4

Meta-analysis

4.1 Summary

In this chapter, we illustrate the use of Bayesian meta-analysis in developing probability distributions on the magnitude of the effects of a medical treatment. These distributions can be used informally to support decision making on patient treatment and on allocation of resources to future trials. They can also become components of a formal decision analysis, of a comprehensive decision model, or of a stochastic optimization. The chapter begins with a brief overview of meta-analysis, with pointers to the extensive, fascinating, and controversial literature. This is followed by an introductory example which looks at the efficacy of tamoxifen in adjuvant treatment of early breast cancer and serves as an introduction to some of the key features of Bayesian meta-analysis. The case study of this chapter deals with the synthesis of evidence from several clinical trials comparing the effectiveness of commonly recommended prophylactic treatments for migraine headaches. The case study is based closely on Dominici *et al.* (1999).

4.2 Meta-analysis

The term 'meta-analysis' originated in psychology. Glass (1976) used it to describe 'the statistical analysis of a large collection of results from individual studies for the purpose of integrating the findings'. In medicine, the practice of formally integrating findings from different studies can be traced back at least to the 1950s (Beecher, 1955). Today, meta-analysis has become a key component of evidence-based medicine. A MEDLINE search of articles published in 1997 yielded 775 articles with the term 'meta-analysis' in the title, abstract or keywords. This is the result of the growth in the number of clinical trials (approximately 8000 new clinical trials begin every year, according to Olkin 1995a) and of the desire to use accruing evidence as early as possible in improving health care decisions. Meta-analysis is also becoming widely applied beyond randomized clinical trials, for example in epidemiological research. Interesting discussions of the role of meta-analysis

in clinical research and decision making will be found in L'Abbé *et al.* (1987) and Gelber and Goldhirsch (1991).

Moher and Olkin (1995) summarize the reasons for the success of meta-analysis:

> Why the dramatic increase in the number of published meta-analyses? Two examples may help address the question. A meta-analysis published in 1990 described the efficacy of corticosteroids given to mothers expected to deliver prematurely (Crowley *et al.*, 1990). The results of the meta-analysis indicated that corticosteroids significantly reduced morbidity and mortality of these infants. This analysis convincingly showed that such evidence was available at least a decade earlier ([i.e.] 1980). Had a meta-analysis been conducted when the evidence became available, much unnecessary suffering might have been avoided.
>
> In an article on meta-analyses published in *JAMA* in 1992 (Antman *et al.*, 1992), the research team showed that textbook recommendations for the treatment of patients with suspected myocardial infarction often lagged behind the empirical evidence by as much as 10 years. The group also noted that, at times, the opinion of experts writing these books was in sharp contrast to the empirical evidence.

In general, meta-analyses can offer important advantages over more traditional narrative approaches to the overview of scientific evidence (Chambers and Altman, 1994). These advantages result primarily from the systematic, explicit, and quantitative nature of the synthesis provided by meta-analysis; from the possibility of assessing uncertainty about the results of the synthesis; and from the increase in sample sizes deriving from the combination of studies. From the point of view of the philosophy of science, meta-analysis is a novel paradigm for scientific investigation, reflecting the scientific community's adaptation to the information explosion of the last few decades. Goodman (1998) hails meta-analysis as 'one of the most important and controversial developments in the history of science'.

As with many promising new paradigms, meta-analysis has given rise to misuses and controversy. Criticisms of meta-analysis are both conceptual and methodological. Randomized clinical trials are modeled on controlled experimentation and are generally agreed to address in a satisfactory way the question of the causal relationship between treatment and outcomes. Difficulties may arise: the sample size can be small, the measurable outcomes may problematic, the protocol may be difficult to implement. But the paradigm is thought to be sound. The design of a meta-analysis study is different from that of a randomized clinical trial. Some authors (such as Erwin, 1984) consider these differences to be sufficient to question the possibility of

inferring causal relations from meta-analysis of clinical trials (even though it is possible to infer such relations from the individual trials). A related point is made by Charlton (1996), who writes that 'the prestige of meta-analysis is based upon a false model of scientific practice'; meta-analysis, in his view, cannot be considered a hypothesis testing activity, and should be confined to effect size estimation.

Another common criticism is epitomized by the slogan 'many bad studies don't make a good one'. The argument is this: for a given clinical question there either is a single, critical, well-conducted, study, with a large enough sample, that can provide guidance to physicians, or there is not. If there is not, that is often because existing trials are conflicting, diverse, or not sufficiently well conducted. Meta-analysis, it is argued, should not be used to attempt to settle a clinical question in the presence of vexing primary data problems.

Methodologically, critical issues are search strategies, publication bias, and study heterogeneity. Some meta-analyses are conducted by gathering primary data from the study investigators (Simes, 1986; Early Breast Cancer Trialists Collaborative Group (EBCTCG), 1990). This is an appropriate and effective strategy, but it is not always feasible. Most meta-analyses are based on published results. The key elements of a meta-analysis of published results are the criteria used for searching for, evaluating, and selecting articles. Concerns about the quality of these criteria in applied research are common (Cook *et al.*, 1995; Sacks *et al.*, 1987; Chambers and Altman, 1994). Guidelines for rigorous procedural methodology have been put forward by several authors (Simes, 1986; Deeks *et al.*, 1997) and by *ad hoc* working groups in clinical trials (Moher *et al.*, 1999) and epidemiology (Stroup *et al.*, 2000).

Publication bias arises because scientific journals selectively publish studies with statistically significant results. For example, in 1986–87 about 76% of the articles published in the *New England Journal of Medicine* used statistical tests; 88% of these tests rejected the null hypothesis. These proportions are even higher in the experimental psychology literature (Sterling *et al.*, 1995). It is clear that the results of published studies are not a representative sample of the results of all studies, and that this may systematically bias the result of a meta-analysis which considers only published results; for example, a small treatment effect is likely to become artificially magnified. Diagnostics for the presence of publication bias are based on observing a relationship between sample size and effect size (Duval and Tweedie, 2000).

Because the aims of a meta-analysis are typically broader that those of the individual trials being reviewed, it is likely that any sizable meta-analysis will have several sources of between-trial heterogeneity. These include differences in specific treatment regimens, patient eligibility criteria, baseline disease severity, and outcomes. Complete homogeneity is not necessarily a desirable goal. Moher and Olkin (1995) note that 'sometimes, too much homogeneity of studies will stifle generalizations to a larger population. On the other hand, too much study heterogeneity will weaken the results. Thus, there has to

be an understanding of the sources of the heterogeneity.' Similar views are expressed by Thompson (1994): 'discussion of heterogeneity in meta-analysis affects whether it is reasonable to believe in one overall estimate that applies to all the studies encompassed, implied by the so called fixed effect method of statistical analysis. Undue reliance may have been put on this approach in the past, causing overly simplistic and overly dogmatic interpretation.' Statistical models and techniques for quantifying heterogeneity and for developing and interpreting summary estimates are illustrated extensively in the rest of this chapter.

There are also issues surrounding the quality of reporting of meta-analyses. Jadad and McQuay (1996) carried out a systematic review of methodology in 74 meta-analyses of analgesic interventions. They found that: 'Ninety percent of the meta-analyses had methodological flaws that could limit their validity. The main deficiencies were lack of information on methods to retrieve and to assess the validity of primary studies and lack of data on the design of the primary studies'. They also found that 'meta-analyses of low quality produced significantly more positive conclusions'. Sacks *et al.* (1987) identify six content areas thought to be important in the conduct and reporting of meta-analyses: study design, combinability, control of bias, statistical analysis, sensitivity analysis, and problems of applicability. Moher and Olkin (1995) review the issues and lay the groundwork for developing standards for the reporting of meta-analyses.

In this chapter we will discuss some statistical tools for meta-analysis. What is their role amidst this controversy? I will take a pragmatic view: there are decisions to be made today, and we should make them using the best available evidence. The limitations of mathematical modeling as a means of synthesis can be serious, but a well-conducted analysis can often limit publication bias, incorporate heterogeneity, and provide practical guidance.

4.3 Bayesian meta-analysis

Statistical methods have been developed for meta-analysis and are continuously being refined in response to the increasing demand for meta-analysis and the increasing complexity of the meta-analyses performed. Olkin (1995b) provides a historical perspective; Sutton *et al.* (1998) provide a comprehensive and up-to-date review; Hasselblad and McCrory (1995) give a more concise practical guide; Stangl and Berry (2000) provide a collection of state-of-the-art applications. A classic and beautiful book on the subject is that by Hedges and Olkin (1985). Reviews of software include Sutton *et al.* (2000) and Normand (1995).

Here we will be concerned with Bayesian methods, whose use is well established in the statistical literature (DuMouchel and Harris, 1983; Berry, 1990; DuMouchel, 1990; Eddy *et al.*, 1990) and is gaining acceptance in the medical literature as well (Baraff *et al.*, 1993; Berry, 1998; Biggerstaff

et al., 1994; Brophy and Joseph, 1995; Sorenson and Pace, 1992; Tweedie *et al.*, 1996). Many interesting situations can now be modeled using software packages such as BUGS. Detailed applications of meta-analysis are illustrated by Smith *et al.* (1995; 2000).

Interest in Bayesian meta-analysis is motivated by several desiderata:

a) providing decision makers with summaries of evidence in the form of probability distributions, given all available evidence. This input is appropriate for a subsequent decision model. For example, Pallay and Berry (1999) demonstrate this using Bayesian meta-analysis to assess the worthiness of a phase III trial.
b) developing approaches for modeling trial heterogeneity, and devising summary measures that are relevant to decision making in the presence of trial heterogeneity. For example, it is important to address the question of the probability that a patient receiving drug A survives longer than a patient receiving drug B for a patient from a future, or unobserved, or hypothetical trial. This is a question of prediction. Bayesian random effects and hierarchical models (Lindley and Smith, 1972; Raudenbush and Bryk, 1985; Morris and Normand, 1992; Carlin and Louis, 2000) provide a flexible and practical framework for developing predictive models.
c) modeling unobserved aspects of the data generation and reporting processes. Examples include modeling publication bias (Silliman, 1997; Givens *et al.*, 1997), missing covariates (Lambert *et al.*, 1997), and partially reported results (Dominici *et al.*, 1999).
d) realistically assessing uncertainty. Bayesian simulation-based methods do not need to rely on asymptotic approximations and can straightforwardly accommodate uncertainty about nuisance parameters, often leading to more conservative and accurate statements about uncertainty in the overall conclusions. For example, Carlin (1992) finds that, in the meta-analysis of 2×2 tables, Bayesian estimates of parameter uncertainty are more accurate than the corresponding empirical Bayes estimates.

The limitations of Bayesian meta-analysis are related primarily to the added complexity of implementation. Depending on the application and the state-of-the-art in the field, elicitation of prior information may also become complex or controversial.

4.4 Tamoxifen in early breast cancer

4.4.1 Background

To illustrate some of the interesting features of Bayesian meta-analysis in a simple and common situation, let us consider a case in which each study reports a 2×2 table of successes and failures for both a treatment and a control group. This situation is exemplified by the data of Table 4.1, taken

Table 4.1 Summary results for the tamoxifen and control group in 14 randomized clinical trials, as reported by EBCTCG (1998a). Breast cancer recurrence rates are abbreviated as Rec.

	Tamoxifen			Control			
Study s	Rec. x_1^s	Total n_1^s	Odds of rec. $\frac{x_1^s}{n_1^s - x_1^s}$	Rec. x_0^s	Total n_0^s	Odds of rec. $\frac{x_0^s}{n_0^s - x_0^s}$	Odds ratios $\frac{x_1^s/(n_1^s - x_1^s)}{x_0^s/(n_0^s - x_0^s)}$
1	55	97	1.310	67	101	1.971	0.665
2	137	282	0.945	187	306	1.571	0.601
3	505	927	1.197	590	915	1.815	0.659
4	62	123	1.016	74	140	1.121	0.907
5	99	239	0.707	118	236	1.000	0.707
6	50	130	0.625	49	107	0.845	0.740
7	185	311	1.468	200	319	1.681	0.874
8	186	303	1.590	187	307	1.558	1.020
9	148	325	0.836	178	325	1.211	0.691
10	25	79	0.463	38	86	0.792	0.585
11	223	344	1.843	224	350	1.778	1.037
12	183	937	0.243	185	936	0.246	0.985
13	2	12	0.200	0	8	0.000	
14	129	434	0.423	159	449	0.548	0.771

from the overview of clinical trials of adjuvant tamoxifen for women with early breast cancer carried out by the EBCTCG (1998a). Since the mid-1980s this group has been responsible for thorough and influential reviews of randomized clinical trials of all treatments for early breast cancer. Their analysis is based on patient-level data obtained from study investigators, rather than on published summaries. As a result it is likely to be very robust to publication bias. Extensive effort has been devoted to the grouping of trial results according to relevant clinical characteristics, such as type and duration of treatment, dose of drug, use of other therapies in conjunction with tamoxifen, patient prognostic factors, and type of outcome (recurrence or death). Data collection and data checking procedures are described in detail in EBCTCG (1990; 1998a).

The data of Table 4.1 refer to 14 randomized clinical trials, as reported by EBCTCG (1998a). Each trial compared a group receiving tamoxifen for an average of about one year with a control group not receiving tamoxifen. We consider the endpoint to be breast cancer recurrence. There is variation in odds of recurrence among trials. In this case, this variation is mainly attributable

to different follow-up times at the time of the overview, and different patient selection criteria. The ratios of the odds of recurrence in the two arms also vary (see also Figure 4.4 below). Modeling this variation is the main focus of this section.

4.4.2 Modeling heterogeneity

One approach to analyzing several 2×2 tables is to first perform a preliminary test for heterogeneity. If the null hypothesis (of homogeneity) is not rejected, it is common to proceed with analyses that effectively pool patients as though they all belonged to the same trial. If the null hypothesis of homogeneity is rejected, it is common to declare the studies too dissimilar to be combined, and stop. This general two-step approach can be implemented using both Bayesian and frequentist tools. For example, classical tests for heterogeneity are reviewed by Sutton *et al.* (1998), while a Bayesian approach using Bayes factors is proposed by Berry (1999).

When the meta-analysis is carried out to support decision making, this approach has several limitations. One problem is that, unless the number of trials is very large, tests have insufficient power to detect heterogeneity. Also, when the trials are sufficiently similar in clinical design, a moderate amount of heterogeneity can be desirable, because it can give a sense of how well the conclusion of trials can be extrapolated to other clinical settings and populations. The relevant quantities for clinical decisions refer to predictions for future patients. How will these be affected by the choice of a homogeneity versus a heterogeneity analysis? If heterogeneity is indeed small, the results will not deviate much. But the larger the heterogeneity, the more important it is to acknowledge it. So in an analysis whose goal is to assist clinical decision making, the scientific modeling issue is not so much whether the trials are homogeneous or not. In the vast majority of meta-analyses they are not. The practical issue is whether there is enough *prima facie* evidence of heterogeneity to justify the small additional trouble of implementing a heterogeneity model.

A step toward quantifying the heterogeneity of trials, and incorporating this into decision making if necessary, is to use a model with two levels: one describes the population of trials; the other describes the subpopulations of patients within each trial. These multilevel models are usually called hierarchical, because of the nesting of patients within trials (see also Figure 4.1 below). To set up a two-level hierarchical model for the tamoxifen example, we begin with some notation. Studies will be indexed by s. The sample sizes in the control and treatment arms of study s are n_0^s and n_1^s, respectively. The numbers of observed recurrences x_0^s and x_1^s. All trials are randomized, and therefore the sample sizes carry no information about efficacy. In study s, the probability distribution of each 2×2 table is completely characterized by two parameters: the probability of recurrence in the control arm (π_0^s) and in the treatment arm (π_1^s).

It is useful to reparameterize these into a measure of baseline recurrence rate (π_0^s) and a measure of the efficacy of treatment. Two measures of the efficacy of treatment are the log odds ratios

$$\lambda^s = \log\left(\frac{\pi_1^s}{1-\pi_1^s}\right) - \log\left(\frac{\pi_0^s}{1-\pi_0^s}\right), \qquad s = 1, \ldots, 14,$$

and the log relative risks

$$\log\left(\frac{\pi_1^s}{\pi_0^s}\right), \qquad s = 1, \ldots, 14;$$

both of these become smaller the more effective the treatment. In our discussion we will model the log odds ratios λ^s. Approaches based on modeling the log relative risks can be carried out along the same lines and require only small modifications to the procedure. One additional difficulty is that for any given baseline recurrence rate π_0^s, the range of values for the relative risk is constrained: the relative risk cannot be greater than $1/\pi_0^s$. In parametric models, the choice of parameterization can be important, especially when the number of studies is not large. One choice criterion is interpretability. For example, it may be more natural to specify distributional assumptions or to elicit expert opinion on one parameterization.

Let us consider the (π_0^s, λ^s) parameterization. The log odds ratios λ^s are likely to vary from study to study, but on the other hand are also likely to reflect the actual underlying efficacy of tamoxifen. Knowing the value of the log odds ratios in one study is informative about log odds ratios in other studies, despite heterogeneity. Assuming that all log odds ratio are the same (as assumed by the fixed effects models) ignores heterogeneity. Assuming that they are independent would ignore a key commonality. We need a compromise between independence and equality.

One way to achieve this compromise is to postulate a hypothetical population of studies, each with a different log odds ratio λ^s, and then learn about the features of this population from the data. We will discuss the simple case $\lambda^s \sim N(\theta, \tau^2)$. Here θ represents the overall mean log odds ratio and plays a role similar to the common log odds ratio in a model without heterogeneity; τ measures the study-to-study variability in the log odds ratios. Both θ and τ will be unknown and inferred from the data. In this way, it is the observed data that help us decide where we should position our analysis in the continuum between independence and equality of the effects. Alternative distributional assumptions are also possible. For example, assuming that the λ^s have a Student t or a double exponential distribution would result in inferences that are more robust to unusually large or small effects.

In the tamoxifen example, knowing the baseline rate π_0^s of one study is unlikely to inform us about the baseline rate of other studies. The determinants of the baseline rates are likely to be more strongly related

to study design and current follow-up than to underlying clinical factors. Therefore, it can be reasonable to assume that the π_0^s, unlike the λ^s, are independent of each other, and forgo modeling a second-stage distribution. In other examples it may be appropriate to model a second-stage distributions for both the λ^s and the π_0^s.

In this analysis we are assuming that studies are conditionally independent given θ and τ. A consequence of this is that we are assuming that the studies are exchangeable, or in other words that that there are no study features that can help us predict whether the log odds ratios in a study is more or less likely to be, say, larger than average. A challenge to this assumption comes from the possibility that the log odds ratios might be correlated with baseline rates across studies. A simple diagnostic for this assumption is to plot empirical log odds ratios versus empirical recurrence probabilities in the control arms, and look for relationships. For example, if the treatment began to take effect only after a certain number of months, we could observe a positive relationship, with trials having longer follow-up displaying both a higher mortality and larger effect. In our case, the plot reveals no relationship.

Figure 4.1 summarizes the interconnections among the variables in the model. All variables are probabilistically dependent. However, the model is specified via conditional independence assumptions that simplify model-building, interpretation and computing. For example, the data in study 1 are conditionally independent of θ and τ given λ^1. However, via λ^1, the data in study 1 do provide information about θ and τ. In general, unknown parameters can be thought of as vehicles for information to percolate across the model. If the value of λ^1 were revealed to us by an oracle, the data in study 1 would not provide any information about θ and τ.

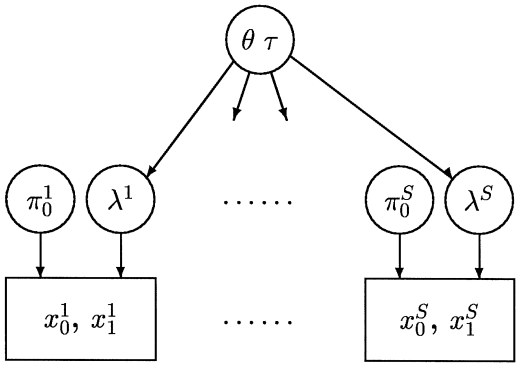

Figure 4.1 Graphical summary of the relationships among the parameters and observations in the hierarchical model considered in this section. Circles represent unknown parameters; rectangles represent data. Links represent probabilistic dependence relationships. The parameters τ and θ describe the population of study-specific log odds ratios. Each study also has a baseline recurrence rate π_0^s, but in this model these are independent of each other. This is reflected in the fact that they are not linked by belonging to a higher-level population.

The Bayesian model is completed by prior distributions on the π_0^s, θ, and τ. Our overall strategy here is to specify a vague prior distribution on the π_0^s and θ, which are well estimated based on the data, and a more informative prior distribution for τ. All priors are proper, in that they yield 1 when integrated over all possible parameter values. In the case of τ, it is important to specify a proper prior, as familiar choices of noninformative improper priors may lead to improper posterior distributions (DuMouchel, 1990).

More specifically, we assume that the π_0^s are independent and uniformly distributed over $(0, 1)$, and that θ follows a normal distribution with large standard deviation (say, 10) and mean zero. Because of the large standard deviation, the choice of the prior mean is unimportant. Because τ reflects sources of variability that are likely to be common to other groups of similarly heterogeneous trials, a possible strategy is to resort to other meta-analyses to gather a priori information about the likely magnitude of τ. Smith *et al.* (2000) illustrate this technique using a large collection of unrelated meta-analyses. In our case, because the focus is solely on the trials of Table 4.1, we could consider two other groups of trials of tamoxifen for the same type of patients, where tamoxifen was administered for two and five years instead of one. One caveat is that the heterogeneity may increase with the magnitude of the effect, so that heterogeneity of trials with longer treatment may be slightly higher.

A convenient choice of functional form for the distribution of τ^2 is to specify an inverse gamma distribution. We can use the parameterization with density function $f(x) = b^a x^{-a-1} e^{-b/x}/\Gamma(a)$ for $x > 0$. The mean is $b/(a-1)$ and the variance $b^2/(a-1)^2(a-2)$. We selected hyperparameters $a = 3$, the integer value giving the most diffuse finite-variance distribution, and $b = 0.1$, to approximately match the dispersion from the other groups of trials. One way of interpreting this prior specification in the simpler, nonhierarchical, problem is that it roughly corresponds to the information provided by $a = 3$ previous observations whose sum of squared deviations from the mean is $b = 0.1$. Specifying $a = 3$ limits the sensitivity of the analysis to the specified value of b, which is likely to be overridden by the experimental information. Because reasonable choices of prior distributions for τ can lead to different results, a sensitivity analysis, comparing the results obtained under different priors, is generally useful. While we do not illustrate it in this section, we will return to this point in Section 4.6.

In summary, we specified the following model:

$$
\begin{array}{lll}
\text{PRIOR} & \theta \sim N(0, 100) & \\
 & \tau^2 \sim \text{IG}(3, 0.1) & \\
\text{STUDIES} & \pi_0^s \sim U(0,1) & s = 1, \ldots, S \\
 & \lambda^s | \theta, \tau^2 \sim N(\theta, \tau^2) & \\
\text{PATIENTS} & x_0^s | \pi_0^s \sim \text{Bin}(\pi_0^s, n_0^s) & \\
 & x_1^s | \pi_1^s \sim \text{Bin}(\pi_1^s, n_1^s) &
\end{array}
$$

There are two parameters of primary interest. The first is the mean log odds ratio θ in the population of trials. Inference about θ addresses the question of the size of is the effect shown by the trials. The second is the log odds ratio λ^{S+1} in a future, unobserved, or hypothetical trial based on the available population of trials. The latter predictive distribution is directly relevant for clinical decisions. In particular, a possible future trial is a single woman who is faced with the decision about tamoxifen as an adjuvant therapy for early breast cancer. The benefit she will receive can be regarded as the next observation from the population of trial benefits.

Models for combining 2×2 tables in meta-analysis while acknowledging heterogeneity have been proposed by several authors. A seminal paper in this area is DerSimonian and Laird (1986). There are also Bayesian hierarchical models that share a similar overall structure and goals with the model of this section. Carlin (1992) and Smith *et al.* (2000) reparameterize the π^s into the corresponding logits, and assign normal conjugate distributions to them; Berry (1998) considers a Poisson, rather than binomial, sampling scheme, and models the event rates (or hazard rates) in the control group using a gamma with unknown parameters, and the log of the ratio of hazard rates in the two groups by a normal. A significant difference between the strategy used here and all these is that while the efficacy parameters (log odds ratios) are modeled hierarchically, here the baseline event rates are not.

4.4.3 Computing

To evaluate the posterior and predictive distributions we will use an MCMC algorithm, sampling in turn from τ, θ, and the pairs (π_0^s, λ^s). The full conditional distributions are

$$\theta \sim N\left(\left(\frac{1}{\tau^2}\sum_s \lambda^s\right)\left(\frac{1}{100}+\frac{S}{\tau^2}\right)^{-1}, \left(\frac{1}{100}+\frac{S}{\tau^2}\right)^{-1}\right)$$

$$\tau^2 \sim \mathrm{IG}\left(3+\frac{S}{2}, 0.1+\frac{1}{2}\sum_s(\lambda^s-\theta)^2\right)$$

$$(\pi_0^s, \lambda^s) \sim K(\pi_0^s)^{x_0^s}(1-\pi_0^s)^{n_0^s-x_0^s}\left(1+\frac{1-\pi_0^s}{\pi_0^s}e^{-\lambda^s}\right)^{-n_1^s}\left(\frac{1-\pi_0^s}{\pi_0^s}e^{-\lambda^s}\right)^{n_1^s-x_1^s} \times e^{-\frac{1}{2\tau^2}(\lambda^s-\theta)^2},$$

where K is the normalizing constant.

The first two full conditional distributions are the same as the posterior distributions for the mean of a normal population with known variance and for the variance of a normal population with known mean. These can be sampled directly. General expressions which do not depend on the specific hyperparameters chosen here, and details of the derivations, can be found in (Bernardo and Smith, 1994).

The full conditional distributions of the recurrence probabilities are not immediately recognizable. To sample from them we can use a Metropolis–Hastings step within the MCMC method (Tierney, 1994). There are many ways to implement this step. We used the strategy of approximating the conditional covariance matrix of (π_0^s, λ^s) and using it in a symmetric random walk Metropolis algorithm (see Section 3.2.3). First, we approximated the binomial component of the full conditional (the top line) with a bivariate normal. Using the central limit theorem, the statistics x_t^s/n_t^s are independent with limiting distributions

$$\sqrt{n_t^s}\left(\frac{x_t^s}{n_t^s} - \pi_t^s\right) \to N(0, \pi_t^s(1-\pi_t^s)), \qquad t = 0, 1.$$

To simplify the presentation we denote this by

$$\frac{x_t^s}{n_t^s} \sim N\left(\pi_t^s, \frac{\pi_t^s(1-\pi_t^s)}{n_t^s}\right), \qquad t = 0, 1.$$

Using the multivariate Delta method (described in a similar application in (Bishop *et al.*, 1975, pp. 486ff.), we can transform the implied joint distribution into an approximate distribution for $x_0^s/n_0^s, \hat{\lambda}^s$, where $\hat{\lambda}^s = \log(x_1^s/(n_1^s - x_1^s)) - \log(x_0^s/(n_0^s - x_0^s))$ is the maximum likelihood estimate of the log odds ratio λ^s based on the study-specific information alone. This asymptotic distribution is

$$\begin{bmatrix} \frac{x_t^s}{n_t^s} \\ \hat{\lambda}^s \end{bmatrix} \sim N\left(\begin{bmatrix} \pi_t^s \\ \lambda^s \end{bmatrix}, \begin{bmatrix} \frac{\pi_0^s(1-\pi_0^s)}{n_0^s} & -\frac{1}{n_0^s} \\ -\frac{1}{n_0^s} & \frac{1}{\pi_0^s n_0^s} + \frac{1}{(1-\pi_0^s)n_0^s} + \frac{1}{\pi_1^s n_1^s} + \frac{1}{(1-\pi_1^s)n_1^s} \end{bmatrix}\right) \equiv N(m_0, V_0).$$

Substituting this approximate normal likelihood for the binomial terms into the full conditional distribution of (π_0^s, λ^s) leads to the product of two normal terms. Replacing V_0 with a point estimate $\hat{V}_0$ and combining the quadratic forms on the exponents of these terms, we obtain a normal approximation to the full conditional. The variance of this approximation is

$$V_1 = \left(\hat{V}_0^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{\tau^2} \end{bmatrix}\right)^{-1},$$

while the mean is

$$m_1 = \hat{V}_0^{-1} m_0 + \begin{bmatrix} 0 \\ \frac{\theta}{\tau^2} \end{bmatrix}.$$

In general, $\hat{V}_0$ can be estimated by the study-specific maximum likelihood, that is, by substituting empirical frequencies for recurrence probabilities. In this example, however, $\hat{V}_0$ was obtained by adding 1 to each count before computing the empirical frequencies. In this way, probabilities on each arm

```
model;
{
   for( i in 1 : Num ) {
      x0[i] ~ dbin(p0[i],n0[i]) ;
      x1[i] ~ dbin(p1[i],n1[i]) ;
      p0[i] ~ dunif(0,1);
      logit(p1[i]) <- logit(p0[i])+lambda[i] ;
      lambda[i] ~ dnorm(theta,tau);
   }
   theta ~ dnorm(0.0, 0.01);
   tau ~ dgamma(3, 0.1);
}
```

Figure 4.2 BUGS command-line code for the model specification in Section 4.4.2.

are approximated by $(x + 1)/(n + 2)$. This is a somewhat *ad hoc* procedure, but it is useful for avoiding infinite variance estimates for the log odds when no recurrences are observed. There is a Bayesian interpretation for it: $(x+1)/(n+2)$ is the posterior mean in a single-study analysis when assuming a uniform prior distribution on the recurrence probabilities.

This approximation to the full conditional distribution provides an excellent fit for studies with larger sample sizes, but is less accurate for small studies, and turns out to be quite poor in study 13, which only has 8 and 10 patients in the two arms. If all studies were large, we could use this distribution to develop an independent proposal Metropolis–Hastings sampler, or even use the approximation directly in lieu of the actual full conditional. In this data set, however, it is safer to implement a symmetric random walk proposal, in which a proposed value is obtained by perturbing the current value. Our approximation is still crucial to determining the covariance matrix of the perturbation.

When utilizing a random walk proposal, we have the flexibility of using a variety of symmetric distributions to generate a new parameter value. In this case, two natural choices are the normal and Student's t. We used the latter here, to allow for better sampling of the tails of the full conditional distributions in studies with small sample sizes. Student's t is centered at the current parameter value, has scale matrix dV_1, and degrees of freedom given by the smallest sample size in all study arms (in our case 8). d is a scale factor that needs to be adjusted based on the application to hand. Values between 1 and 4 should provide good performance and are unlikely to need adjustment.

The model considered here, as well as many other useful models for combining 2×2 tables can be fit using the software package BUGS, mentioned in Section 3.2.3. Figures 4.2 and 4.3 show how to enter the data and how to

```
list( x1 = c( 55, 137, 505,  62,  99,  50, 185,
              186, 148, 25, 223, 183,  2, 129),
      n1 = c( 55, 137, 505,  62,  99,  50, 185,
              186, 148, 25, 223, 183,  2, 129),
      x0 = c( 67, 187, 590,  74, 118,  49, 200,
              187, 178, 38, 224, 185,  0, 159),
      n0 = c(101, 306, 915, 140, 236, 107, 319,
              307, 325, 86, 350, 936,  8, 449),
      Num=14)

list(lambda = c(0,0,0,0,0,0,0,0,0,0,0) ,
     tau = 1,
     theta = 0,
     p0 = c(0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5))
```

Figure 4.3 BUGS command-line code for the dataset of Table 4.1, to be used in conjunction with the model specification in Figure 4.2.

represent the model of this section. In this framework it is also straightforward to explore the sensitivity of results to alternative choices of distributions for the λs. BUGS can handle, for example, the double exponential and Student's t.

4.4.4 Results

Results are summarized in Figure 4.4 in terms of study-specific relative risks. A similar figure in terms of log odds ratios can be produced using the same MCMC output. Horizontal bars represent 98% probability intervals. Their width reflects the accuracy with which relative risks can be estimated within each study. For example, even though studies 3 and 12 have approximately the same sample size, the posterior probability distributions of the study-specific relative risks have very different intervals. This is the result of their different baseline event rates: study 3 has an event rate closer to 1/2. Studies with wider intervals contribute less to the overall conclusions. One way to see this is to realize that their relative risks will fluctuate more widely in the simulation.

Within each study, the relative risk is π_1^s/π_0^s. For the next trial we can derive the predictive distribution of the log odds ratios directly from the model. We then convert this distribution into a distribution for the relative risk by fixing the baseline rate (in this example to 1/2) and performing a change of variables. In practical situations, the baseline rates of an adverse event may be known based on patient prognostic factors. Then the predictive distribution of the absolute recurrence rate π_1^s can be derived from the baseline rates and the distribution of log odds ratios or relative risks. If the baseline rate is not known exactly, but a probability distribution is available, the same analysis applies with straightforward modifications.

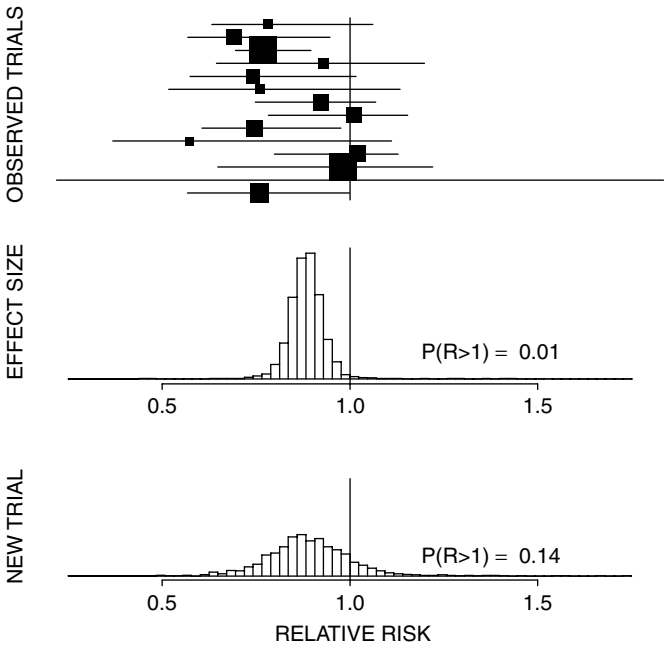

Figure 4.4 Summary of meta-analysis results for the tamoxifen example. The horizontal scale is relative risk, that is, the ratio of recurrence rates in the treatment to the control group. Values smaller than one correspond to a beneficial treatment. Squares in the top panel represent empirical relative risks. Their area is proportional to the overall sample size in the trial. Horizontal lines cover 98% of the probability distributions of π_1^s/π_0^s, with 1% of the mass removed on either side. The interval for study 13 is wider than the chosen limits of the figure. The middle panel is the posterior distribution of the overall relative risk, assuming a baseline recurrence rate of 50%. The bottom panel is the predictive distribution of the relative risk in a future trial, assuming again a baseline recurrence rate of 50%. The probability that the average effect in a trial is less than one is less than 0.99, but the probability of a benefit in a future patient is 0.86. Both probabilities are independent of the assumed baseline rate.

It is interesting to contrast predictions for a future patient with inferences about the average effect size. The probability of a benefit in a future patient is 0.84. (This is independent of the baseline rate, which only affects the magnitude of the relative risk.) On the other hand, the probability that the average relative risk in the trials is less than unity is 0.99. The posterior mean of both θ and λ^{s+1} is -0.2. However, the posterior standard deviations are respectively 0.13 and 0.45.

This example illustrates how two-stage hierarchical models can be used to address study heterogeneity, one of the most relevant criticisms of meta-analysis, and to develop probability distributions for the quantities that are relevant in clinical decision making, that is, predictions for future

patients. Ignoring study heterogeneity can be deleterious, and modeling heterogeneity is not difficult. Because conclusions can be sensitive to the prior distribution of τ, modeling heterogeneity requires careful consideration of what is a priori a reasonable amount of heterogeneity, based on past experience. For an example in breast cancer screening, see Berry (1998). We will return to this issue later in this chapter when we discuss sensitivity analysis.

4.5 Combining studies with continuous and dichotomous responses

A challenging problem in meta-analysis occurs when study responses, while similar, are not directly comparable. For example, some studies may report continuous outcomes, say the change in a functionality scale, while other studies consider a binary representation of a similar response, for example by recording whether there is an improvement in patient functionality. A simple approach in this situation is to dichotomize the continuous responses and proceed as in the simpler all-binary case, for example as in Section 4.4. This approach is practical, but it has limitations: for example, the choice of cutoff point for the dichotomization may be arbitrary, and a loss of information will occur in the dichotomization of the continuous response. The case study in this chapter will call for a strategy that overcomes both of these difficulties. Hasselblad and Hedges (1995) propose continuous scales based on the logits of the observed frequencies in the binary studies. These overcome the discretization problem in a nice way.

Here we discuss a Bayesian alternative, which has the added advantages of fitting conveniently within a hierarchical framework and leading to a simple solution to the problem of ranking treatments. In this section I present the Bayesian strategy in a simple case. The discussion is based on Dominici and Parmigiani (2000). The key step is to think of the binary responses as the result of dichotomizing some underlying unobserved continuous response. For example, whether there is an improvement in patient functionality can be approximately though of as a dichotomization of the change on a functionality scale, with yes corresponding to a positive change. The continuous scale can be used as a common underlying scale for all the studies. In this way we preserve the full information from the studies reporting continuous variables, we do not need to choose any arbitrary cutoff points, and we incorporate uncertainty arising from the heterogeneous nature of the responses. Technically, there is a problem in that this scale is missing in some studies and will have to be inferred from the observed binary response. But, as we have seen, MCMC methods are well suited to this task. Simulation-based methods for handling uncertainty about missing data in Bayesian analysis are discussed in detail by Tanner (1991) and Gelman *et al.* (1995).

4.5.1 A simulated example

By way of a simple concrete illustration of this idea, we now consider a simulated example, chosen to highlight the differences between the latent variable approach and alternative analyses that dichotomize the first two studies and then proceed as in the all-binary case. Consider four studies, indexed by s, each comparing a treatment arm ($t = 1$) with a placebo (or control) arm ($t = 0$). For each study, 20 observations were simulated in each arm. Studies 1 and 2 record a continuous response, while studies 3 and 4 record a binary response. Summaries of simulated data from the four studies are shown in Table 4.2.

We will use the notation y^s_{ti} for the observed response of the ith individual ($i = 1, \ldots, 20$) assigned to arm t in studies 1 and 2, and the notation x^s_{ti} for the binary outcomes corresponding to the observation in studies 3 and 4. To build a simple model for these data, assume that the continuous responses are normally distributed, that is

$$y^s_{ti} \mid \theta_t, \sigma^2 \sim N(\theta_t, \sigma^2), \quad t = 0, 1, \quad s = 1, 2,$$

where θ_0 and θ_1 are the placebo and treatment population means, and σ^2 is a common population variance. For studies 3 and 4, introduce the latent variables y^s_{ti}, where

$$y^s_{ti} \mid \theta_t, \sigma^2 \sim N(\theta_t, \sigma^2), \quad t = 0, 1, \quad s = 3, 4,$$

and $x^s_{ti} = 1$ whenever $y^s_{ti} > 0$. Therefore the binary observations have sampling distribution

$$P\{x^s_{ti} = 1 \mid \theta_t, \sigma^2\} = 1 - \Phi\left(-\frac{\theta_t}{\sigma}\right), \quad t = 0, 1, \quad s = 3, 4,$$

Table 4.2 Summary statistics from the simulated data with continuous and dichotomous responses. $\bar{y}^s_t$ and v^s_t are the sample mean and variance in arm t of study s.

Study				
	$\bar{y}^s_0$	v^s_0	$\bar{y}^s_1$	v^s_0
1	0.0218	0.0087	−0.0162	0.0075
2	0.0015	0.0103	0.1119	0.0105
	$\sum_i x^s_{0i}$	$\sum_i x^s_{1i}$		
3	12	8		
4	10	17		

where Φ is the standard normal cdf. It is convenient to indicate by $y_t^s = (y_{t1}^s, \ldots, y_{t20}^s)$ the vectors of continuous patient measurements in arm t of study s. In studies 3 and 4 this is not observed directly.

If the binary responses arise in practice as a dichotomization of the continuous responses, this model describes directly the data generation mechanism. Otherwise, the latent variables can be interpreted as hypothetical continuous outcomes consistent with the observed discretized outcomes. Using latent normal variables to model binary observation is discussed in greater generality by Carlin and Polson (1992) and Chib and Greenberg (1998). Because the point of this section is to illustrate the latent variables approach, we consider a fixed effects model, ignoring for the moment potential study heterogeneity.

When continuous and discrete responses are believed to differ systematically, an offset parameter α could be added to the model. The distribution of the continuous variables in studies 1 and 2 could be specified as

$$y_{ti}^s \mid \theta_t, \sigma^2 \sim N(\theta_t + \alpha, \sigma^2), \quad t = 0, 1, \quad s = 3, 4,$$

where α can be interpreted as the difference in measured efficacy between the two types of response. In order better to focus on the mechanics of the latent variable approach, we do not consider this case here.

In the model we have specified, all parameters are identified. In more general formulations, one may consider both treatment effects and variances to be study-specific. However, in studies with binary responses, we cannot make inference on both the mean and the variance of the underlying latent variables from the observed binary xs. Therefore, identifying restrictions are necessary, for example specifying a common variance σ^2 in all the studies.

Our Bayesian formulation is completed by specifying prior distributions for θ_0, θ_1, and σ^2. In the absence of more specific information, we choose conjugate, vague priors: $\theta_t \overset{\text{ind}}{\sim} N(0, 25)$, $t = 0, 1$, and $\sigma^2 \sim IG(a, b)$, independent of the θs. We use the parameterization of the inverse gamma with density function $f(x) = b^a x^{-a-1} e^{-b/x}/\Gamma(a)$ for $x > 0$, with mean $b/(a-1)$ and variance $b^2/(a-1)^2(a-2)$. We select hyperparameters $a = 3$, the integer value giving the most diffuse finite-variance prior distribution, and $b = 0.5$, selected to produce a mean of 1.

We can draw inferences on any of the unknowns using the joint posterior distribution

$$p(\theta_0, \theta_1, y_0^3, y_1^3, y_0^4, y_1^4 \mid \text{data}), \tag{4.1}$$

where data $= \left(y_0^1, y_1^1, y_0^2, y_1^2\right)$. Neither (4.1) nor its marginals are available in closed form. A practical choice for determining the marginal distribution, marginal probabilities, and other summaries of interest is again to draw a sample of values from (4.1). This can be done using a Gibbs sampler (Gelfand and Smith, 1990), based on partitioning the unknown parameters and missing

data into groups and sampling each group in turn, given all the others. Observe that, if the y_{ti}^s are known for $s = 3, 4$, then posterior inference and simulation can be obtained easily using routine normal theory (Bernardo and Smith, 1994). The y_{ti}^s are, of course, unknown; however, given the data x_{ti}^s, the conditional distributions of y_{ti}^s given $x_{ti}^s = 1$ (0) are normal $N(\theta_t, \sigma^2)$ truncated to the positive (negative) values. With this in mind, we can write the full conditional distributions as follows. For the latent variables in studies 3 and 4:

$$\begin{aligned} y_{ti}^s \mid x_{ti}^s, \theta_t, \sigma^2, \text{data} &\sim N(\theta_t, \sigma^2) I_{y_{ti}^s > 0}(y_{ti}^s) \qquad \text{for } i : x_{ti}^s = 1, \\ y_{ti}^s \mid x_{ti}^s, \theta_t, \sigma^2, \text{data} &\sim N(\theta_t, \sigma^2) I_{y_{ti}^s \le 0}(y_{ti}^s) \qquad \text{for } i : x_{ti}^s = 0. \end{aligned}$$

For the parameters:

$$\begin{aligned} \theta_t \mid \sigma^2, \text{data},\ y_0^3, y_1^3, y_0^4, y_1^4 &\sim N(a_1, s_1^2) \\ \sigma^2 \mid \theta_0, \theta_1, \text{data},\ y_0^3, y_1^3, y_0^4, y_1^4 &\sim \text{IG}(a_1, b_1) \end{aligned}$$

where $t = 0, 1$ and

$$\begin{aligned} a_1 &= \left[\frac{1}{\sigma^2} \sum_{s=1}^{4} n_t^s + \frac{1}{s^2} \right]^{-1} \left[\frac{1}{\sigma^2} \sum_{s=1}^{4} n_t^s \bar{Y}_t^s + \frac{a}{s^2} \right], \\ s_1^2 &= \left[\frac{1}{\sigma^2} \sum_{s=1}^{4} n_t^s + \frac{1}{s^2} \right]^{-1}, \\ a_1 &= a + \sum_{s=1}^{4} \sum_{t} n_t^s, \\ b_1 &= b + \frac{1}{2} \sum_{s=1}^{4} \sum_{t} \sum_{i=1}^{n_t^s} (y_{ti}^s - \theta_t)^2, \end{aligned}$$

where n_t^s is the sample size in arm t of study s.

Implementing the Gibbs sampler above, we can use sampled values from the chain to estimate the marginal densities of quantities of interests, such as the effect size difference $\theta_1 - \theta_0$. The posterior probability of a negative effect size difference, that is, $P(\theta_1 - \theta_0 \le 0 \mid \text{data})$, is 0.021, while the 95% probability interval is $(0.004, 0.181)$.

It is interesting to contrast these results with those obtained by dichotomizing the continuous variables at 0 and then combining studies, which leads to a two-sample comparison of proportions. Our prior specification for the θs and σ^2 implies a U-shaped, approximately independent prior on such proportions. Replacing this prior with a product of noninformative priors of the form $1/p(1-p)$, the joint posterior on the proportions is approximated by a product of beta densities. The probability that the proportion of successes is greater in the treatment group than in the placebo is 0.076. The 95% equal-tail probability interval for the difference in proportions is $(-0.041, 0.263)$,

which includes 0. While the two priors are not equivalent, they are sufficiently similar that we can attribute much of the difference in tail probabilities to the loss of efficiency resulting from the dichotomization of the observed continuous data. Similarly, the p-value of the two-sided Mantel–Haenszel test obtained by dichotomizing studies 1 and 2 with cutoff point at 0 is 0.1936. Therefore, at a level of 0.05, the latent variables approach leads to the conclusion that there is a significant difference between the treatment and the placebo effect, while the dichotomized approach, either in the Bayesian or frequentist formulation, does not. While this difference in testing outcomes naturally depends on the specific settings of the simulated example, the loss of efficiency it illustrates is general.

4.6 Migraine headache

4.6.1 Background and goals

Chronic migraine headache is a common condition, it is difficult to treat, and it has substantial impact on public health and productivity (Ziegler, 1990; Lipton and Stewart, 1993). A recent study of prevalence in the US population concluded that '8.7 million females and 2.6 million males suffer from migraine headache with moderate to severe disability. Of these, 3.4 million females and 1.1 million males experience one or more attacks per month' (Stewart *et al.*, 1992). A wide range of drug and nondrug headache treatments are available, and there is wide disagreement about which are most effective (Pryse-Phyllips *et al.*, 1997). The recent consensus conference of the Canadian Medical Association (Pryse-Phyllips *et al.*, 1997) emphasized that migraine headache continues to be inadequately managed. Problems are particularly severe for prophylactic treatment, where a plethora of small clinical studies and trials offer conflicting evidence about treatment efficacy.

The Agency for Health Care Policy and Research (AHCPR) supported a large-scale systematic review of various categories of headache treatment. A team of clinicians reviewed the literature to select and abstract studies for a comprehensive evidence report. In our discussion we will focus on two types of drug therapy for migraine headache: (beta-blockers, and calcium-channel blockers. We draw data from the evidence reports by Goslin *et al.* (1998a; 1998b), which include 19 studies of four beta-blockers and eleven studies of four calcium-channel blockers. The drugs and their abbreviations are summarized in Table 4.3. Each study includes two or three treatment groups, so no study compares all the drugs directly. In 13 out of the 19 beta-blockers studies, and in 9 out of the 11 calcium-channel blockers studies, one of the experimental groups is given a placebo. The other studies consider drug-to-drug comparisons.

A meta-analysis of clinical trials can help bring existing evidence to bear on the controversy about which treatments are most effective. Often, because

Table 4.3 Summary of abbreviations for the migraine headache drugs examined in Section 4.6.

Calcium-channel blockers		Beta-blockers	
CB0	Placebo	BB0	Placebo
CB1	Verapamil	BB1	Propanolol
CB2	Nifedipine	BB2	Timolol
CB3	Flunarizine	BB3	Nadolol
CB4	Nimodipine	BB4	Metoprolol

of the relatively low response rate to prophylaxis, several treatments are prescribed in sequence to the same patient. To support better-informed choices of treatment sequences, it is important to provide a ranking of treatments, and uncertainty statements about this ranking. Also, despite the large number of published studies on the subject, relatively few medications for prophylactic treatment of migraine headache have been subjected to adequate clinical trial (Pryse-Phyllips *et al.*, 1997), and new trials will be necessary to resolve many of the remaining uncertainties. In our discussion we will present a model whose goal is to support clinical treatment decisions, and potentially help guide the planning of new trials. The critical statistical aspects of these goals are the synthesis of this complex and diverse information, the estimation of treatment effects on a common scale, and the relative ranking of treatments, both within classes and overall.

In developing a statistical model to address these goals we face several challenges. First, because individual studies only include a subset of treatments, and do not always include a placebo, we need to make indirect comparisons among treatments that may never have been tested together in the same trial.

Second, while most studies report continuous treatment effects for each treatment, some report only differences in effectiveness for pairs of treatments and others report only 2×2 contingency tables for dichotomized responses. Also, all the published responses that we have available have been previously standardized by dividing observed treatment differences by the estimated population subject-to-subject standard deviation within each study, to give them a common dimensionless scale (the primary sources reported effects on a variety of scales, including ordinal measures of well-being).

Finally, studies differ in the modalities of administration of the treatments, and in the criteria for defining migraine headache and admitting patients. On the other hand, the fact that the drugs' mechanisms are similar suggests that treatment effects are likely to be as well. Also, similarity is likely to be stronger within treatment classes than across classes. Because of the sparseness of the

study/treatment matrix, and the relatively small size of some of the studies, it is important to exploit this similarity to 'borrow strength' from other studies in estimating each effect – borrowing more heavily from treatments of the same class.

Here we present an analysis that addresses these challenges using a combination of two techniques: hierarchical Bayesian modeling and latent variables modeling. Hierarchical modeling is used to model study heterogeneity and differential borrowing of strength within and across drug classes. In view of the previous paragraph), and the absence of study-specific covariate information to quantify effect variations, both studies and treatments present features that would make a traditional fixed effects model inadequate for the analysis. As we have seen in Section 4.4, hierarchical models offer a convenient way of modeling study heterogeneity. Latent variables modeling is used to account for study results having been standardized, some studies reporting only differences between treatments, and some studies reporting only dichotomized responses. We create a latent scale common to all studies for combining information from studies that report results in different forms. Our strategy will be similar in spirit to that of Section 4.5.

This model permits us to synthesize this heterogeneous information and to make inferences about treatment effects and the relative ranks of treatments, without ignoring important components of uncertainty. Estimation, ranking, model validation, and sensitivity analysis are all implemented through simulation-based methods. Our analysis is based on Dominici *et al.* (1999), who present a model sharing the same principles and motivation. Compared to what is presented here, their model is more complex in that it includes a third group of treatments, the biofeedback treatments. These are not drug treatments, and the borrowing of strength from the treatment effects requires an additional level in the hierarchical structure. Our discussion also differs in a number of small details of model, prior, and sensitivity analysis specification.

4.6.2 Data

The study effects and standard deviations from the 30 trials selected in the AHCPR preliminary evidence report (Goslin *et al.*, 1998a; 1998b) are summarized in Table 4.4. Despite the relatively large number of trials, the data are sparse. Four of the treatments only appear in two trials.

The treatments under consideration are indexed by $t = 0, \ldots, 8$; $t = 0$ denotes the placebo groups in both drug classes. $\mathcal{T}_{\text{bb}} = \{1, \ldots, 4\}$ is the set of indexes for the beta–blockers and $\mathcal{T}_{\text{cb}} = \{5, \ldots, 9\}$ is the set of indexes for the calcium-channel blockers, both in the same order as in Table 4.4. The studies are indexed by $s = 1, \ldots, 30$, again in the order of Table 4.4. Each (t, s) combination is referred to as an arm. For each $s \in \mathcal{S}$, let $\mathcal{T}^s \subset \mathcal{T}$ be the set of treatments included in study s. For example, from rows 3 and 20 of Table 4.4, we see that $\mathcal{T}^3 = \{0, 1, 2\}$ and $\mathcal{T}^{20} = \{0, 7\}$.

Table 4.4 Reported study results for the 30 studies. Rows correspond to studies. Columns in each subtable correspond to treatments, with beta blockers in the top table and calcium-channel blockers in the bottom table. Entries in the tables are observed standardized treatment effects z. In parentheses are the standard deviations $\sqrt{q}$. A $\circ$ indicates a crossover study. A $\star$ indicates studies reporting 2×2 tables. In study 5 there were 0 successes and 8 failures in the placebo group, and 6 successes and 14 failures in the treatment group. In Study 23 there were 14 successes and 15 failures in the placebo group, and 23 successes and 6 failures in the treatment group.

Study	n_s	BB0	BB1	BB2	BB3	BB4
1	52	−1.262 (1.3)	−0.687 (0.56)			
2	65	−0.016 (0.98)				0.508 (1.03)
3	240	−1.204 (1.21)	−0.874 (0.88)	−0.725 (0.87)		
4	58		−0.79 (0.92)		−1.34 (1.06)	
5	28	★			★	
6	94	−0.425 (1.03)		0.425 (0.97)		
7	32	○				0.497 (1) ○
8	40		−0.341 (1.02)			0.308 (0.98)
9	14	○		0.497 (1) ○		
10	64	−1.516 (1.1)	−0.953 (0.89)			
11	67		−1.611 (1.04)			−1.462 (0.95)
12	27		0.146 (1.03)		−0.158 (0.97)	
13	56		○			0.929 (1) ○
14	56	−2.033 (1)	−1.513 (1)			
15	55	−0.241 (1.28)	1.74 (0.71)			
16	34	−1.314 (1.18)	−0.902 (0.78)			
17	40		0.501 (1.11)			−0.751 (0.81)
18	30	○	0.497 (1) ○			
19	59	0.181 (0.93)				0.388 (1.07)

Study	n_s	CB0	CB1	CB2	CB3	CB4
20	42	0.23 (0.9)			0.449 (1.09)	
21	29	○			0.497 (1) ○	
22	20	−1.581 (1)		−1.423 (1)		
23	29	★			★	
24	24	−1.364 (1.12)	−0.987 (0.87)			
25	30	0.325 (1.17)			0.949 (0.8)	
26	28	0.069 (0.94)			0.526 (1.05)	
27	14	○	1.141 (1) ○			
28	78			0.408 (1.03)	0.556 (0.97)	
29	25				0.237 (1.25)	0.39 (0.63)
30	50	0.485 (1.2)				0.924 (0.82)

For studies with continuous outcome that include estimated individual treatment effects, we know the standardized treatment effect estimates z_t^s and the standardized variances q_t^s. These quantities are derived from original continuous data that are not available for our analysis, and will be denoted by Y. Specifically, we denote by Y_{ti}^s the response of the ith of the n_t^s individuals assigned to treatment t in study s, and by Y_t^s the sample mean for treatment t in study s. Then the sample variance for arm t of study s is

$$v_t^s = \frac{1}{n_t^s - 1} \sum_{i=1}^{n_t^s} (Y_{ti}^s - Y_t^s)^2,$$

and the pooled estimate of the variance in study s is a weighted average of the individual arms' sample variances, given by

$$v^s = \frac{\sum_{t \in \mathcal{T}^s} (n_t^s - 1) v_t^s}{\sum_{t \in \mathcal{T}^s} (n_t^s - 1)}.$$

The standardized quantities that are available to us are defined as

$$z_t^s = \frac{Y_t^s}{\sqrt{v^s}},$$
$$q_t^s = \frac{v_t^s}{v^s}.$$

This is a common strategy for reporting study effects in meta-analysis (Hedges and Olkin, 1985).

A small number of studies originally measured and reported an ordinal scale. These were subsequently transformed to a continuous scale using cumulative ranks. Only the transformed outcomes are available from the evidence report, so these studies are treated as continuous. The five studies indicated by ∘'s in Table 4.4 adopted a crossover design, in which each patient receives both treatments at different points in time (Piantadosi, 1997). For those we only have available the estimated differences of standardized treatment effects, $z_t^s - z_{t'}^s$, between the two arms (usually between a treatment and a placebo), and the overall sample size $n_t^s + n_{t'}^s$. Study 27 (also indicated by a ∘ in Table 4.4) reports only the differences between the two arms, but the design is a standard randomized trial. For brevity we will sometimes refer to these as *incomplete studies*. Finally, two studies (indicated by ⋆'s in Table 4.4) treat the response as dichotomous, and report 2×2 contingency tables. We use the notation x_t^s for the number of successes in arm t of study s.

4.6.3 Sampling distributions and latent variables

Our overall strategy consists of two main steps: creating a common scale for the response to treatment in all studies; and developing a hierarchical model

for this common scale. For the continuous studies, and the so-called incomplete studies, the natural scale is given by the original, though unreported, values of Y. In crossover studies we introduce additional latent variables representing the unobserved effects in each of the arms. For the binary response studies we introduce, as in Section 4.5, latent normally distributed auxiliary variables and regard the reported binary responses as dichotomizations of the unobserved variables. In both cases our inference is based on an imputation of possible unreported responses for a hypothetical two-armed trial with continuous response, consistent with the abbreviated trial reports, and reflecting the uncertainty about the latent quantities.

The main structural assumption of our model is that the average response in arm (t, s) is the sum of two components: a study effect μ^s that represents the differences in responses across studies, and captures differing patient populations, protocol variations, and so on; and a treatment effect θ_t. This provides a reasonable flexible structure and at the same time allows comparisons of treatment effects across studies. In this setting the effect of a drug is assumed to be the same across studies. While it is technically feasible to model interactions between study and treatment efficacy, the data of Table 4.4 are too sparse to reliably estimate the resulting model. An alternative to interaction terms is hierarchical modeling of the treatment effects across studies, as was done in Section 4.4. Again, the sparsity of the data and the large number of drugs make it prohibitive to estimate such a model. For many drugs we would need to estimate a population of drug-specific effects based on two studies.

Our assumptions about the sampling distribution of the responses Y_t^s are as follows:

(i) the sample averages Y_t^s are approximately normally distributed with mean $\theta_t + \mu^s$ and variance σ_{ts}^2/n_t^s;
(ii) the sample variances v_t^s are approximately distributed as

$$\frac{\sigma_{ts}^2}{(n_t^s - 1)} \chi^2_{n_t^s - 1}.$$

The structure of the model for the complete continuous studies is shown in Figure 4.5. Each study provides information about as many θ_t as there are treatments in that study, and about the overall offset parameter μ^s which captures an overall shift in response for that study. In addition to the θ_t and μ^s, each study arm informs us about its own variance parameter. Because only standardized quantities are observed, there is no information about the original scale of the data. In other words, the empirical evidence remains equally likely if we multiply all the Ys by an arbitrary factor, say 2. Defining

$$\sigma_s^2 = \frac{\sum_{t \in \mathcal{T}^s} (n_t^s - 1)\sigma_{ts}^2}{\sum_{t \in \mathcal{T}^s} (n_t^s - 1)}$$

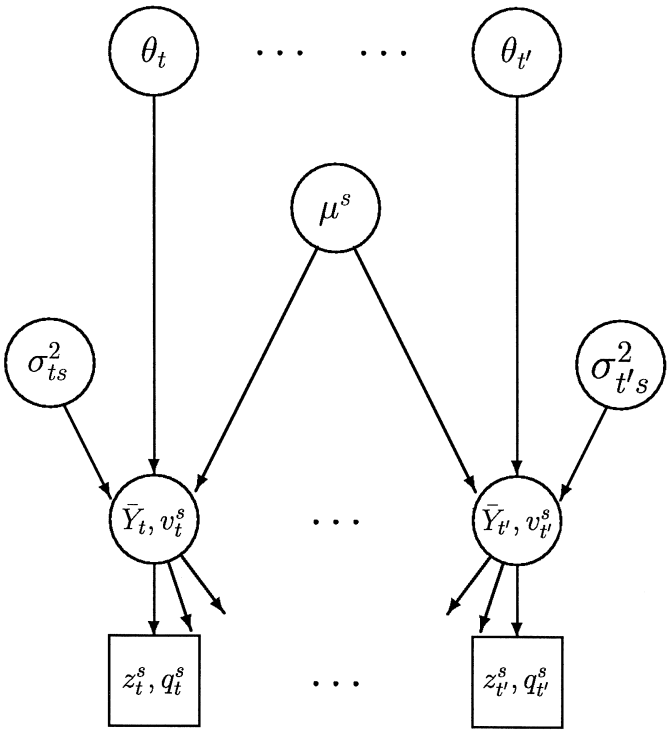

Figure 4.5 Summary of notation and conditional independence assumptions for the component model for complete study s. The squares with zs and qs are the observed standardized data; the circles with Ys and vs are the unobserved sufficient statistics for the latent normal response.

and $\gamma_t^s = \sigma_{ts}^2/\sigma_s^2$, the distribution of the observed zs and qs is independent of σ_s^2 given the γs. We can therefore set $\sigma_s^2 \equiv 1$ and interpret θ_t, μ^s, and σ_{ts}^2 as parameters of a model for the standardized sample averages Y_t^s/σ_s. By this convention the parameters σ_{ts}^2 are constrained to have a weighted average of unity within each study. In the incomplete studies there is no information about the γs, which will be assumed equal.

The hierarchical model of Figure 4.5 is specified in terms of unobserved variables Y_t^s, which are related to observed quantities in different ways for different studies. Anticipating our MCMC implementation, we give the conditional distribution of the latent Ys, given the reported observations. In the complete studies the unobserved v^s are distributed as

$$v^s \sim \frac{\sum_{t\in\mathcal{T}^s} \sigma_{ts}^2 \chi^2_{n_t^s-1}}{\sum_{t\in\mathcal{T}^s}(n_t^s-1)}.$$

Using v^s and the reported statistics, one can reconstruct all other unobserved statistics by way of $Y_t^s = z_t^s\sqrt{v^s}$ and $v_t^s = q_t^s v^s$.

The incomplete studies report only the standardized sample mean differences $(z_t^s - z_{t'}^s)$ and total sample sizes $(n_t^s + n_{t'}^s)$. Using the additional assumption that variances and sample sizes are equal in the two arms, v^s is distributed as

$$v^s \sim \frac{1}{(n_t^s + n_{t'}^s - 2)} \chi^2_{n_t^s + n_{t'}^s - 2}$$

and the reported difference is given by $\Delta_{tt'}^s \equiv (Y_t^s - Y_{t'}^s) = (z_t^s - z_{t'}^s)\sqrt{v^s}$. The conditional distribution of the missing sufficient statistics given the model parameters and the reported statistics is available from routine calculations with the normal distribution:

$$Y_t^s | \Delta_{tt'}^s \sim N\left(\mu^s + \frac{n_t^s \theta_t + n_{t'}^s \theta_{t'} + n_{t'}^s \Delta_{tt'}^s}{n_t^s + n_{t'}^s}, \frac{\sigma_s^2}{n_t^s + n_{t'}^s}\right).$$

For studies reporting only 2×2 tables we treat the data x_{ti}^s as indicators of the events $Y_{ti}^s > 0$, so that $P(x_{ti}^s = 1 | \theta_t, \mu^s, \sigma_{ts}) = \Phi((\theta_t + \mu^s)/\sigma_{ts})$. The constraint $\sigma_s^2 = 1$ still applies here. The conditional distribution of Y_{ti}^s given $x_{ti}^s = 1$ is $N(\theta_t + \mu^s, \sigma_{ts}^2)$, truncated to the positive values. When $x_{ti}^s = 0$, the same normal is truncated to the negative values.

4.6.4 Treatment and study variation

We now discuss the hierarchical model for the latent variables Y. The distributions of Y are controlled by treatment effects θ, the study effects μ and the arm-specific variances σ^2. We model hierarchically the θs and the μs, but not the σ^2s, as discussed. The overall structure of the model is shown in Figure 4.6.

The μs are modeled as independent and identically distributed random variables, independent of θ. To capture potential clustering in the distribution of the study effects we used a mixture of normal distributions. Each mixture component has an unrestricted variance, but the component means are constrained to give $E(\mu^s) = 0$. Based on diagnostic plots, discussed in Section 4.6.9, we choose a two-component mixture, that is

$$\mu^s | \omega, \alpha, \sigma_{\mu 1}^2 \sigma_{\mu 2}^2 \overset{\text{iid}}{\sim} \omega N(\alpha, \sigma_{\mu 1}^2) + (1 - \omega) N\big(-\omega\alpha/(1 - \omega), \sigma_{\mu 2}^2\big).$$

The resulting variance of the study effects is $\sigma_\mu^2 = \omega\sigma_{\mu 1}^2 + (1 - \omega)\sigma_{\mu 2}^2 + \omega\alpha^2/(1 - \omega)$. Here the mixture weight ω, the offset parameter α, and the component variances $\sigma_{\mu 1}^2$ and $\sigma_{\mu 2}^2$ are all unknown and must be estimated from the data. For identifiability, α is defined to be the positive one of the two component means, a constraint enforced through its prior distribution. This distributional specification, via the unknown parameters $\omega, \alpha, \sigma_{\mu 1}^2$, and $\sigma_{\mu 2}^2$, induces a dependence among the μ^s.

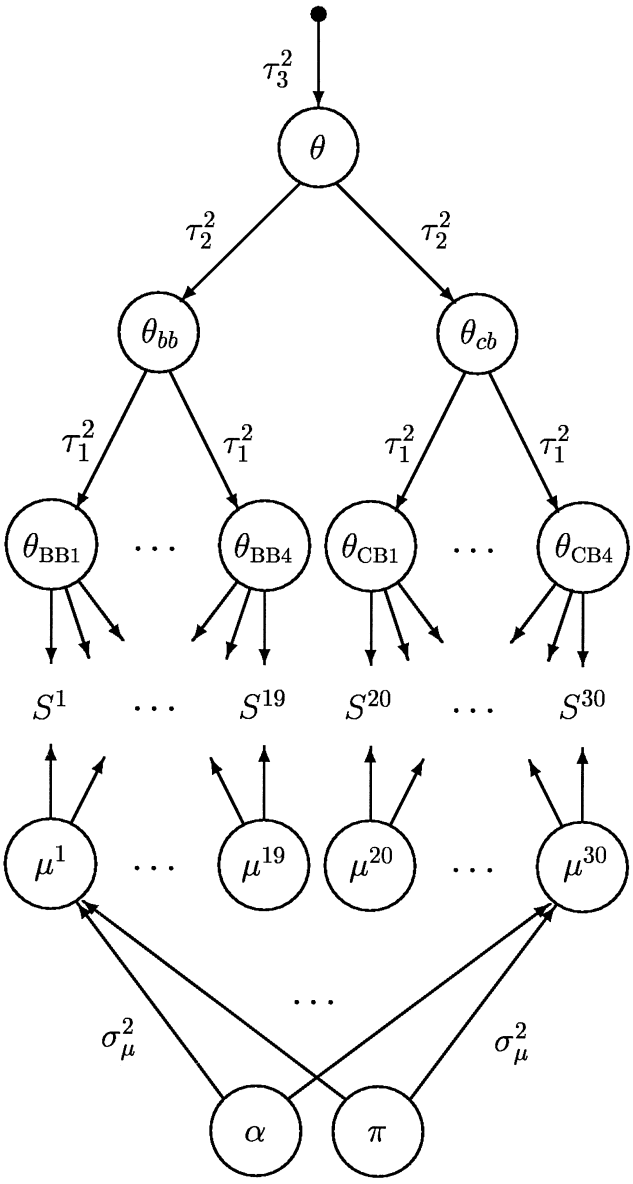

Figure 4.6 Summary of notation and conditional independence assumptions. $S^1, \ldots, S^{30}$ indicate the component models for each study, represented in Figure 4.5. Study and treatment effects are represented by nodes, while variance components are indicated next to the arc representing the corresponding conditional distribution. For example, τ_2^2 is the conditional variance of θ_{bb} given θ.

The response scale origins may differ across studies; to obtain a meaningful comparison, we set the placebo effect θ_0 to zero. For studies without a placebo arm, the study effect μ^s also embodies any difference in scale origins. The vector θ of treatment effects for nonplacebo arms is modeled as multivariate normal, with a dependence structure suggested by the clinical evidence of stronger similarity of treatments within classes, and weaker similarity for treatments across classes. We accomplish this with the two-level structure illustrated in the top portion of Figure 4.6.

Specifically, we define an overall mean parameter θ, and treatment class mean parameters $\theta_{\rm bb}$ and $\theta_{\rm cb}$. Correspondingly, we have variance components τ_1^2 for the variation of effects within the same treatment class; τ_2^2 for the variation of class effects $\theta_{\rm bb}$ and $\theta_{\rm cb}$ with respect the overall mean parameter θ; and finally, τ_3^2 for the variation of the overall mean parameter θ with respect to its mean of 0. The marginal variance of each treatment effect is $\tau_1^2+\tau_2^2+\tau_3^2$. Our hierarchical specification implies correlations among the treatment effects. For treatments t, t' within the same class we have

$$\mathrm{cor}(\theta_t, \theta_{t'}) = \rho_0 = \frac{\tau_2^2 + \tau_3^2}{\tau_1^2 + \tau_2^2 + \tau_3^2};$$

while for treatments in different classes we have

$$\mathrm{cor}(\theta_t, \theta_{t'}) = \rho_1 = \frac{\tau_3^2}{\tau_1^2 + \tau_2^2 + \tau_3^2}$$

As desired, $0 \le \rho_1 \le \rho_0 \le 1$.

Introducing these parameters extends the more common model where all treatment effects are drawn from the same distribution, which corresponds to setting $\rho_0 = 0$, or $\tau_2^2 = \tau_3^2 = 0$. Another special case is the separate analysis of the two treatment classes, with common variance τ_1^2 in all classes, which corresponds to $\rho_0 = 1$ and $\rho_1 = 0$, or $\tau_2^2 \to \infty$ and $\tau_3^2 = 0$. Our correlation structure generates differential borrowing of strength in the effect size estimates within and across classes. Because one of the main goals is to make inferences about relative ranking, it seems appropriate to incorporate clinical knowledge about the ensemble of treatments. In a different context, such as deciding whether to approve an individual treatment, other approaches may be more appropriate.

In summary, the distributional assumptions of our latent variables and hierarchical model parameters are as follows:

$$\begin{aligned}
\theta_{\rm bb}, \theta_{\rm cb} &\overset{\rm iid}{\sim} N(\theta, \tau_2^2) \\
\theta_t &\overset{\rm ind}{\sim} \begin{cases} N(\theta_{\rm bb}, \tau_1^2), t \in \mathcal{T}_{\rm bb} \\ N(\theta_{\rm cb}, \tau_1^2), t \in \mathcal{T}_{\rm cb} \end{cases} \\
\mu^s &\overset{\rm iid}{\sim} \omega N(\alpha, \sigma_{\mu 1}^2) + (1-\omega) N\big(-\omega\alpha/(1-\omega), \sigma_{\mu 2}^2\big) \\
Y_t^s &\overset{\rm ind}{\sim} N(\theta_t + \mu^s, \sigma_{ts}^2/n_t^s) \\
v_t^s &\overset{\rm ind}{\sim} \big(\sigma_{ts}^2/(n_t^s - 1)\big)\chi^2_{n_t^s - 1}.
\end{aligned} \tag{4.2}$$

Here each distribution is conditional on all the parameters higher in the list and hierarchy, although it may not depend on all of them as a result of the conditional independence assumptions represented in Figure 4.6.

While both this analysis and that of tamoxifen studies in Section 4.4 use hierarchical models, they differ in two important respects which highlight the flexibility of hierarchical models in this context. The first is that in the headache application the study effects are dependent, while in the tamoxifen application each study had a separate baseline success rate in the control arm and there was no borrowing of strength across studies for those parameters. The reasons for this difference are the much larger number of studies in the headache application and the smaller sample size of most of them. Both strategies could be defended in both analyses. The second is that the treatment effects are assumed to be constant across studies in the headache example, while they are allowed to be heterogeneous and modeled hierarchically in the tamoxifen application. In the headache example the hierarchical model describes variation of the treatment effect of similar drugs within the same class, while in the tamoxifen example the hierarchical model describes variation of the treatment effect of the same drug in different studies.

4.6.5 Prior distributions

For a Bayesian analysis of this model we must specify distributions for all unknown parameters not included in the list (4.2). We will refer to these as priors, even though it is somewhat arbitrary to define the boundary between prior and likelihood in multilevel hierarchical models. For example, it is not unusual to think of the distributions on θ_{bb} and θ_{cb} as part of the prior as well.

An attractive and practical approach for choosing prior distributions on high-level parameters in complex hierarchical models such as this is to specify a dispersed but proper baseline prior distribution, and to supplement the baseline analysis with additional sensitivity analyses. In Section 4.6.10 we present sensitivity analyses for four possible departures from the following baseline prior specification:

Overall effect mean	θ	$\sim N(0, \tau_3^2)$	
Variance (treatment effects)	τ_j^2	$\sim \text{IG}(a_j, b_j)$,	$j = 1, 2$
Variance (study effects)	$\sigma_{\mu i}^2$	$\sim \text{IG}(a_\mu, b_\mu)$,	$i = 1, 2$
Offset (study effects)	α	$\sim N(0, b_\alpha) I_{\alpha > 0}$	
Mixture weight (study effects)	ω	$\sim U(0, 1)$	
Variance (individual effects)	σ_{ts}^2	$\sim \text{IG}(a_\sigma, b_\sigma)$,	$t = 1, \ldots, 8$
			$s = 1, \ldots, 30.$

The as and bs are fixed hyperparameters.

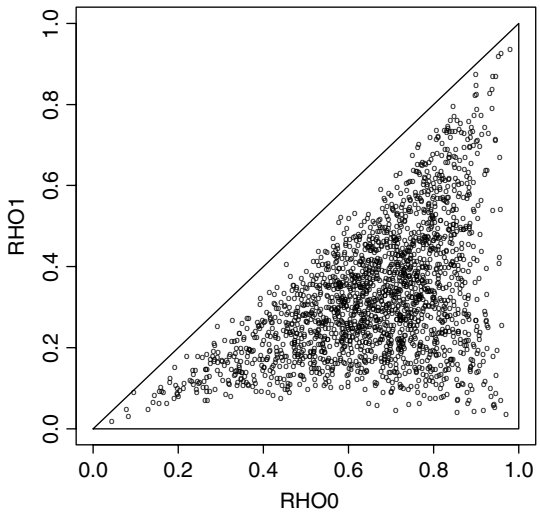

Figure 4.7 A sample from the joint prior probability distribution of the correlations ρ_0 and ρ_1 in the baseline specification.

Unless otherwise implied by the hierarchical structure, we used independent prior distributions. We use the parameterization of the inverse gamma with density function $b^a x^{-a-1} e^{-b/x}/\Gamma(a)$ for $x > 0$, with mean $b/(a-1)$ (for $a > 1$) and variance $b^2/(a-1)^2(a-2)$ (for $a > 2$). For our baseline analysis we selected hyperparameters $a_j = a_\mu = 3$, the integer value giving the most diffuse finite-variance prior distributions to our variance components, and $b_j = b_\mu = 2/3$, to ensure that the marginal expected treatment effect variance is $\tau^2 = \sum_{j=1}^{3} \tau_j^2 = 1$, appropriate for these standardized quantities, apportioned equally to the three levels of the hierarchy. A sample from the resulting distribution of the correlations ρ_0 and ρ_1 is shown in Figure 4.7. The prior distributes mass throughout the region. We assign the study effect offset α a half-normal distribution with variance $b_\alpha = 3$, and give improper reference distributions $(1/\sigma_{ts}^2)$ to the study-specific individual effect variances, since the sample sizes are sufficient to support a noninformative analysis of these parameters. We adopt the conservative approach of using prior mean 0 for the overall treatment and placebo effect mean θ, suggesting no prior information that any of the treatments is better or worse than the placebo.

4.6.6 Computing

We evaluate the posterior distribution of the parameters and unreported observations is Markov chain Monte Carlo simulation (Tanner and Wong, 1987; Tierney, 1994). Most of the full conditional distributions, including those of the latent sufficient statistics, are available in closed form and simple

to sample from, making the Gibbs sampling variation of MCMC (Gelfand and Smith, 1990) viable. The identifying constraint $\sigma_s^2 = 1$ makes closed-form expressions for the study-specific variances unavailable, so we use a Metropolis step (Tierney, 1994). The normal mixture parameters for the study effects are updated by augmenting the data with a vector of latent mixture component indicators, as done by Diebolt and Roberts (1994) and West and Turner (1994) for their unconstrained cases. The constraint $E(\mu^s) = 0$ is linear in α and so the full conditional distribution of α is normal, truncated to the positive half-line. At each iteration the latent variables are simulated from the distributions given in Section 4.6.3.

It is important to check graphical diagnostics for convergence of the chain, as implemented, for example, in CODA (Best *et al.*, 1995). In this application, the MCMC sampler mixes well, and a few hundred iterations typically suffice for convergence. Because iterations are not expensive, we present results based on a subset of 6000 equally spaced draws from the last 30 000 iterations of a single chain of 35 000 iterations.

4.6.7 Results

Our objective is to make inferences about treatment effectiveness, identifying which treatment class seems best and how well each treatment works, by computing the posterior distribution of the treatment effect vector θ, given the reported values z_t^s and q_t^s. The posterior distributions of the treatment effects θ's are shown in Figure 4.8. All θs have positive posterior means, indicating effective treatments. Overall, the probability that none of the treatments has a positive effect is smaller than 0.0002; the probability that at least one of the treatments has an effect larger than 1 is 0.46; the probability that all treatments have a positive effect is 0.35.

To emphasize the contribution of individual studies and to validate the assumption of homogeneity of effects across studies, Figure 4.8 also shows study-specific summaries $Y_t^s - \mu^s$. These are the quantities that contribute to the estimation of the treatment effect within each study. Because the μ^s are unknown, we replaced them with their posterior means. So each corresponds to a study including the treatment in question, and is positioned at the posterior mean of $Y_t^s - \mu^s$.

An important contribution of a meta-analysis to patient management is to quantify how much is known about which treatments are best. This can be expressed by a probability distribution over possible rankings. In addition to supporting treatment decisions, an accurate assessment of uncertainty may help guide the planning of new trials. In a simulation-based approach the ranking problem can be solved easily even in the presence of complex random effects and missing data that may induce significant dependencies in the joint distribution of the effects under comparison. Ranking is done simply by reporting the empirical frequency with which each treatment is

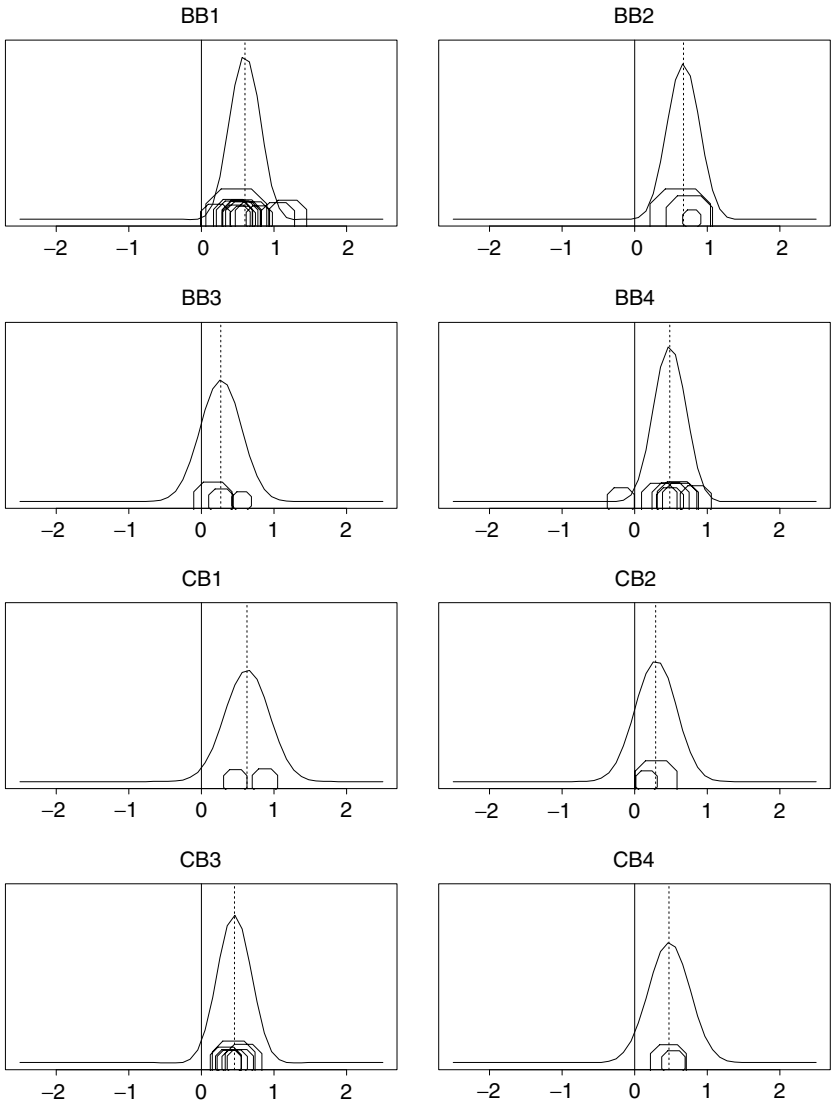

Figure 4.8 Marginal posterior distribution of treatment effects for the beta-blockers (top four) and calcium-channel blockers (bottom four) classes. Labels identify treatments; each circle corresponds to a study that includes the treatment in question, and is positioned at the posterior mean of $Y_t^s - \mu^s$; areas of circles are proportional to the sample sizes n_t^s. Posterior means and no-effect points are indicated by vertical dashed and solid lines, respectively.

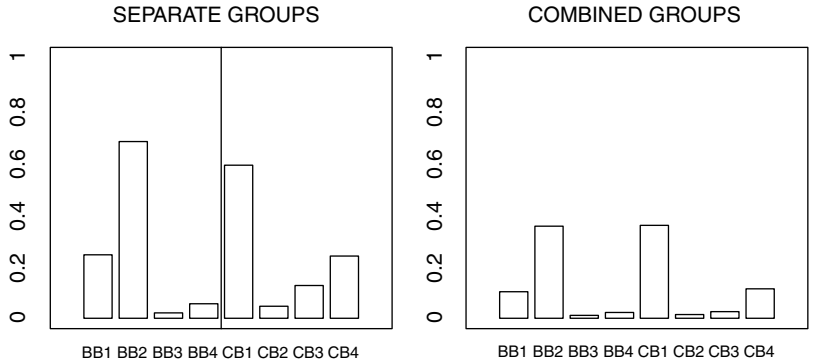

Figure 4.9 Probabilities that each treatment effect is the largest, by treatment class (left), and all treatments combined (right). Substantial uncertainty remains about which treatments are most effective, both within classes and overall.

better than its competitors within the same treatment group. In Figure 4.9, we show the probability that each treatment effect is the largest. To convey a sense of the remaining uncertainty about the choice of the best treatment, we also report an overall ranking. Labeling the drugs in Figure 4.9 from left to right as 1 to 4, the most probable ranking is $1, 4, 3, 2$, with an estimated posterior probability of 20.6%, followed by $4, 1, 3, 2$ and $4, 3, 2, 1$, with estimated posterior probabilities of 15% and 11%, respectively. None of the studies considered here compares treatments across classes, so that our comparisons across classes rely heavily on assumptions about how populations, and particularly placebo groups, compare across studies.

A comparison of the overall class effects θ_{bb} and θ_{cb} helps us address the issue of which drug class is likely to perform better. Figure 4.10 displays histograms of samples from the posterior distributions of the overall class effects, along with the corresponding prior densities. There is substantial overlap between the two posterior distributions, as expected, suggesting that considerable uncertainty remains about which class of treatments is most effective; in each case the prior is considerably more diffuse than the posterior, suggesting that the data offer evidence about treatment means.

In contrast, the data provide relatively little information about the correlations. The posterior distribution, shown in Figure 4.11, concentrates on high values for the within-class correlation coefficient ($E[\rho_0 \mid \text{data}] = 0.774$, with posterior interquartile range (IQR) $(0.674, 0.842)$), and moderate values for the related-class coefficient ($E[\rho_1 \mid \text{data}] = 0.394$, with posterior IQR $(0.274, 0.505)$); recall that the prior distributions had means $E[\rho_0] = 0.666$, $E[\rho_1] = 0.333$, with prior IQRs $(0.569, 0.788)$ and $(0.207, 0.419)$, respectively. In the figure, dashed vertical lines represent prior means, and solid vertical lines represent posterior means.

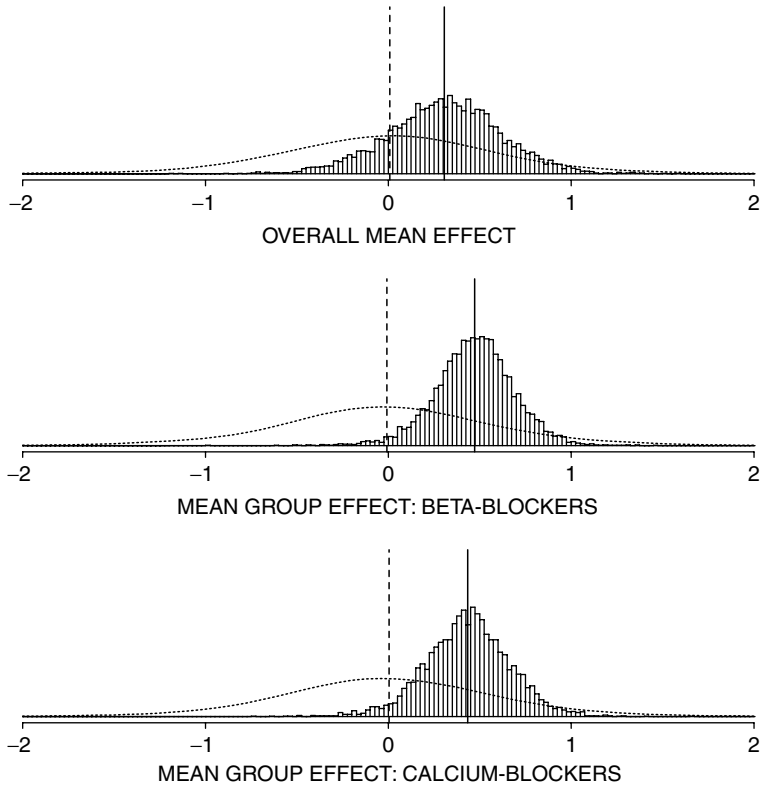

Figure 4.10 Posterior distribution of the overall effect θ and class effects θ_{bb} and θ_{cb} for the two treatments classes. The dotted lines are the prior distributions. Posterior means and no-effect points are indicated by vertical solid and dashed lines, respectively.

The approach and results presented here can potentially provide valuable input for the design of new trials. The treatment rankings can assist in choosing treatment arms; the treatment effect estimates can help in selecting sample sizes. Our hierarchical model can also provide the basis for more formal simulation-based optimal design. For any fixed choice of treatment, one can simulate from the predictive distribution of $Y_t^{s'}$ in a future study $s' \notin \mathcal{S}$. Simulated values can in turn be used to evaluate the information provided by a trial design (Müller, 1998). Incorporating two treatment classes within the same model enables one to consider future trials comparing treatments across classes.

4.6.8 Comparisons with alternative approaches

We now consider a comparison of our results with those obtained using two alternative model specifications. The first is a separate hierarchical analysis of

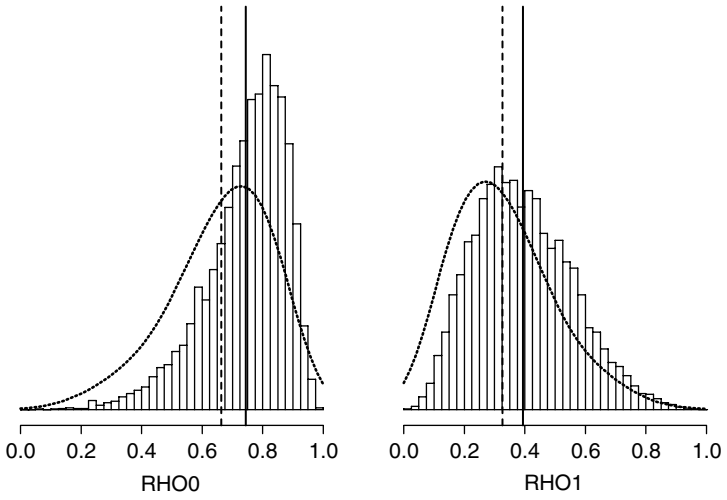

Figure 4.11 Prior and posterior distributions of correlation coefficients ρ_0 and ρ_1 of treatment effects within and across drug classes.

the two treatment classes, without the constraint of a common τ_1^2. In this case there is generally less shrinkage toward a common mean for the biofeedback treatments, and slightly higher posterior standard deviations. The variance component τ_1^2, measuring the heterogeneity of treatment effects within classes, is estimated with substantially greater precision in the combined analysis, where the parameter is common across classes. Because of the small number of treatments, it is important to incorporate information about the variability of effects in similar classes of treatments when estimating variances of treatment effects. Ranking within classes is not changed substantially. Ranking across classes is not possible in this analysis.

The second comparison is with a linear mixed model with study-specific random effects, estimated using the maximum likelihood approach proposed by Hasselblad (1998). Dichotomous outcomes measures were converted to effect sizes using the method of Hasselblad and Hedges (1995). This results in an estimate of the difference in effect sizes, so these studies are represented as contrasts, rather than effects, in modeling the arm-specific means. Crossover studies are treated similarly. Figure 4.12 compares the point estimates obtained using the two approaches. The point estimates are close, differing by less than 0.3. The Bayesian model systematically shrinks the estimates more strongly toward the center of the distribution of estimates. Shrinkage occurs in differing degrees to different effect sizes, the effect sizes that are less accurately estimated being pulled further in. This determines a reversal of ranking between verapamil and timolol for the most effective treatment and between nadolol and nifedipine for the least effective. The Bayesian model also features smaller variances reflecting the correlation within the

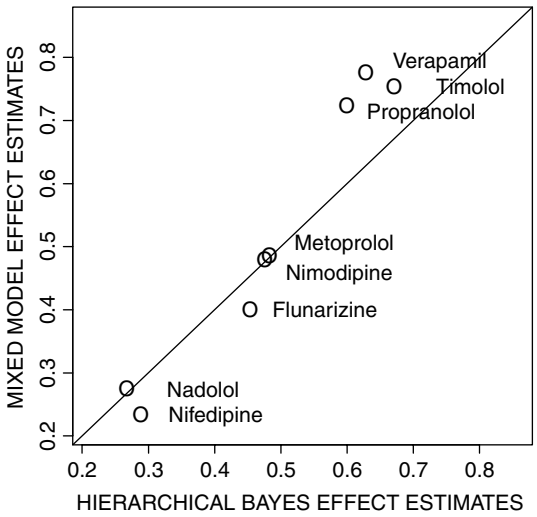

Figure 4.12 Comparison of Bayesian hierarchical effect size estimates (horizontal scale) and maximum likelihood effect size estimates using a linear mixed model.

treatment classes. However, the Bayesian model reflects the uncertainty from the incomplete studies, which tends to increase the variance. The two effects are roughly balanced for treatments in incomplete studies, but in other treatments the effect standard deviations can be reduced by a factor of 2. The linear mixed model approach does not lend itself easily to generating ranking probabilities.

Our results are consistent with those of a non-Bayesian meta-analysis with study-specific random effects, when combining the evidence from relatively large studies. However, they differ in the smaller studies by shrinking the study-specific reported treatment effects towards the mean effect for similar studies and towards the skeptical prior mean effect of zero. As a result of shrinkage, our analysis leads to conclusions that are both more precise in estimating individual treatments and more conservative in drawing conclusions about ranking. We regard this as a scientific justification for the medical community's natural reluctance to be swayed by striking results from small studies, and as a systematic tool for basing inference on a growing body of evidence.

4.6.9 Graphical model checking

When building a complex multilevel model, it becomes especially important to check the validity of assumptions about latent variables and parameters at higher levels in the hierarchy. An effective strategy is to inspect sequences of diagnostic plots of summary measures from the full conditional distributions.

Ideally, summaries should constructed to be conditionally independent and identically distributed if the model assumptions are satisfied. Here we discuss and illustrate this strategy in the context of validating model assumptions about the study effect distributions, and about the overall fit of the model. Another illustration of this idea is presented by Meulders *et al.* (1998).

We begin by considering the distribution of the study-specific random effects μ^s shown in Figure 4.13. The offset parameter α has posterior mean 0.27 and IQR $(0.13, 0.39)$; the mixture weight ω has posterior mean 0.44 and IQR $(0.36, 0.52)$. The two component-specific variances $\sigma^2_{\mu 1}$ and $\sigma^2_{\mu 2}$ have posterior means 1.00 and 1.01, and IQR's $(0.88, 1.11)$ and $(0.65, 1.25)$, respectively. Figure 4.13 illustrates the distributions of the individual study effects μ^s within the three treatment classes. Effects can be large, and are clearly heterogeneous across studies. Studies providing only treatment differences offer no direct evidence about individual study effects, leading to the most widely dispersed distributions. Studies reporting 2×2 tables also have dispersed distributions of random effects, but these can stray far from 0 for informative studies; for example, study 5 has a smaller study effect than study 23, as a result of its lower success rate in the placebo group. Study 15 has the largest positive study effect. Study 11 has one of the lowest study effects. This reflects the fact that both the BBI and the BB4 arms of that study have a low response, while they generally perform well elsewhere (Figure 4.8).

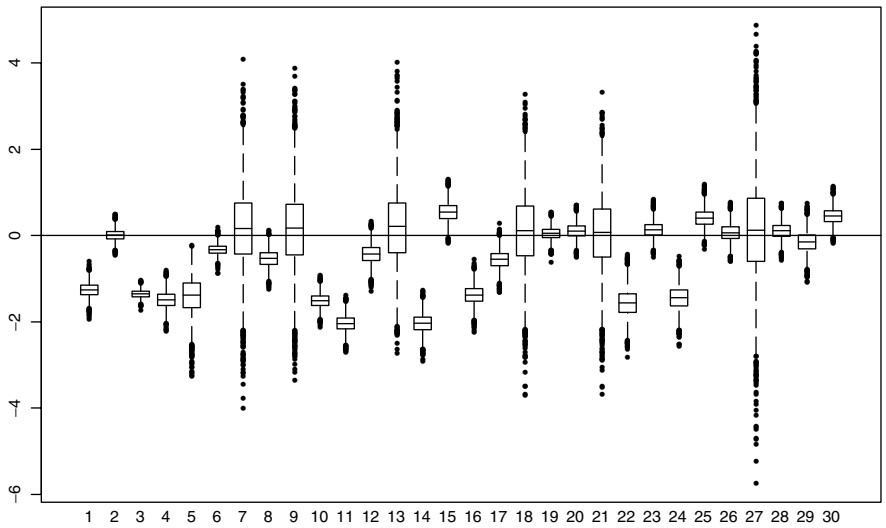

Figure 4.13 Posterior distribution of study effects. Study numbers are listed under each boxplot. Studies 7, 9, 13, 18, 23, and 27 report the estimated differences in effect sizes between the two arms. Studies 5 and 23 report 2×2 contingency tables.

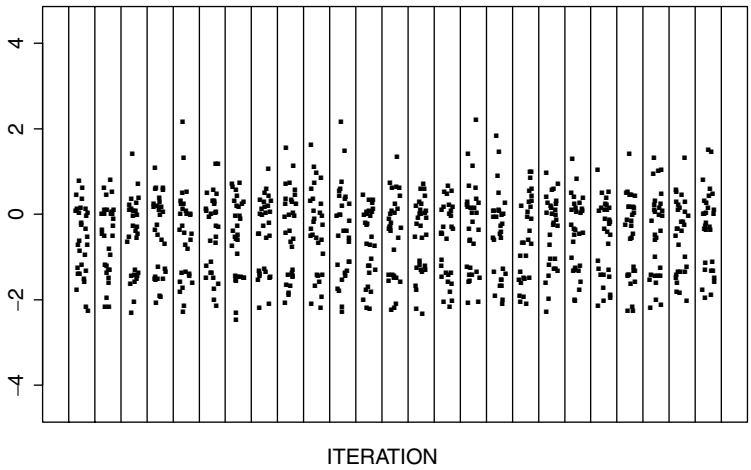

Figure 4.14 A random sample of distributions of the diagnostic summaries m. Each vertical stripe corresponds to one iteration in the Markov chain. At most iterations, the ms exhibit clustering into two distinct subgroups, supporting the choice of a two-component mixture for the distribution of the study effects.

When updating the study effects μ_s, given all other unknowns in the model, the data-dependent quantities

$$m_s = \frac{\sum_{t \in \mathcal{T}^s} (Y_t^s - \theta_t) n_t^s / \sigma_{ts}^2}{\sum_{t \in \mathcal{T}^s} n_t^s / \sigma_{ts}^2}$$

play the role of the observed sample means. For each iteration of the chain we can use the empirical distribution of the m's to assess the validity of the posited distribution of the μ's, much as when validating a likelihood function based on an empirical sample. We generated repeated samples of distributions of m's (Figure 4.14). Recurring features are skewness to the left and heavy tails. Inspecting these distributions and Kolmogorov–Smirnov statistics quantifying their nonnormality led us to replace the simple normal model specified initially with the two-component mixture adopted here. Direct inspection of the distributions of μs would not be as sensitive a diagnostic for two reasons: draws would be from the marginal posterior distribution of the μs rather that the conditional we are validating; and draws would depend directly on the current distributional assumption about the μs. We examined similar plots for the treatment effects. Because of the relatively small number of treatments, the data do not provide detailed information about the shape of the higher-level distribution for which the normal assumption appears to fit well.

To assess the overall fit of the model we considered the latent residuals $\epsilon_t^s = Y_t^s - \theta_t - \mu^s$ and studied their posterior predictive distributions. At each iteration of the chain we simulate data Y_t^s conditionally on the current

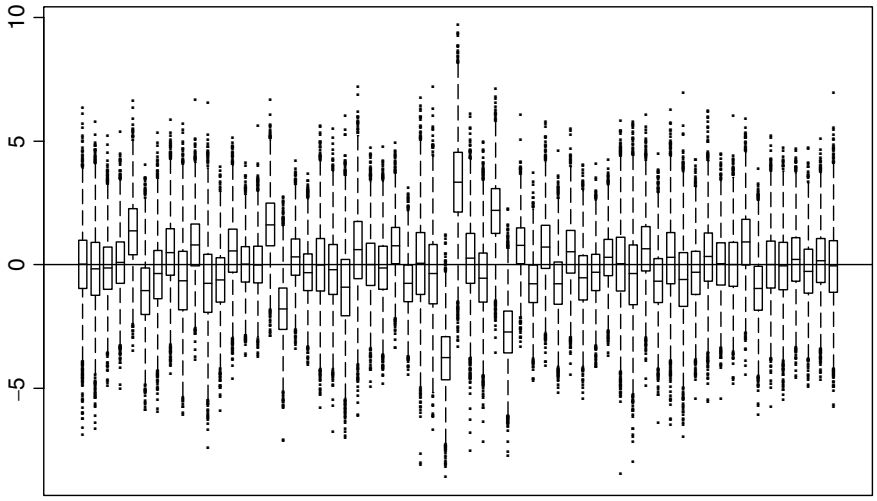

Figure 4.15 Sample from the posterior predictive distributions of the residuals. Each boxplot corresponds to an arm. Studies are ordered from left to right. Within each study, arms are in the same order as in Table 4.4. The vertical scale is $\eta_t^s = (Y_t^s - \theta_t - \mu^s)\sqrt{n_t^s}/\sigma_{ts}$, the standardized deviation of the latent sufficient statistic from its predicted mean. When 0 is in the tail of one of these distributions, the model does not fit well in that study arm.

values of the model parameters, and compute the resulting residuals; at the end of the MCMC simulation the collection of simulated residuals is a sample from their unconditional predictive distribution. For 0 to lie in the tail of one of these distributions is an indication that the model does not fit well in that study arm. We use boxplots to assess this informally, but more formal methods, such as predictive p-values, are also available (Meng, 1994; Rubin, 1984). To accommodate the different variances and sample sizes in the studies, we study the standardized latent residuals $\eta_t^s = \epsilon_t^s\sqrt{n_t^s}/\sigma_t^s$.

Figure 4.15 shows a sample of η_t^s from all arms. The fit of the model is very close to the data, with all boxplots including 0 within their whiskers. The predictive distributions display the heavy tails and other departures from normality that are characteristic of mixing over complex model specifications. In all cases 0 is within the whiskers of the boxplot, indicating that 0 is not an outlying observation in the distribution of residuals. There are three consecutive pairs of residuals in which the middle 50% of the boxplots are fully on opposite sides of 0. These correspond to studies 14, 31, and 33. In study 33, for example, the two arms have effects that are both discordant with the general tendency of the same treatments in other studies.

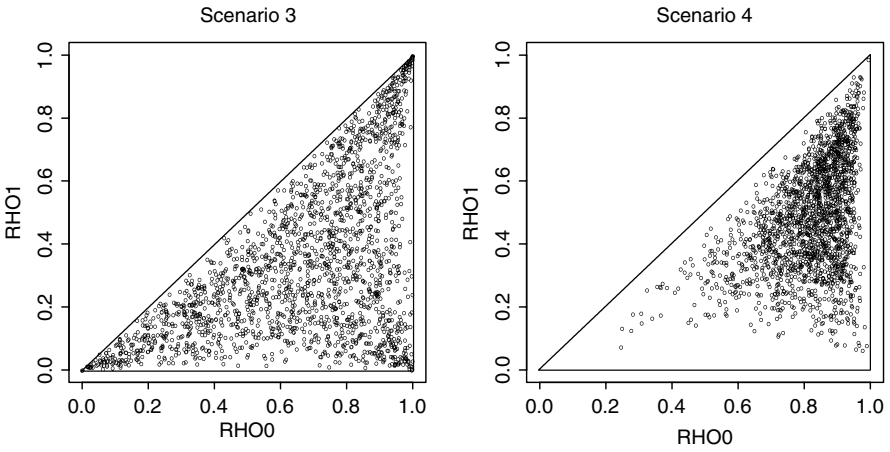

Figure 4.16 Samples from the joint prior probability distribution on the correlations ρ_0 and ρ_1 in scenarios 3 and 4.

4.6.10 Sensitivity to prior specification

The influence of our choice of prior hyperparameters can be addressed using a sensitivity analysis. Our strategy is to select four alternative scenarios for departure from our baseline hyperparameter prior distributions of Section 4.6.5, and evaluate under each of them the outcome of primary interest, in this case the ranking of treatments. The four scenarios are:

1. As the baseline, except that the study effects variance is set to be larger, by specifying $a_\mu = 6/5$ and $b_\mu = 20$ for both mixture components;
2. As the baseline, except that the study-specific variances are set to be close to unity, by specifying $(a_\sigma, b_\sigma) = (11, 10)$;
3. As the baseline, except the correlations ρ are given highly dispersed U-shaped prior distributions, by specifying $a_1 = a_2 = a_3 = 6/5$ and $b_1 = b_2 = b_3 = 1/15$;
4. As the baseline, except the variance component means increase with their place in the hierarchy. The specification is $a_1 = a_2 = a_3 = 3$ and $(b_1, b_2, b_3) = (1/3, 2/3, 1)$. This choice preserves the property that $E[\tau^2] = 1$, which is demanded by the standardization of the effect sizes.

Samples from the joint prior distribution on the correlations ρ_0 and ρ_1 in scenarios 3 and 4 are shown in Figure 4.16. Scenario 3 is nearly uniformly distributed, while scenario 4 favors higher correlations, and therefore a greater borrowing of strength from treatment to treatment.

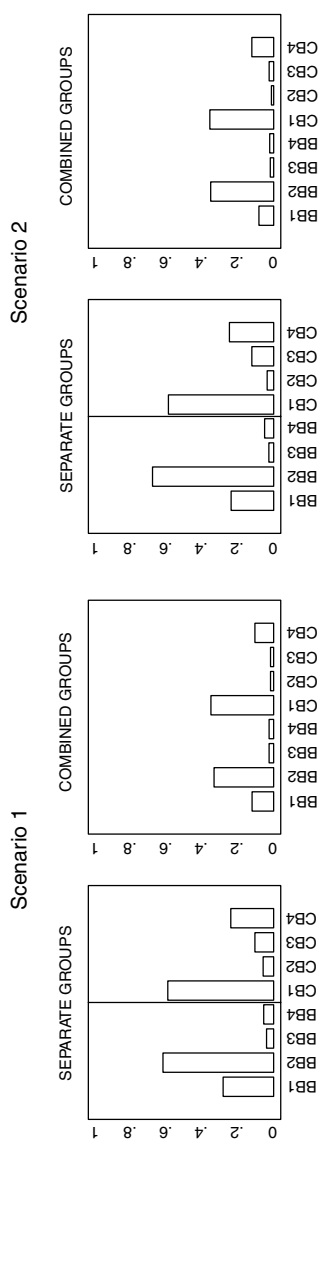

Figure 4.17 Probability that each treatment effect is the largest, by treatment class, and all treatments combined, under the four scenarios considered in our sensitivity analysis.

The probabilities that each treatment effect is the largest, by treatment class, and all treatments combined, under the four scenarios considered in our sensitivity analysis, are shown in Figure 4.17. The ranking of treatments is stable across these four scenarios, and therefore it does not appear that the conclusions about ranking are sensitive to the prior specification.