three half-normal prior distributions for a model parameter τ which have the properties that:
(a) the mean of τ is 1.5;
(b) the median is 3; and
(c) the probability of τ being greater than 1 is 5%.

5.10. For the magnesium meta-analysis in Example 3.13 calculate and plot DuMouchel and uniform shrinkage prior distributions for the random-effects standard deviation τ.