8

Evidence Synthesis

8.1 INTRODUCTION

It is unusual for a policy question to be informed by a single study. Interest in more diffuse areas, such as health-care delivery or broad public health measures, means that health-care evaluations become more realistically complex and there is an inevitable demand to make use of the huge volume of published and unpublished evidence. A quantitative synthesis of multiple studies has become known as a *meta-analysis*, whose procedures for randomised trials have become increasingly formalised by the Cochrane Collaboration (Section A.2). This has led to parallel developments for observational studies (Stroup *et al.*, 2000), and in the context of social science by the Campbell Collaboration (Section A.2).

A Bayesian approach to such 'standard' meta-analyses is considered in Section 8.2, emphasising the additional flexibility that arises both from the use of prior information and the adoption of Markov chain Monte Carlo methods for dealing with more complex models (Section 8.2.2). In particular, Section 8.2.3 illustrates the ability to handle the tricky and controversial issue of dependence of the treatment effect on baseline risk. The basic meta-analysis procedure can be further extended to increasingly complex contexts. First, we examine the somewhat specific but useful issue of *indirect comparison* analyses (Section 8.3), which are required when multiple studies have been carried out in which multiple treatments have been compared in different combinations, and we wish to draw inferences about specific treatment contrasts. Second, we examine the broader topic of *generalised evidence synthesis* (Section 8.4), in which studies of possibly different designs are pooled in order to estimate quantities of interest – a wide range of alternative models for pooling are available, broadly following the structure outlined for handling historical data (Section 5.4).

Since the basic methodological procedures were established in Section 3.17, this chapter relies heavily on a series of quite detailed examples, featuring prediction from meta-analyses (Example 8.1), meta-analysis with rare events (Example 8.2), dependence on baseline risk (Example 8.3), indirect comparisons in drug trials (Example 8.4), synthesis of RCTs and observational studies

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

(Example 8.5), and two examples of the synthesis of multiple studies to estimate the effects of a screening programme (Examples 8.6 and 8.7).

Many of the ideas in this chapter were suggested by Eddy *et al.* (1992) under the general label 'confidence profile method', and promulgated with numerous worked examples and accompanying software (FAST*PRO). They used directed conditional independence graphs (Section 3.19.3) to represent the qualitative way in which multiple contributing sources of evidence relate to the quantity of interest, explicitly allowing the user to discount studies due to their potential internal bias or their limited generalisability (Section 7.3). Their analysis was essentially Bayesian, although it was possible to avoid specification of priors and use only the likelihoods. The need to make explicit subjective judgements concerning the existence and extent of possible biases, and the limited capacity and friendliness of the software, have perhaps limited the application of this technique. However, throughout this chapter we show that modern software can allow straightforward implementation of their ideas, and we fully acknowledge their foresight in promoting these concepts.

8.2 'STANDARD' META-ANALYSIS

8.2.1 A Bayesian perspective

A standard classical meta-analysis will comprise a series of K studies each estimating a treatment effect $\theta_k, k = 1, \ldots, K$, by means of a likelihood which can be expressed, possibly approximately, as

$$y_k \sim \mathrm{N}[\theta_k, s_k^2], \tag{8.1}$$

whether the sample variances s_k^2 are generally considered known or estimated. Following the development in Section 3.17, individual estimates of the θ_k can be termed a *fixed-effects* analysis in which there is no pooling; at the other extreme an analysis in which all the θ_k are assumed equal may be termed *pooled-effect*. An intermediate *random-effects* analysis (DerSimonian and Laird, 1986) treats the θ_k as if they were drawn from a population distribution, generally taken as

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

As mentioned in Section 3.17, a variety of classical techniques are available for estimating τ^2; see Sutton *et al.* (2000) and Whitehead (2002) for recent reviews.

From a Bayesian perspective, it is natural to treat meta-analysis as a standard problem of multiplicity (Section 3.17), and follow the approach taken in contexts such as subset analysis (Section 6.8.1), multi-centre trials (Section 6.8.2), multiple N-of-1 studies (Section 6.11) and institutional comparisons (Section 7.4). Thus, if we are willing to treat the trials as exchangeable, the 'true'

treatment effect in each trial is considered a random quantity drawn from some population distribution, in exactly the same manner as the standard random-effects approach to meta-analysis. However, the latter tends to focus on estimating an overall treatment effect, while a full Bayesian approach also concentrates on estimating trial-specific effects and, as we shall see below, permits a variety of useful extensions. A simple 'empirical Bayes' meta-analysis has already been presented in Example 3.13.

The Bayesian approach requires prior distributions to be specified for the mean effect size μ, the between-studies standard deviation τ, and possibly the within-study variances; as in other hierarchical models, specifying default 'reference' priors for τ is not straightforward (Section 5.7.3).

Some of the potential advantages of the Bayesian approach to meta-analysis are rather briefly summarised below (Sutton *et al.*, 2000); of course, many of these issues can also be tackled from a classical perspective, but perhaps with less flexibility.

1. *Unified modelling*. The conflict between fixed- and random-effects meta-analysis is overcome by explicitly modelling between-trial variability (which could be assumed to be small). The 'random-effects' distribution can also be much more flexible than the standard normal assumption, for example partitioned into subgroups within which studies might be assumed equal or exchangeable.
2. *Borrowing strength*. As in all areas of Bayesian hierarchical modelling, an exchangeability assumption leads to each experimental unit 'borrowing' information from the other units, leading to a shrinkage of the estimate towards the overall mean, and a reduction in the width of the interval estimate. This degree of pooling depends on the empirical similarity of the estimates from the individual units.
3. *Exact likelihoods*. It is not necessary to adopt approximate normal likelihoods, although care may be required in dealing with nuisance parameters (Section 8.2.2).
4. *Allowing for uncertainty in all parameters*. The full uncertainty from all the parameters is reflected in the widths of the intervals for the parameter estimates; these will therefore tend to be wider than those from a classical random-effects analysis.
5. *Allowing for other sources of evidence*. Other sources of evidence can be reflected in the prior distributions for parameters, or in pooling multiple types of study (Section 8.4).
6. *Allowing direct probability statements on different scales*. Quantities of interest can be directly addressed, such as the probability that the true treatment effect in a typical trial is greater than 0. It is also possible to make inferences on a variety of scales, such as risk difference, risk ratio and odds ratio (Carlin, 2000; Warn *et al.*, 2002).

7. *Predictions*. The ease of making predictions within a Bayesian framework allows, for example, current meta-analyses to be used in designing future studies. For example, we may use the basic normal model to predict the treatment effect θ^{new} in a new trial by

$$\theta^{\text{new}} \sim \text{N}[\mu, \tau^2]. \tag{8.2}$$

Rather than making predictions based on the 'plug-in' random-effects distribution $p(\theta^{\text{new}}|\hat{\mu}, \hat{\tau})$, we can use the full predictive distribution

$$p(\theta^{\text{new}}|\text{data}) = \int p(\theta^{\text{new}}|\mu, \tau)\, p(\mu, \tau|\text{data})\, d\mu\, d\tau, \tag{8.3}$$

which fully takes into account the uncertainty concerning μ and τ. This may be easily achieved when using MCMC methods by simulating a value θ^{new} at each iteration; the simulated values form a sample from the full predictive distribution (8.3).

It could be argued that this predictive distribution is a more appropriate summary of the treatment than conclusions regarding the mean effect μ. Such a predictive distribution may also be valuable as the basis for power calculations for confirmatory clinical trials (Section 6.5), and could also act as a prior distribution in their analysis. Predictions of effects in future populations are also required if the analysis is to contribute to a policy model, and these may need to be adjusted for different patient characteristics.

8. *Assessing compatibility between meta-analyses and individual clinical trials*. Suppose we have observed data y^{obs} in a new trial and we wish to assess their compatibility with a meta-analysis. We may consider y^{obs} as providing a likelihood term for a new treatment effect θ^{new}, and the issue becomes one of assessing compatibility between a likelihood and a prior $p(\theta^{\text{new}}|\text{data})$ obtained from (8.3). We have already considered such comparisons in Section 5.8, where Box's method was outlined. This compares y^{obs} with the predictive distribution of new data Y^{new}, given by

$$p(Y^{\text{new}}|\text{data}) = \int p(Y^{\text{new}}|\theta^{\text{new}})\, p(\theta^{\text{new}}|\text{data})\, d\theta^{\text{new}}.$$

Specifically, as a form of two-sided P-value, we calculate twice the minimum tail area $2\min(p(Y^{\text{new}} < y^{\text{obs}}|\text{data}),\ p(Y^{\text{new}} > y^{\text{obs}}|\text{data}))$. This is easily achieved when using MCMC by generating θ^{new}, then generating Y^{new} from $p(Y^{\text{new}}|\theta^{\text{new}})$, and counting the proportion of simulated Y^{new}s that exceed or are less than y^{obs}.

Suppose both prior $p(\theta^{\text{new}}|\text{data})$ and likelihood $p(y^{\text{obs}}|\theta^{\text{new}})$ can be assumed approximately normal with distributions $\text{N}[\hat{\theta}^{\text{new}}, \sigma^2/m]$ and $\text{N}[\theta^{\text{new}}, \sigma^2/n]$ respectively. Then Box's procedure is equivalent to a two-sided test based on a standardised comparison

$$Z = \frac{y^{\text{obs}} - \hat{\theta}^{\text{new}}}{\sigma\sqrt{m^{-1} + n^{-1}}}.$$

Example 8.1 illustrates the comparison of predictions Y^{new} from meta-analyses with observed y^{obs} in new trials, to show the conflict may not be as great as is often claimed – see also Berry (2000).

9. *Cumulative meta-analysis*. It is natural to use a cumulative meta-analysis as external evidence when monitoring a clinical trial (Henderson *et al.*, 1995), and cumulative meta-analysis can also be given a Bayesian interpretation as providing a prior distribution (Lau *et al.*, 1995; see also Section 5.4): in this situation the Bayesian approach relies on the assumption of exchangeability of trials but avoids concerns with retaining Type I error over the entire course of the cumulative meta-analysis.
10. *'Meta-regression'*. It is reasonably straightforward to investigate the relationship between treatment effect and study-level factors. For example, suppose we have measured a covariate x_k on each study. Then we could fit the model

$$\theta_k = \theta_k^{\text{adj}} + \beta(x_k - \overline{x}), \tag{8.4}$$

where θ_k^{adj} is the treatment effect adjusted for the covariate and might be assumed to have a population distribution $\theta_k^{\text{adj}} \sim \text{N}[\mu, \tau^2]$. However, particular care is required for examining the relationship with baseline rates (Section 8.2.3).
11. *Publication bias*. It is feasible to model the effects of different degrees of publication bias, although any conclusions must necessarily be somewhat dependent on uncheckable assumptions (Silliman, 1997; Begg *et al.*, 1997; Givens *et al.*, 1997; Smith *et al.*, 2000).

These methods are not restricted to randomised trials and may equally be applied to meta-analyses of case–control and other observational studies, with the usual caveats about adjustment for potential bias.

Example 8.1 *ISIS: Prediction after meta-analyses*

Reference: Higgins and Spiegelhalter (2002).

Background: Example 3.13 described a meta-analysis carried out in 1993 which showed an apparent survival benefit from magnesium sulphate following myocardial infarction. When the ISIS-4 'megatrial' announced its result of no benefit from magnesium, the apparent conflict with the meta-analysis led to a long-running argument – see Higgins and Spiegelhalter (2002) for a recent analysis. Here we derive a predictive distribution for the effect expected in a new trial based on the data available in

the meta-analysis and presented in Example 3.13, and see whether that prediction is really in conflict with the results observed in ISIS-4. We carry out a full Bayesian analysis on all the parameters, and check sensitivity to prior assumptions.

Statistical model: The normal approximation for the log(odds ratios) described in Section 2.4.1 is adopted.

Prior distribution: As a baseline analysis, μ and τ, the between-study mean and standard deviation, are given uniform priors.

Computation/software: MCMC methods implemented using WinBUGS.

Evidence from study: The data contributing to the meta-analysis were given in Table 3.8. In ISIS-4 2216/29 011 (7.6%) deaths were observed in the magnesium arm, slightly in excess of the 2103/29 039 (7.2%) deaths observed under placebo. This corresponds to a log(OR) of $y^{\text{obs}} = 0.06$, with standard deviation 0.03.

Bayesian interpretation: Summaries of the simulated values of μ and τ are given in Table 8.1 under the uniform prior assumptions. It can be seen that the between-trial heterogeneity is poorly estimated from these data in that the 95% interval is extremely wide, and therefore some prior sensitivity might be expected. Nevertheless the 95% interval for the overall odds ratio does exclude 1. The predicted log(OR) θ^{new} in a new trial has an extremely wide interval, and this is reflected in the predictive distribution of the observed log(OR) Y^{new} in a trial of the size of ISIS-4, which has a point prediction of 0.56 but a 95% prediction interval from 0.10 to 2.43. We note that the huge sample size of ISIS-4 means that the distribution of Y^{new} is essentially the same as θ^{new}. The observed log(OR) of $y^{\text{obs}} = 0.06$ lies well within this interval with a one-sided tail area of 0.12; Box's compatibility measure is the probability of observing such an extreme result,

Table 8.1 Comparison of meta-analysis with megatrial. Y^{new} are the results from a further trial that would be predicted from the meta-analysis. The observed data y^{obs} from ISIS-4 are well within the 95% prediction interval.

	Parameter	Median	95% interval		Median OR	95% interval for OR
μ:	mean effect	−0.59	−1.35 to	−0.01	0.56	0.26 to 0.99
τ:	between-trial SD	0.55	0.02 to	1.62		
θ^{new}:	prediction of effect in new trial	−0.58	−2.28 to	0.89	0.56	0.10 to 2.43
Y^{new}:	prediction of log(OR) to be observed in new trial	−0.59	−2.29 to	0.88	0.56	0.10 to 2.43
y^{obs}:	observed log(OR) in ISIS-4	0.06	0.00 to	0.12	1.06	1.00 to 1.13

$2 \times 0.12 = 0.24$. This analysis does not therefore indicate strong conflict between the meta-analysis and the megatrial.

Sensitivity analyses: Six alternative prior distributions for τ give predictive distributions for Y^{new} shown in Figure 8.1. As expected from the discussion in Section 5.7.3, the Gamma(0.001,0.001) (a) (equivalent to a root-inverse-gamma on τ), DuMouchel (e) and half-normal with $\tau_u = 1.0$ (f) tend to support smaller values of τ and hence produce narrower posterior intervals, while the uniform on τ^2 (b) leads to very wide intervals. We note that $s_0 = 0.36$, roughly corresponding to an average of 31 events per trial (in fact a total of 286 events are recorded in Table 3.8, or an average of 36 events per trial).

The resulting one-sided *P*-values $P(Y^{new} < y^{obs})$ ranged from 0.06 (for (a) and (f)) to 0.18 (for (b)), so under no assumption was there particularly strong evidence of incompatibility.

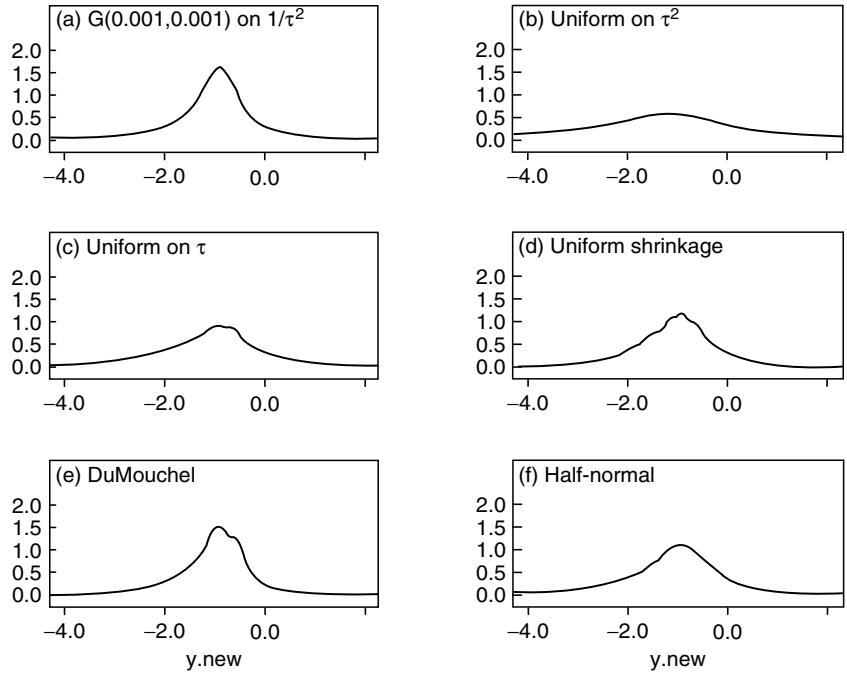

Figure 8.1 Alternative predictive distributions for the observed log(OR) in a trial the size of ISIS-4, arising from six different prior distributions on τ. The actual observed log(OR) was 0.058, and hence was not seriously in conflict with any of the predictive distributions.

8.2.2 Some delicate issues in Bayesian meta-analysis

The Bayesian approach to meta-analysis promises additional flexibility but raises some tricky issues, some of which are generic to hierarchical models and some more specific to this context. These include the following:

The between-study standard deviation τ. Comparative studies show that when there are few studies and hence τ cannot be accurately estimated from the data alone, the prior for this parameter may become important and the empirical Bayes approach, in which the uncertainty about the between-study variability is ignored, tends to provide intervals that are too narrow. Priors on the heterogeneity parameter have already been discussed in Section 5.7.3, in which it was noted that Higgins and Whitehead (1996) use proper priors derived from a series of meta-analyses. It is important to check the sensitivity to the prior on τ – see Example 8.1.

Exact likelihoods and nuisance parameters. The standard normal approximation given in (8.1) may not be appropriate when the studies are small or their results extreme, as the resulting likelihoods may not be approximately normal. For example, suppose in the kth trial there are n_{tk} and n_{ck} in the treatment and control groups respectively, and we observe r_{tk} and r_{ck} deaths. If either n_{tk} and n_{ck} is small, or mortality rates are near 0% or 100%, we may adopt a full binomial model instead of the normal approximation of Section 2.4. Specifically, we assume

$$\begin{aligned} r_{tk} &\sim \text{Bin}[p_{tk}, n_{tk}], \\ r_{ck} &\sim \text{Bin}[p_{ck}, n_{ck}], \end{aligned}$$

where the mortality probabilities are expressed as

$$\begin{aligned} \text{logit}(p_{tk}) &= \phi_k + \theta_k, \\ \text{logit}(p_{ck}) &= \phi_k. \end{aligned} \tag{8.5}$$

Hence ϕ_k is the logit(mortality rate) in the control group of trial k, and the treatment effect θ_k is the log(odds ratio).

The ϕ_k can also be called 'study effects' or 'baseline rates' and require careful handling. Generally they will be considered as nuisance parameters, except in the situation where a relationship between treatment effect and underlying risk is suspected (Section 8.2.3). Eliminating such nuisance parameters is a problem within all schools of statistical inference: see Section 3.18 for a brief review.

In the context of meta-analysis the following methods have been adopted:

- *'Approximate pivotal quantity'*. The standard normal approximation in (8.1) has a distribution which does not depend on the baseline ϕ_k.

- *'Conditional likelihood'*. By conditioning on the value of a statistic we derive a likelihood which depends only on the parameter of interest: see Liao (1999) for a Bayesian application of this procedure in meta-analysis.
- *Prior distributions*. The appropriate joint prior distribution for the ϕ_k and the θ_k presents a particular problem. The 'study effects' ϕ_k might be given independent uniform priors, but a choice must be made between the logit (ϕ_k) and probability (p_{ck}) scale. Random study effects can be assumed if the control group risks are considered exchangeable, but a normal distribution may not be appropriate. Finally, it may be reasonable to assume the ϕ_k and the θ_k are correlated, and hence carry out a 'bivariate meta-analysis' (van Houwelingen *et al.*, 1993). This is essential if one is explicitly investigating the relationship between effect and baseline risk (Section 8.2.3), but it has been argued that it would be appropriate in any situation in which one assumes random ϕ_k. The reasoning is as follows: if the ϕ_k and the θ_k are assumed independent, (8.5) shows that the variance of the treatment risks is forced to be greater than the variance of the control risks. Of course this may be a reasonable assumption, but it should be explicitly acknowledged.

Example 8.2 examines a meta-analysis of trials with rare events, and explores the sensitivity of conclusions to a range of these modelling options.

Example 8.2 *EFM: meta-analyses of trials with rare events*

References: Sutton and Abrams (2001), Sutton *et al*. (2002).

Intervention: Electronic foetal heart rate monitoring (EFM) in labour, with the aim of early detection of altered heart-rate pattern and hence a potential benefit in perinatal mortality.

Aim of study: EFM was gradually introduced in the early 1970s, and early evaluation of its impact in terms of perinatal death was in terms of either non-randomised comparative studies or before–after studies. A large body of evidence was collected which suggested that EFM was indeed clinically effective in reducing the risk of perinatal death. Despite this body of evidence a number of randomised trials were conducted, which were much smaller in terms of sample sizes, but which suggested that there was little benefit, if any, from the use of EFM. Here we consider the evidence from the randomised trials, with emphasis on the difficulties associated with rare events.

Study design: Meta-analysis of nine randomised trials.

Outcome measure: Perinatal mortality, as measured by the odds ratio in deaths per 1000 births, odds ratios less than 1 favouring EFM. We note that Sutton and Abrams (2001) consider the risk difference, which is

directly related to the number needed to treat (NNT) and hence a policy decision (Section 3.14).

Statistical model: There are a number of options for dealing with the nuisance parameters in this model, *i.e*. the control group risks (Section 3.18), acknowledging that the standard normal approximation for the log (odds ratio) likelihood within each study may be inappropriate due to the rarity of perinatal deaths.

(a) *Fixed effects*. A normal approximation to the likelihood for the observed log(odds ratio) (Section 2.4), with the log(odds ratios) θ_k assumed to be independent.

(b) *Approximate normal likelihood, random effects*. A normal approximation to the likelihood for the observed log(odds ratio) (Section 2.4), with the log(odds ratios) assumed to have the distribution $\theta_k \sim \mathrm{N}[\mu, \tau^2]$.

(c,d) *Binomial likelihood, random effects*. An exact binomial model (8.5), with the log(odds ratios) assumed to have the distribution $\theta_k \sim \mathrm{N}[\mu, \tau^2]$. The control group risks are assumed independent, with options (c) and (d) representing different assumptions (see below).

An exchangeable model for the control group risks could also have been adopted.

Prior distribution: μ and τ, the between-study mean and standard deviation, are given uniform priors. For the full binomial models (c) and (d), two alternative priors for each study's control group mortality p_{ck} are considered: (c) p_{ck} is given an independent uniform prior, and (d) $\phi_k = \mathrm{logit}(p_{ck})$ is given an independent uniform prior.

Computation/software: MCMC methods implemented using WinBUGS.

Evidence from study: The randomised data are presented in Figure 8.2. We note that trial 8 has a high mortality rate in the control group, which would cast doubt on a simplistic normal assumption for exchangeable control groups risks. The 0s in trials 3 and 6 also suggest that conclusions may be sensitive to ways of dealing with the nuisance parameters.

Bayesian interpretation and sensitivity analyses: Figure 8.2 shows the estimated odds ratios for each trial and for the population, for each of the four models (a) to (d). The approximate normal random-effects model (b) is consistently more conservative in its estimate than the models using a binomial likelihood, and also more precise. The binomial model (d) with a uniform prior on the logit of the control risks is more conservative than model (c) with a uniform prior on the control risks – this is presumably because model (d) will tend to estimate smaller control risks than model (c) and hence will reduce any apparent benefit of EFM.

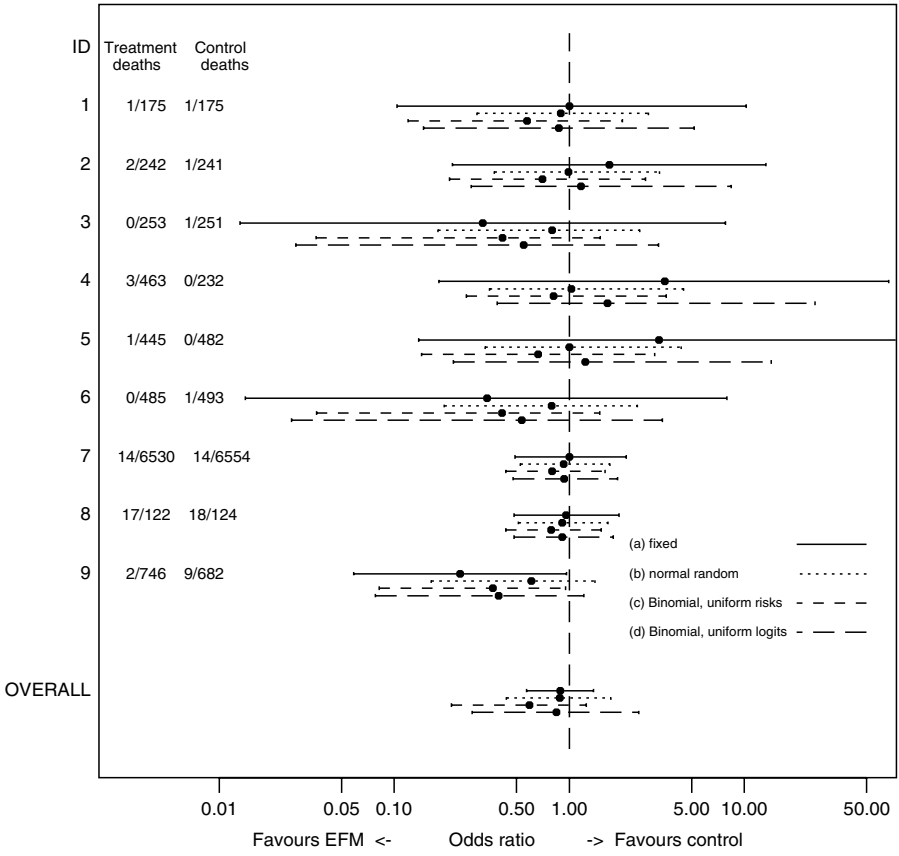

Figure 8.2 Four different models for a meta-analysis of nine trials of electronic foetal monitoring. The rare events lead to considerable sensitivity of conclusions to assumptions concerning the form of the likelihood and prior distributions on nuisance parameters.

Table 8.2 shows that the three random-effects models also give rise to different estimates of τ, although each has a wide interval with the bulk of the density near 0. There is likely to be considerable additional sensitivity to prior assumptions concerning τ.

Comments: This example shows there can be sensitivity to likelihood assumptions as well as prior distributions, and that analyses with rare events have to be handled with care. In particular, the traditional normal approximation, used in so many of our examples, would lead to excessive confidence in the conclusion, whereas the RCTs provide little evidence of efficacy on their own.

Table 8.2 Posterior summaries for between-trial standard deviation τ from three different random-effects models.

Model	Median of τ	95% interval
(b) Approximate normal likelihood, random effects:	0.32	0.01 to 1.50
(c) Binomial likelihood, random effects, uniform on control risks:	0.54	0.01 to 2.25
(d) Binomial likelihood, random effects, uniform on logit control risks:	0.75	0.09 to 2.82

Sutton and Abrams (2001) present both case–control and cohort data addressing this comparison: randomised and observational data could be combined by, for example, using the (possibly discounted) observational data as a prior for the meta-analysis presented above (Hornbuckle *et al*., 2000), or by conducting a generalised evidence synthesis in which different study designs are pooled in a hierarchical model (Section 8.4).

8.2.3 The relationship between treatment effect and underlying risk

The appropriate means of modelling the dependence of effect on baseline risk has been the subject of some controversy. There is general agreement that it is natural to investigate the linear model

$$\theta_k = \theta_k^{\text{adj}} + \beta(\phi_k - \overline{\phi}), \tag{8.6}$$

where θ_k^{adj} is now the treatment effect adjusted for a measure of baseline risk ϕ_k, also known as a 'study effect'. θ_k^{adj} might be assumed to have a distribution

$$\theta_k^{\text{adj}} \sim \text{N}[\mu, \tau^2]. \tag{8.7}$$

We note from (8.6) and (8.7) that the treatment effect θ_k has distribution

$$\theta_k \sim \text{N}[\mu + \beta(\phi_k - \overline{\phi}), \tau^2], \tag{8.8}$$

and hence the treatment effect in any future trial with true baseline risk ϕ can be obtained by substitution in (8.8). In particular, the effect is expected to be 0 when ϕ obeys

$$\phi_0 = \frac{-\mu}{\beta} + \overline{\phi};$$

the solution to this equation is known as the 'breakeven' point. MCMC methods allow inferences to be drawn about this quantity, as demonstrated in Example 8.3. Such models have been investigated by McIntosh (1996), Thompson *et al.* (1997), Sharp and Thompson (2000) and Arends *et al.* (2000).

The controversy arises in the specification of a prior for the 'study effects' ϕ_k. Thompson *et al.* (1997) assume independent priors and hence fixed study effects, but this is strongly criticised by Houwelingen and Senn (1999), who argue that since this introduces an additional nuisance parameter for each trial, the procedure will be 'inconsistent' in the sense that under broad assumptions it will, as the number of trials grows, not tend to give the correct underlying relationship. In their reply the authors claim that fixed study effects are standard methodology, for example in using logistic regression, and will only give misleading conclusions in extreme situations. These alternative approaches are investigated in Example 8.3.

Van Houwelingen and Senn (1999) also make the important point that there will always, in a sense, be dependence between effect and baseline, since if there is no relationship on a logit scale, there would be on an absolute risk scale. An important aim may therefore be to find a scale on which the effect is most independent of baseline.

Example 8.3 *Hyper: Meta-analyses of trials adjusting for baseline rates*

References: Hoes *et al*. (1995) and Arends *et al*. (2000).

Intervention: Drug treatment in mild to moderate hypertension.

Aim of study: To determine whether drug treatment reduced mortality and to see whether the size of the treatment effect depended on the event rate in the control group.

Study design: Meta-analysis of 12 randomised trials with considerable variability in baseline risk.

Outcome measure: All-cause mortality per 1000 patient-years of follow-up.

Statistical model: A random-effects Poisson regression model was assumed. In a similar manner to Section 3.18, for the *i*th study the numbers of deaths r_{ti} and r_{ci} in treatment and control groups are assumed

$$r_{ti} \sim \text{Poisson}(m_{ti}),$$
$$r_{ci} \sim \text{Poisson}(m_{ci}),$$

using the notation of Section 2.6.2. The Poisson means are expressed as

$$m_{ti} = \log(n_{ti}/1000) + \phi_i + \theta_i,$$
$$m_{ci} = \log(n_{ci}/1000) + \phi_i,$$

where n_{ti} and n_{ci} are the patient-years of follow-up in the treatment and control groups. Hence ϕ_i is the log of the rate per 1000 patient-years in the control group of trial i, and the treatment effect θ_i is the log(rate ratio).

The dependence of treatment effect on baseline rate is then modelled exactly as described in Section 8.2.3.

Prior distribution: For the baseline analysis, μ and τ, the between-study mean and standard deviation, are given uniform priors. Following the discussion in Section 8.2.3, two priors are considered for each study's control log(event rate) ϕ_i: independent uniform priors, and exchangeable with a normal distribution

$$\phi_i \sim \mathrm{N}\left[\mu_\phi, \tau_\phi^2\right],$$

where μ_ϕ, τ_ϕ are given uniform priors.

Computation/software: MCMC methods implemented using WinBUGS.

Evidence from study: The data are given in Table 8.3. Figure 8.3(a) shows the observed rate ratios from Table 8.3 plotted against the observed control group rates. There is a clear suggestion of a relationship.

Bayesian interpretation: Figure 8.3(b) shows the estimated rate ratios e^{θ_i} plotted against the estimated control group rates e^{ϕ_i} when adjusting for baseline, assuming independent uniform priors for the ϕ_i. There is clear shrinkage towards the assumed straight line, with the control group rate for centre 2 estimated to be even smaller than that observed. The intersection

Table 8.3 Data from 12 randomised trials of drug treatment for mild-to-moderate hypertension: r is the number of deaths, n is the patient-years of follow-up, and rates are events per 1000 patient-years.

Treatment group			Control group		
r_t	n_t	rate$_t$	r_c	n_c	rate$_c$
10	595.2	16.8	21	640.2	32.8
2	762.0	2.6	0	756.0	0.0
54	5 635.0	9.6	70	5 600.0	12.5
47	5 135.0	9.2	63	4 960.0	12.7
53	3 760.0	14.1	62	4 210.0	14.7
10	2 233.0	4.5	9	2 084.5	4.3
25	7 056.1	3.5	35	6 824.0	5.1
47	8 099.0	5.8	31	8 267.0	3.7
43	5 810.0	7.4	39	5 922.0	6.6
25	5 397.0	4.6	45	5 173.0	8.7
157	22 162.7	7.1	182	22 172.5	8.2
92	20 885.0	4.4	72	20 645.0	3.5

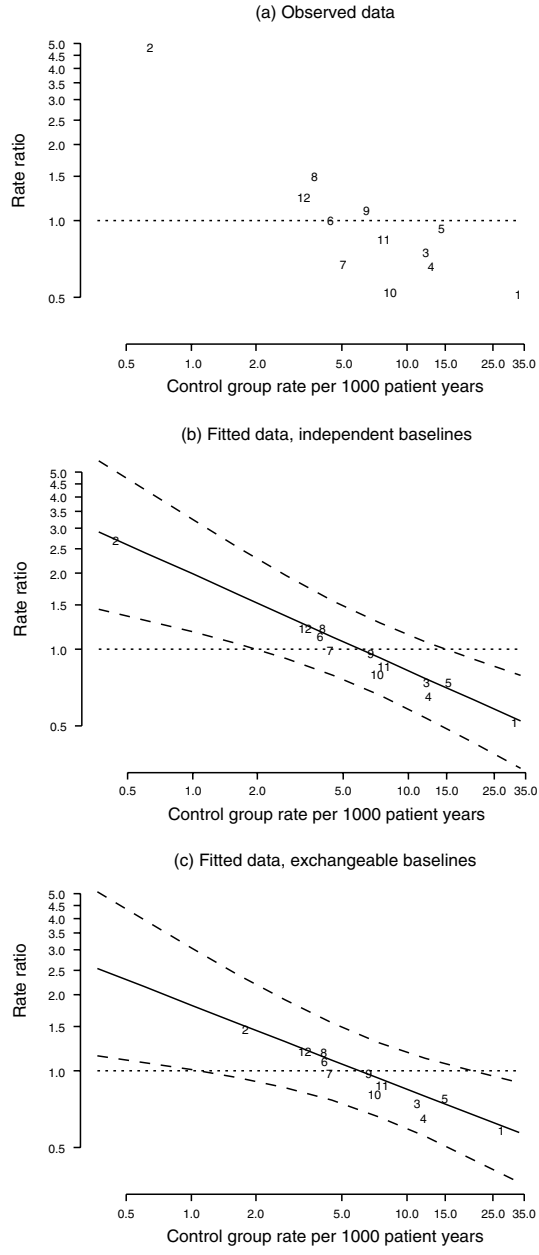

Figure 8.3 Estimated control group rates and rate ratios in 12 studies under different assumptions. (a) can be considered as fixed-effect estimates of control rate and treatment effects. In (b), the treatment effect is assumed linearly related to independent log(control group rates), whereas in (c) the log(control group rates) are assumed exchangeable and hence shrunk towards a common value.

Table 8.4 Results from fitting independent and exchangeable control group rates.

		Independent control rates		Exchangeable control rates	
	Parameter	Median	95% interval	Median	95% interval
β	Dependence on baseline	−0.38	−0.57 to −0.17	−0.33	−0.55 to −0.09
e^{ϕ_0}	'Breakeven' control rate	6.00	3.67 to 8.01	6.06	2.73 to 8.80
τ	Residual SD	0.10	0.01 to 0.28	0.10	0.01 to 0.30

of the upper and lower prediction intervals with the null rate ratio 1 corresponds to the interval for e^{ϕ_0}, the control group rate at which there is no treatment effect. The corresponding estimates are shown in Table 8.4.

Figure 8.3(c) shows the consequences of assuming the control rates are exchangeable: the estimates are shrunk towards a common value, particularly the smaller study 2. The reduced spread in the control group rates with the exchangeable analysis has resulted in increased uncertainty.

After adjusting for baseline risk, there is very little residual between-study heterogeneity suggesting it may be reasonable to set $\tau = 0$ and assume all heterogeneity is explained by baseline risk.

Sensitivity analyses: Alternative priors for the between-study standard deviation τ have little influence on this analysis.

Comments: Acknowledging functional dependence of treatment and baseline rates brings about a reduction in the apparent gradient, compared with that obtained by plotting the raw data. Assuming exchangeable control group rates brings some shrinkage but has little influence on the conclusions. There is little residual variability around the fitted line.

8.3 INDIRECT COMPARISON STUDIES

Suppose that a number of experimental interventions are investigated in a series of studies, where each study compares a subset of the interventions with a control group. We would like to draw inferences on the treatment effects compared with control, and possibly also make comparisons between treatments that may well never have been directly compared. We shall call these *indirect* comparisons, although the term *mixed* comparisons has also been used. Song *et al.* (2003) carry out an empirical investigation and report that such comparisons arrive at essentially the same conclusions as 'head-to-head' comparisons.

A specific application arises in the context of 'active control' studies. Suppose an established treatment C exists for a condition, and a new intervention T is

being evaluated. The efficacy of T would ideally be estimated in randomised trial with a placebo P as the control group, but because of the existence of C this may be considered unethical. Hence C may be used as an 'active control' in a head-to-head clinical trial, and inferences about the efficacy of T may have to be estimated indirectly, using past data on comparisons between C and P.

Let ϕ_{jk} represent the expected response (on an appropriate scale) of treatment j being given in study k, where the control is labelled as $j = 0$. A simple model might express ϕ_{jk} as

$$\phi_{jk} = \phi_k + \theta_{jk}, \tag{8.9}$$

where ϕ_k denotes a 'study effect' and θ_{jk} a treatment effect in the kth study. It is often convenient to set $\theta_{0k} = 0$, so that we can interpret ϕ_k as the response in the control group. Equation (8.9) needs to be further constrained in order to estimate parameters: we might assume a common treatment effect across all studies $\theta_{jk} = \theta_j$, or a random effect in which the θ_{jk} are assumed drawn from some population distribution, say, $\theta_{jk} \sim \mathrm{N}[\theta_j, \tau_j^2]$ (Higgins and Whitehead, 1996; Hasselblad, 1998). A variety of models are possible for the distributions of the ϕ_k and θ_{jk}: Higgins and Whitehead (1996) point out that if we wish the contrasts between all possible treatment pairs (including control) to have the same distribution, then we need to assume a multivariate normal distribution for the θ_{jk} with a particular correlation structure. Example 8.4 re-examines a published example of such an analysis.

Example 8.4 *Blood pressure: Estimating effects that have never been directly measured*

Reference: Gould (1991).

Intervention: Alternative therapies for lowering blood pressure.

Aim of study: To estimate the contrast between two therapies that have never been compared head-to-head. Gould (1991) suggests such an inference could then be used to design a direct comparison study.

Available evidence: Table 8.5 displays the results from a set of eight crossover experiments comprising randomised comparisons and single-arm studies (Gould, 1991), showing mean and standard deviation of change in blood pressure, and sample size in each group. Four treatments (control, *A*, *B* and *C*) have been given, but there has been no direct comparison between treatments *A* and *B* and it is this contrast that is of particular interest.

Statistical model: Let y_{jk} be the mean response recorded in Table 8.5 for the jth treatment in the kth study. We assume

Table 8.5 Sample sizes m, mean and standard deviation of responses under each treatment given in eight studies: e.g. study 1 compared A with C, while study 2 randomised between control and B in a 1:2 ratio. The problem is to compare treatments A and B.

	Control ($j = 0$)			$A(j = 1)$			$B(j = 2)$			$C(j = 3)$		
Study	m	Mean	SD	m	Mean	SD	m	Mean	SD	m	Mean	SD
1				41	8.90	7.49				39	6.05	10.28
2	47	5.51	8.72				100	6.21	8.02			
3	53	3.75	7.07	54	10.20	9.39						
4	47	3.04	9.20	44	8.43	8.17						
5	30	2.97	7.69				32	6.53	7.80			
6	69	3.99	8.04									
7	68	5.28	7.58									
8	67	3.34	8.01									

$$y_{jk} \sim \mathrm{N}\left[\phi_{jk}, \frac{\sigma^2}{m_{jk}}\right],$$

and assume $\phi_{jk} = \phi_k + \theta_j$ (8.9), where $\theta_0 = 0$ so that ϕ_k is the response in the control group in study k (although there was not necessarily an actual control in the kth study) and θ_1, θ_2, θ_3 measure the mean effects of A, B, C over placebo, respectively. Some of the studies have only a single arm, and if we assume fixed study effects then these will contribute no information (except in contributing to the estimate of σ^2). Since all the studies were carried out in a common research programme by the same investigators, it may be reasonable to adopt exchangeable study effects ϕ_k, with

$$\phi_k \sim \mathrm{N}\left[\mu_\phi, \tau_\phi^2\right].$$

The treatment effects θ_1, θ_2, θ_3 are taken as independent fixed effects. We may use the following distribution theory to obtain a likelihood for σ (Section 2.6.5). The observed standard deviations s_{jk} have the property

$$\frac{(m_{jk} - 1)s_{jk}^2}{\sigma^2} \sim \chi^2_{m_{jk}-1},$$

and hence $(m_{jk} - 1)s_{jk}^2 \sim \Gamma((m_{jk} - 1)/2,\ 1/(2\sigma^2))$.

Prospective analysis?: No.

Table 8.6 Posterior summaries.

	Parameter	Median	SD	95% interval
μ_δ	Control mean	4.01	0.50	3.00 to 4.98
θ_1	*A*	9.37	0.79	7.87 to 10.98
θ_2	*B*	6.10	0.87	4.28 to 7.73
θ_3	*C*	6.92	1.08	4.83 to 9.07
$\theta_1 - \theta_2$	*A* vs. *B*	3.28	1.16	1.08 to 5.68
σ	sampling sd	8.18	0.22	7.79 to 8.63
τ_ϕ	between-study sd	0.46	0.48	0.02 to 1.78

Prior distribution: Uniform distributions are given to $\log(\sigma)$, μ_ϕ, τ_ϕ and each of the θ_j.

Loss function or demands: None specified.

Computation/software: MCMC implemented in WinBUGS, with inferences based on 10 000 iterations after a burn-in of 1000.

Bayesian interpretation: The results are shown in Table 8.6, revealing the between-study standard deviation τ_ϕ to have a wide interval. The indirect analysis allows a posterior distribution to be obtained for $\theta_1 - \theta_2$ which might be used in designing a suitable trial for a direct comparison of *A* and *B*.

8.4 GENERALISED EVIDENCE SYNTHESIS

As noted when discussing observational studies in Chapter 7, in some circumstances randomised evidence will be less than adequate due to economic, organisational or ethical considerations (Black, 1996). Considering all the available evidence, including that from non-randomised studies, may then be necessary or advantageous. Droitcour *et al.* (1993) describe the limitations of using either RCTs or databases alone, in that RCTs may be rigorous but restricted, whereas databases have a wider range but may be biased. They introduce what they term *cross-design synthesis*, an approach for synthesising evidence from different sources, with the aim 'not to eliminate studies of overall low quality from the synthesis, but rather to provide the information needed to compensate for specific weaknesses'. Although not a strictly Bayesian approach, they are essentially explicitly modelling potential biases (Section 7.3), and then attempting to generalise the results of clinical trials for broader populations. Rubin (1992) emphasises pooling evidence through modelling in order to 'build and extrapolate a response surface', which models the true treatment effect conditional on both the design of the study and subgroup factors.

Cross-design synthesis was outlined in a report from the US General Accounting Office (General Accounting Office, 1992), but a *Lancet* (1992) editorial was

critical of this approach, suggesting it would deflect attention from carrying out serious controlled trials: this was denied in a subsequent reply by Chelimsky *et al.* (1993). A commentary by Begg (1992) suggested they had underestimated the difficulty of the task, and appeared to assume that randomised trials and databases could be reconciled by statistical adjustments, whereas selection biases and differences in experimental rigour could not be eliminated so easily. A non-Bayesian case study is provided by Belin *et al.* (1995) who combine observational databases in order to evaluate interventions to increase screening rates, but need to impute missing data in some studies.

One must clearly be very cautious in such an endeavour, balancing the desire to make use of all available evidence with due acknowledgement of potential weaknesses. It is not a purely technical exercise, and must be carried out in loose collaboration with subject-matter experts. Nevertheless, it is natural to take a Bayesian approach to the synthesis of multiple study designs, in which relationships are assumed between some underlying parameters of the different studies. Such relationships may involve a huge variety of both deterministic models and probabilistic dependence, and again fall naturally into the taxonomy of relationships already explored in the use of historical data (Section 5.4)

(a) *Irrelevance*. It is always an option, possibly on purely subjective grounds, to declare certain studies irrelevant to the issue under study.
(b) *Exchangeable*. Typically we may be able to classify our studies according to a 'type', say randomised, case–control or cohort: this naturally leads to hierarchical exchangeability assumptions, which can specifically allow for the quantitative within- and between-study-type heterogeneity, and incorporate prior beliefs regarding qualitative differences between the various sources of evidence. Figure 8.4 shows a stylised graphical representation of a possible model, in which treatment effects are assumed exchangeability within study type, and also that mean study effects are exchangeable. Examples of this approach include Prevost *et al.* (2000) who pool randomised and non-randomised studies on breast cancer screening (Example 8.5), Larose and Dey (1997) who similarly assume open and closed studies are exchangeable, and Dominici *et al.* (1999) who examine migraine trials and pool open and closed studies of a variety of designs in a four-level hierarchical model. There is a clearly a difficulty in making such exchangeability assumptions, since there are few study types and hence little information on the variance component. Prior assumptions may be very important, and priors for the degree of 'similarity' between alternative designs might be empirically informed by studies comparing the results of RCTs and observational data, such as listed in Section 7.3.
(c) *Potential biases* and (d) *Equal but discounted*. Both biases and discounting can be incorporated into a model for between- and within-study-type variation such as that shown in Figure 8.4.

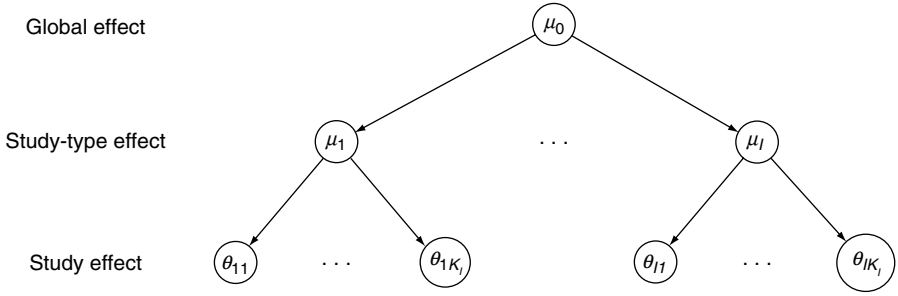

Figure 8.4 Hierarchical model in which the effects θ_{ij} in studies of type i are assumed exchangeable with mean μ_i, and the study-type effects μ_i are assumed exchangeable with mean μ_0.

(e) *Functional dependence*. Suppose we are interested in drawing inferences on a quantity f about which no direct evidence exists, but where f can be expressed as a deterministic function of a set of 'fundamental' parameters $\theta = \theta_1, \ldots, \theta_N$. For example, f might be the response rate in a new population made up of subgroups about which we do have some evidence. More generally, we might assume we have available a set of K studies in which we have observed data $y_1, \ldots, y_K$ which depend on parameters $\psi_1, \ldots, \psi_K$, where each ψ_k is itself a function of the fundamental parameters θ. This structure is represented graphically in Figure 8.5. This situation sounds very complex but in fact is rather common, when we have a lot of studies, each of which informs part of a jigsaw, and which need to be put together to answer the question of interest. See Example 8.6 for a case where the fundamental parameters have directly relevant evidence, and Example 8.7 in which the fundamental parameters have only indirect evidence.

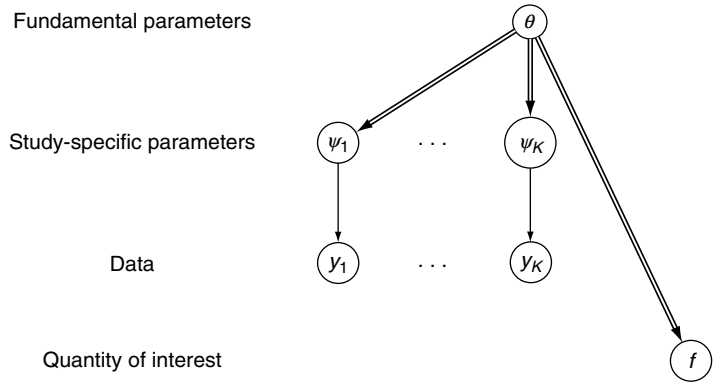

Figure 8.5 Data y_k in each of K studies depend on parameters ψ_k, which are known functions of fundamental parameters θ. We are interested in some other function f of θ, and so need to propagate evidence from the y_k.

(f) *Equal*. It is of course possible to assume the treatment effect is common across studies of different designs. For example, Li and Begg (1994) present a non-Bayesian analysis of pooling controlled and single-arm studies, in which each is assumed to have a common treatment effect but the study effect is taken as random – this is essentially an application of the indirect comparison models considered in Section 8.3, in which some of the studies are non-comparative since only one treatment is given.

Such models allow enormous room for imagination and complexity, and graphical representations (Spiegelhalter, 1998) have been found to be very useful in clarifying the underlying structure. There is also considerable flexibility in the logical and stochastic assumptions: for example, Dominici *et al*. (1999) assume that between-study variability follows a 'mixture of normals' distribution to allow for skewness. Nevertheless, such analyses may be controversial, since there may be strong dependence on assumptions and there is concern that including studies with 'poor' designs will weaken the analysis. Careful sensitivity analyses are clearly vital, and perhaps one reason for the limited uptake of such syntheses is that they are not seen as 'clean' methods, with each analysis being context-specific, less easy to set quality markers for, easier to criticise as subjective and so on.

Example 8.5 *Screen: generalised evidence synthesis*

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

Intervention: Mammographic screening for breast cancer.

Aim of study: Breast cancer has the potential to be particularly amenable to screening in that RCTs and observational studies clearly indicate that prognosis is extremely good for early stage tumours, especially in women over 50 years of age. In order to assess the magnitude of this potential benefit, a number of RCTs and observational studies have been conducted world-wide. Whilst it is accepted that RCTs provide a 'gold standard' by which to assess efficacy, it has been argued that the inclusion of observational evidence may help in the estimation of effectiveness that may be seen in a potential population. However, observational studies are often subject to various biases and therefore any synthesis must be flexible enough to allow these to be incorporated. This study therefore developed a hierarchical Bayesian model in which prior opinions regarding the relative plausibility of different sources of evidence may also be included.

Study design: Synthesis of evidence from five RCTs and five observational studies which evaluated screening in women over 50.

Outcome measure: Breast cancer mortality per 1000 patient-years.

Statistical model: The three-level model follows that shown in Figure 8.4. Let y_{ik} be the observed log(risk ratio) in the ith study of type k, where $k = 1$ (RCT), 2 (observational), and σ_{ik}^2 its associated variance. Then we assume

$$\begin{aligned} y_{ik} &\sim \text{N}[\theta_{ik}, \sigma_{ik}^2], \\ \theta_{ik} &\sim \text{N}[\mu_k, v_k^2], \\ \mu_k &\sim \text{N}[\mu_0, \tau^2]. \end{aligned} \tag{8.10}$$

The θ_{ik} represent the underlying effect, on the log(risk ratio) scale, in the ith study of type k. The θ_{ik} are distributed about an overall effect for the kth type of study, μ_k, with v_k^2 representing the between-study variability for those studies of type k. At the third level of the model the study-type effects are distributed about an overall population effect, μ_0, with τ^2 representing the between-study-type variability. As with many other meta-analytic models the level 1 variances, σ_{ik}^2, can be replaced by the estimated sample variances s_{ik}^2, derived in this case using the methods described in Section 2.4.3. In this case prior distributions are required for μ_0, τ^2 and the v_k^2.

Prospective analysis?: No.

Prior distribution: A prior distribution for each of the v_k^2 is derived using the techniques described in Section 5.7.3. We assume we are 95% sure that the true underlying risk ratio for a study of a particular type will be within a range from four times to a quarter the overall risk ratio of that type, which means that the upper 95% point of the prior distribution for each v_k is $\log(16)/(2 \times 1.96) = 0.71$. A half-normal distribution (Section 2.6.7) $v_k \sim \text{HN}[0.36^2]$ has this property.

In a similar manner a prior for the between-type variance, τ^2, can be derived from assuming 95% belief that the underlying risk ratio for a particular study type will be less than double or more than half the overall population effect. On this basis, a half-normal prior distribution $\tau \sim \text{HN}[0.18^2]$ is obtained.

For μ_0, the overall population effect, a relatively vague prior distribution is specified on the basis that the overall relative risk is unlikely to exceed 500 in favour of either screening or control, and therefore a prior distribution for μ_0 has standard deviation $\log(500)/1.96 = 3.17$, or $\mu_0 \sim \text{N}[0,10]$.

Loss function or demands: None used.

Computation/software: MCMC in WinBUGS.

Evidence from study: Figure 8.6 displays the observed risk ratios (together with 95% confidence intervals) for the five RCTs and five observational studies.

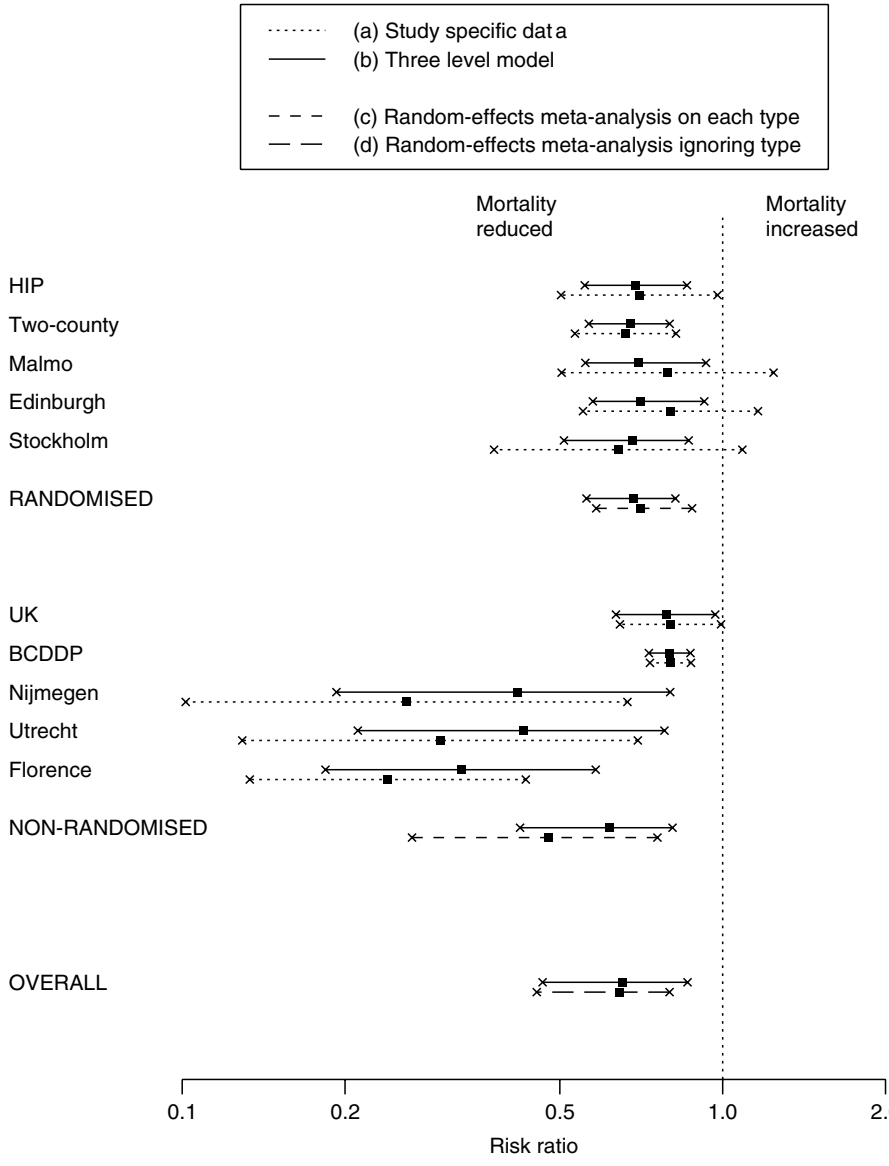

Figure 8.6 Observed risk ratio of breast cancer mortality in RCTs and observational studies in women over 50, together with Bayesian estimates of overall synthesis.

Bayesian interpretation: Figure 8.6 also displays the results, in terms of estimates and 95% intervals, of applying model (8.10) using the prior distributions derived above. In terms of the individual study estimates

there is the usual shrinkage towards the overall study-type estimates, the degree of shrinkage dependent upon the within-study variances, and towards the overall population estimate for the study-type overall estimates. The overall population estimate is very little different from the overall RCT estimate, but the 95% interval for the population effect is considerably larger than that for the RCTs. The key point is that the effect of synthesising both RCT and observational evidence has not been to change our overall estimate of the effectiveness of breast cancer screening, but rather to be less certain about this estimate.

Sensitivity analysis: Table 8.7 shows the results of changing the prior distributions for the variance parameters used in the analysis above, together with that for μ_0, the overall population effect. As an alternative to the prior distributions described above for the variance parameters, uniform distributions over the range 0 to 5 are assumed on a standard deviation scale, and the prior distribution for μ_0 is made even more diffuse. The prior distribution for τ has the largest effect on the estimates for μ_0, μ_1 and μ_2, which is due to the fact that there are only two study types in this example, and therefore relatively little data on which to estimate τ^2.

A further sensitivity analysis was undertaken by Prevost *et al*. (2000) regarding the plausibility of introducing the observational evidence at all into the analysis. In a manner similar to the discounting of historical evidence (Section 5.4), they considered letting v_2, the between-study standard deviation for the observational studies, be a function of v_1 the between-study standard deviation of the RCTs, *i.e.* $v_2 = a \times v_1$. In this

Table 8.7 Sensitivity analysis of estimates of population risk ratio, e^{μ_0}, pooled risk ratio for randomised studies, e^{θ_1}, and pooled risk ratio for observational studies, e^{θ_2} (95% credible interval), under different prior distributions.

Prior for τ	Prior for $v_j (j = 1, 2)$	Prior for μ_0	
		N(0,10)	N(0,10 000)
HN(0.033)	HN(0.125)	e^{μ_0}: 0.65 (0.46, 0.86) e^{θ_1}: 0.68 (0.56, 0.82) e^{θ_2}: 0.62 (0.42, 0.81)	0.65 (0.47, 0.90) 0.68 (0.56, 0.83) 0.61 (0.41, 0.84)
	U(0,5)	e^{μ_0}: 0.65 (0.44, 0.92) e^{θ_1}: 0.69 (0.53, 0.85) e^{θ_2}: 0.62 (0.39, 0.88)	0.65 (0.44, 0.92) 0.69 (0.53, 0.85) 0.62 (0.39, 0.88)
U(0,5)	HN(0.125)	e^{μ_0}: 0.61 (0.24, 1.47) e^{θ_1}: 0.70 (0.57, 0.88) e^{θ_2}: 0.52 (0.30, 0.80)	0.80 (0.19, 13.15) 0.70 (0.56, 0.87) 0.49 (0.26, 0.80)
	U(0,5)	e^{μ_0}: 0.59 (0.15, 1.47) e^{θ_1}: 0.70 (0.57, 0.85) e^{θ_2}: 0.50 (0.22, 1.00)	0.67 (0.28, 3.64) 0.70 (0.58, 0.86) 0.52 (0.21, 0.99)

case *a* can be used to represent beliefs about the relative credibility of the two types of evidence. As an illustration they consider placing a N[3,1] prior distribution on *a*, which corresponds to prior beliefs that the RCTs could be 'valued' three times as highly as the observational studies, but that is also consistent with them being valued as much as five times the observational studies or in fact on an equal basis with the RCTs. Re-estimating the overall population relative risk incorporating this prior distribution yields an estimate of 0.66 with 95% credible interval from 0.47 to 0.92. As with the main three-level analysis above, the point estimate is similar to the overall population relative risk, but the uncertainty surrounding this estimate is now greater than both one based on only the RCTs and a full Bayesian three-level model.

Comments: A wide range of models could be applied to these data. For example, an alternative approach would be to use the observational evidence as a prior distribution for a likelihood based on only the RCT evidence. The model could also be extended to include covariates, and allow prediction on new populations. Nevertheless, there may be difficulties in overcoming suspicion of non-randomised studies, in spite of downweighting and sensitivity analysis.

Example 8.6 *Maple: estimating complex functions of parameters*

Reference: This example forms Chapter 27 of Eddy *et al.* (1992).

Intervention: Neonatal screening for maple syrup urine disease (MSUD), an inborn error in amino acid metabolism, the early detection of which should lead to reduced rates of retardation.

Aim of study: To estimate the probability of retardation without screening, and the change in retardation rate associated with screening. The latter is denoted $e_d = \theta_n - \theta_s$, where θ_n is the retardation rate in those not screened, and θ_s is the rate in those screened.

Study design: Modelling exercise using results from multiple epidemiological cohort studies.

Outcome measure: Expected retardations.

Statistical model: The data described above are all assumed to arise from binomial distributions with the appropriate parameters. The functional relationships shown in Table 8.8 then exist.

The graphical model is shown in Figure 8.7, using the graphical tool for WinBUGS.

Table 8.8 Model and notation for maple syrup urine disease example.

Factor	Notation	Derivation
Probability of MSUD	r	
Prob. of early detection with screening	ϕ_s	
Prob. of early detection without screening	ϕ_n	
Prob. of retardation with early detection	θ_{em}	
Prob. of retardation without early detection	θ_{lm}	
Prob. of retardation for a case of MSUD who is screened	θ_{sm}	$\phi_s\theta_{em} + (1 - \phi_s)\theta_{lm}$
Prob. of retardation for a case of MSUD who is *not* screened	θ_{nm}	$\phi_n\theta_{em} + (1 - \phi_n)\theta_{lm}$
Expected retardations per 100 000 newborns who are screened	$100\,000\theta_s$	$\theta_{sm}\ r$
Expected retardations per 100 000 newborns who are *not* screened	$100\,000\theta_n$	$\theta_{nm}\ r$
Change in retardations due to screening 100 000 newborns	e_d	$\theta_s - \theta_n$

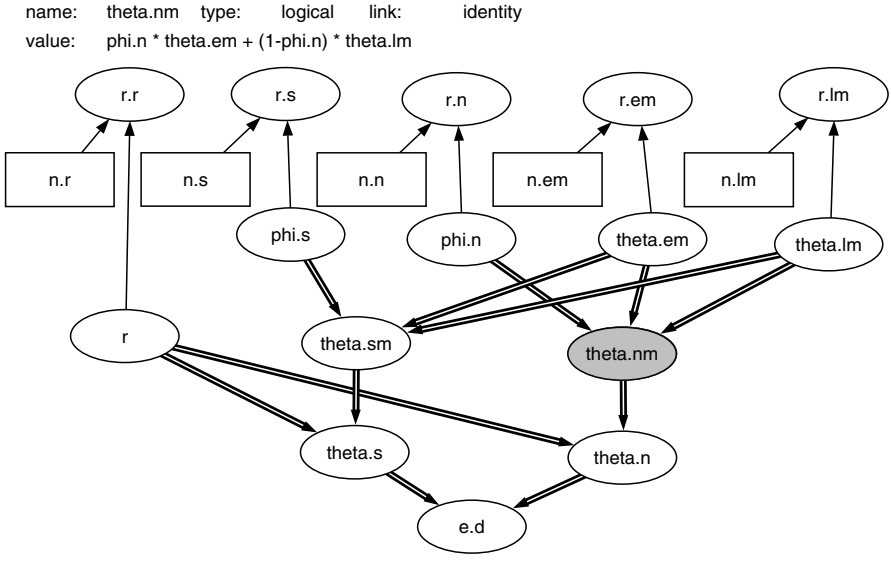

Figure 8.7 A graphical model underlying the maple syrup urine disease example. The observed data at the top of the graph depend on denominators and unknown proportions. The quantities of interest are functions of those proportions, where a double arrow corresponds to a deterministic function. This illustration is taken from WinBUGS, and shows the logical definition of node θ_{nm}, the probability of retardation for a case of a MSUD patient who is not screened.

Prospective analysis?: No.

Prior distribution: The prior distributions for all the binomial parameters used by Eddy *et al*. are the 'non-informative' Jeffreys priors, *i.e*. Beta[0.5, 0.5] (Section 5.5.1).

Loss function or demands: None.

Computation/software: MCMC analysis using WinBUGS; 100 000 iterations were carried out.

Evidence from study: There was no direct evidence on the change in retardation rate in screened and unscreened populations. The data shown in Table 8.9 were used, as provided by Eddy *et al*. (1992).

Bayesian interpretation: The posterior distribution of e_d had the properties shown in Table 8.10. Eddy *et al*. display a normal approximation to the posterior distribution for e_d, with an estimate of -0.35 (95% interval from -0.69 to -0.19). Our wider interval accurately reflects the skewed posterior distribution.

Comments: This example illustrates the synthesis of evidence from multiple studies, with appropriate allowance for the uncertainty of the parameter estimates. Further extensions could include allowance for various biases and uncertainty on the inputs to the model.

Table 8.9 Data used in maple syrup urine disease example.

Factor	Notation	Outcomes	Observations
Probability of MSUD	r	7	724 262
Prob. early detection with screening	ϕ_s	253	276
Prob. early detection without screening	ϕ_n	8	18
Prob. retardation with early detection	θ_{em}	2	10
Prob. retardation without early detection	θ_{lm}	10	10

Table 8.10 Results for maple syrup urine disease example.

Parameter	Notation	Posterior mean	95% credible interval
Expected retardations per 100 000 newborns who are *not* screened	θ_n	0.65	(0.25, 1.27)
Change in expected retardations due to screening 100 000 newborns	e_d	-0.35	(-0.77, -0.11)

Example 8.7 *HIV: synthesising evidence from multiple sources and identifying discordant information*

Reference: Ades and Cliffe (2002).

Intervention: Alternative strategies for screening for HIV in pre-natal clinics: *universal* screening of all women, or *targeted* screening of current intravenous drug users (IDUs) or women born in sub-Saharan Africa (SSA).

Aim of study: To determine the optimal policy, taking into account the costs and benefits. However, Ades and Cliffe (2002) point out that the formulation is not wholly realistic as the decision to screen universally throughout England has now been taken, and in any case a strategy of targeted testing may not be politically acceptable.

Study design: Synthesis of multiple sources of evidence to estimate parameters of the epidemiological model shown in Figure 8.8. The relevant fundamental parameters are described in Table 8.11. However, direct evidence is only available for a limited number of these parameters.

Outcome measure: SSA and IDU women will be screened under both universal and targeted strategies, and hence the only difference between the strategies comprises the additional tests and additional cases detected

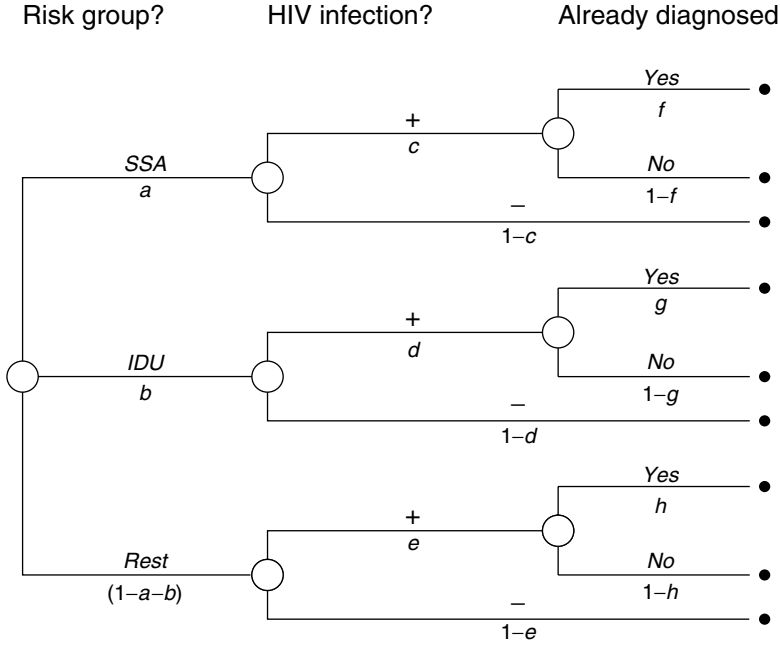

Figure 8.8 Probability tree showing how the proportions of women in different risk groups can be constructed.

Table 8.11 Definition of fundamental parameters in HIV model.

Label	Parameter
a	Proportion of women born in sub-Saharan Africa
b	Proportion of women who are intravenous drug users
c	HIV infection rate in SSA
d	HIV infection rate in IDUs
e	HIV infection rate in non-SSA, non-IDUs
f	Proportion HIV already diagnosed in SSA
g	Proportion HIV already diagnosed in IDUs
h	Proportion HIV already diagnosed in non-SSA, non-IDUs

in the non-SSA, non-IDU group. Additional tests per 10 000 women comprise those on non-SSA, non-IDU women who are not already diagnosed, and so the rate is given by $10\,000(1-a-b)(1-eh)$. The rate of new HIV cases detected is $10\,000(1-a-b)e(1-h)$.

Statistical model and evidence from study: Table 8.12 summarises the data sources available – full details and references are provided by Ades and Cliffe (2002) who also describe their efforts to select sources which are as 'independent' as possible.

Table 8.12 Available data from relevant studies, generally only allowing direct estimation of functions of fundamental parameters of interest.

Data items and sources	Parameter being estimated	Data
1 Proportion born in SSA, 1999	a	11 044 / 104 577
2 Proportion IDU last 5 years	b	12 / 882
3 HIV prevalence, women born in SSA, 1997–8	c	252 / 15428
4 HIV prevalence in female IDUs, 1997–9	d	10 / 473
5 HIV prevalence, women not born in SSA, 1997–8	$\dfrac{db+e(1-a-b)}{1-a}$	74 / 136 139
6 Overall HIV seroprevalence in pregnant women, 1999	$ca+db+e(1-a-b)$	254 / 102 287
7 Diagnosed HIV in SSA women as a proportion of all diagnosed HIV, 1999	$\dfrac{fca}{fca+gdb+he(1-a-b)}$	43 / 60
8 Diagnosed HIV in IDUs as a proportion of non-SSA diagnosed HIV, 1999	$\dfrac{gdb}{gdb+he(1-a-b)}$	4 / 17
9 Overall proportion HIV diagnosed	$\dfrac{fca+gdb+he(1-a-b)}{ca+db+e(1-a-b)}$	87 / 254
10 Proportion of infected IDUs diagnosed, 1999	g	12 / 15
11 Prop of serotype B in infected women from SSA, 1997–8	w	14 / 118
12 Prop of serotype B in infected women not from SSA, 1997–8	$\dfrac{db+we(1-a-b)}{db+e(1-a-b)}$	5 / 31

The crucial aspect is that there is no direct evidence concerning the vital parameters e and h for the low-risk group, and hence their value must be inferred indirectly from other studies. For this reason the parameter w is introduced which is not part of the epidemiological model: the assumption that the low-risk group has the same prevalence of subtype B as SSA women, and that all IDU women are subtype B, allows use of data source 12 on non-SSA women.

Prior distribution: Uniform priors for all proportions are adopted.

Computation/software: MCMC methods implemented using WinBUGS.

Bayesian interpretation: The posterior estimates and intervals for the proportions underlying the studies are given in Table 8.13, together with the quantities of interest.

Sensitivity analyses: Here we focus on the consistency of data sources rather than the usual analysis of sensitivity to model assumptions. We have synthesised all available data, but the results may be misleading if we have included data that do not fit our assumed model. A simple way of assessing possible conflict is to compare the observed proportion in the 12 sources with that fitted by the model, and it is apparent that the observation for source 4 is only just included in the 95% interval, while the data for source 12 lie wholly outside its estimated interval. This is only a crude method, since a source may strongly influence its estimate, so a better procedure is to leave each source out in turn, re-estimate the model, and then predict the data we would expect in a source of that

Table 8.13 Estimates of parameters underlying the available data. Estimates of quantities of interest in selecting a screening strategy are also shown.

	Quantity	Observed proportion	Estimate	95% interval	*P*-value (excl 4)
1	Proportion SSA	0.106	0.106	0.104 to 0.108	0.47
2	Proportion IDUs	0.0137	0.0088	0.0047 to 0.149	0.46
3	HIV prevalence in SSA	0.0163	0.0172	0.0155 to 0.0189	0.27
4	HIV prevalence in IDUs	0.0211	0.0120	0.0062 to 0.0219	0.004
5	HIV prevalence non-SSA	0.000544	0.000594	0.000478 to 0.000729	0.35
6	Overall HIV prevalence	0.00248	0.00235	0.00217 to 0.00254	0.21
7	SSA as proportion of all diagnoses	0.717	0.691	0.580 to 0.788	0.50
8	IDU as proportion of non-SSA diagnoses	0.235	0.298	0.167 to 0.473	0.40
9	Proportion HIV diagnosed	0.343	0.350	0.296 to 0.408	0.47
10	Proportion IDU already diagnosed	0.800	0.747	0.517 to 0.913	0.44
11	Prop subtype B in SSA	0.119	0.111	0.065 to 0.171	0.43
12	Prop subtype B in non-SSA, 1997–8	0.161	0.285	0.201 to 0.392	0.23
	Additional tests per 10 000, $10\,000(1-a-b)(1-eh)$		8856	8789 to 8898	
	Additional HIV cases detected, $10\,000(1-a-b)e(1-h)$		2.49	1.09 to 3.87	

size. This predictive distribution, easily obtained using MCMC methods, is then compared to the observed data and a *P*-value calculated in a parallel manner to Box's test of prior/data compatibility described in Section 5.8 (although here we seek to criticise the data rather than the 'prior' based on the remaining studies). We may term these 'cross-validatory *P*-values'.

Removing data source 4 from the analyis leads to the cross-validatory *P*-values shown in Table 8.13. The small *P*-value for source 4 shows its lack of consistency with the remaining data, whereas the predictions for the remaining data seem quite reasonable. Removing source 4 from the analysis leads to an estimate of 8810 (8717 to 8872) for additional tests per 10 000, and 2.73 (1.31 to 4.12) for additional HIV cases detected, so the removal of this divergent source does not in fact have much influence on the conclusions. The estimates for the fundamental parameters are presented in Table 8.14.

Comments: Example 9.5 extends this example to include cost-effectiveness analysis.

Table 8.14 Estimates of fundamental parameters in HIV model, ignoring evidence from source 4.

Label	Parameter	Median	95% interval
a	Proportion of women born in SSA	0.106	0.104 to 0.108
b	Proportion of women who are IDUs	0.013	0.007 to 0.022
c	HIV infection rate in SSA	0.0172	0.0156 to 0.0189
d	HIV infection rate in IDUs	0.0046	0.0015 to 0.012
e	HIV infection rate in non-SSA, non-IDUs	0.00051	0.00039 to 0.00065
f	Proportion HIV already diagnosed in SSA	0.32	0.24 to 0.40
g	Proportion HIV already diagnosed in IDUs	0.78	0.55 to 0.93
h	Proportion HIV already diagnosed in non-SSA, non-IDUs	0.40	0.22 to 0.67

8.5 FURTHER READING

Sutton *et al.* (2000) review the whole area of meta-analysis and Bayesian methods in particular: other reviews are provided by Jones (1995), Normand (1999) and Hedges (1998). See also the book edited by Stangl and Berry (2000).

Empirical Bayes approaches for meta-analysis have received most attention in the literature until recently, largely because of computational difficulties in the use of fully Bayesian modelling (Raudenbush and Bryk, 1985; Stijnen and van Houwelingen, 1990). However, the full Bayesian hierarchical model has been investigated extensively by DuMouchel and Harris (1983), DuMouchel (1990),

DuMouchel and Waternaux (1992) and Abrams and Sansó (1998) using analytic approximations, and also using MCMC methods (Morris and Normand, 1992; Smith *et al.*, 1995). Carlin (1992), for example, considers meta-analyses of both clinical trials and case–control studies; he examines the sensitivity to choice of reference priors, and explores checking the assumption of normal random effects. There have been many comparative studies of the full Bayesian approach, including trials (Rogatko, 1992; Su and Po, 1996; Tunis *et al.*, 1997) and observational studies (Biggerstaff *et al.*, 1994; Su and Po, 1996; Tweedie *et al.*, 1996).

Tutorial articles on the confidence profile method include Eddy (1989), Eddy *et al.* (1990a, 1990b) and Shachter *et al.* (1990). The method has been used in meta-analysis of the benefits of antibiotic therapy (Baraff *et al.*, 1993), mammography in women aged under 50 (Eddy *et al.*, 1988) and angioplasty (Adar *et al.*, 1989).

8.6 KEY POINTS

1. A unified Bayesian approach appears to be applicable to a wide range of problems concerned with evidence synthesis.
2. The Bayesian approach provides a natural structure for many subtle issues that arise in meta-analyses, such as adjusting for baseline risk.
3. Priors on nuisance parameters can be important when there is limited evidence, such as when there are rare events or few studies.
4. 'Indirect' comparisons enable one to infer comparisons where there is limited or no head-to-head evidence.
5. Generalised evidence synthesis is likely to become increasingly important as evidence from disparate studies is used in the construction of health-policy models.
6. Complex synthesis models make extensive use of assumptions, only some of which can be empirically checked, and careful sensitivity analysis is vital.

EXERCISES

8.1. Repeat the analysis in Example 3.13 but using a full Bayesian analysis as in Section 8.2, using WinBUGS. Given the relatively small number of studies, it is important to consider the sensitivity of the posterior results to the prior distribution for the between-study variability (Section 5.7.3): explore the options illustrated in Example 8.1.

8.2. Table 8.15 is adapted from Berry (2000) and presents the results of six RCTs which evaluated cholesterol reduction compared to control in terms of coronary deaths in patients who had previously suffered a myocardial infarction.

Table 8.15 RCTs evaluating cholesterol reduction compared to control in terms of coronary deaths in patients who had previously suffered a myocardial infarction.

	Intervention		Control	
Study	Deaths	Total	Deaths	Total
CDP	398	2224	535	2789
Newcastle	25	244	44	253
Edinburgh	34	350	35	367
Stockholm	47	279	73	276
Oslo	37	206	50	206
MRC	35	322	37	323

(a) Obtain and compare the posterior distribution for the overall pooled odds ratio using a random-effects meta-analysis based on: (i) a normal approximation to the likelihood arising from the observed log(odds ratio) and standard error in each RCT; (ii) modelling the events in the two arms of each RCT using binomial distributions.

(b) In each case assess the sensitivity of the results to the prior distribution assumed for the between-study variability, as in Example 8.1.

(c) An additional large-scale RCT (4S) was reported after those in Table 8.15, in which 111 deaths occurred out of 2221 patients in the intervention arm, and 189 deaths occurred out of 2223 patients in the control arm. The observed effect in the 4S trial was considered to be in conflict with that of those in Table 8.15. Obtain the predictive distribution based on the six RCTs in Table 8.15 for a future RCT and therefore assess whether the assertion that there was a conflict was in fact warranted, and in particular whether the sensitivity analyses considered in (a) affect this assessment.

8.3. Geddes *et al.* (2000) consider a meta-analysis of 23 RCTs which compared the use of atypical anti-psychotic drugs with haloperidol in patients with schizophrenia. The summary data are shown in Table 8.16 with the relevant dose. Evaluate whether there is evidence for an effect of dose on treatment effect.

8.4. Using the techniques described in Section 8.2.3, investigate the extent to which the effect of diuretic therapy on risk of pre-eclampsia considered in Exercise 3.12 depends upon the baseline level of risk.

8.5. In Example 8.2 a meta-analysis of nine RCTs evaluating the effect of electronic foetal heart rate monitoring on perinatal mortality was presented. In addition to the nine RCTs, Sutton and Abrams (2001) also considered evidence from the seven non-randomised comparative studies and ten before–after studies which are presented in Table 8.17 together with the results for the RCTs. Explore the effect that consideration of both randomised and non-randomised evidence has on the conclusions obtained in Example 8.2 when: (a) the non-randomised evidence is

Table 8.16 Standardised effect sizes and associated standard errors (SE) for 23 RCTs evaluating comparing atypical anti-psychotic drugs with haloperidol in patients with schizophrenia.

Study	Standardised effect size	SE	Dose
1	−0.014	0.158	12.0
2	−0.070	0.150	15.0
3	−0.191	0.136	15.0
4	−0.663	0.312	8.0
5	−0.488	0.320	20.0
6	+0.455	0.254	11.0
7	−0.273	0.250	20.0
8	+0.129	0.309	6.0
9	−0.109	0.142	10.0
10	−0.779	0.330	22.5
11	−0.765	0.225	7.6
12	−0.214	0.214	7.5
13	−0.775	0.437	13.5
14	+0.216	0.116	16.0
15	+0.018	0.105	10.0
16	−0.406	0.145	20.0
17	−0.234	0.146	17.5
18	−0.112	0.075	10.0
19	−0.294	0.147	16.0
20	−0.469	0.131	17.5
21	−0.903	0.365	20.0
22	−0.237	0.048	12.5
23	+0.049	0.099	9.4

considered as *prior* evidence, either at 'face value' or downweighted; and (b) when both the randomised and non-randomised sources of evidence are considered within a single hierarchical model following the methods of Section 8.4 and Example 8.5. You will need to make some explicit prior assumptions about the size of the potential bias of the non-randomised studies, and conduct suitable sensitivity analysis.

8.6. In addition to the 17 single-arm studies evaluating either radiotherapy alone (RTx) or radiotherapy together with adjuvant chemotherapy (RTx+Chm) following surgery for childhood medulloblastoma reported in Table 5.7, Sutton *et al.* (2000) also considered six RCTs comparing the two interventions and summarised in Table 8.18. Using the prior distribution for the difference in 5-year survival rates between the two therapies in Exercise 5.6, together with the RCT evidence in Table 8.18, obtain a posterior distribution for the difference: (a) using the evidence from the single-arm studies at 'face value'; (b) possibly downweighting the uncontrolled evidence or allowing for bias; (c) modelling both the randomised and non-randomised sources of evidence within a single model following the methods of Section 8.4.

Table 8.17 RCTs, non-randomised comparative studies and before–after studies evaluating electronic foetal heart rate monitoring (EFM) in terms of perinatal mortality.

Study	Year of publication	EFM		Control	
		Deaths	Total	Deaths	Total
RCTs					
1	1976	1	175	1	175
2	1976	2	242	1	241
3	1978	0	253	1	251
4	1979	3	463	0	232
5	1981	1	445	0	482
6	1985	0	485	1	493
7	1985	14	6 530	14	6 554
8	1987	17	122	18	124
9	1993	2	746	9	682
Non-randomised					
1	1973	2	1 162	17	5 427
2	1973	0	150	15	6 836
3	1975	1	608	37	6 179
4	1977	1	4 210	9	2 923
5	1978	1	554	3	692
6	1979	0	4 978	2	8 634
7	1982	10	45 880	45	66 208
Before–after					
1	1975	4	991	0	1 024
2	1975	7	1 161	9	1 080
3	1975	14	11 599	1	1 950
4	1976	15	4 323	1	3 529
5	1977	53	4 114	21	3 852
6	1978	35	15 357	6	7 312
7	1980	19	4 240	2	4 503
8	1980	15	6 740	5	8 174
9	1984	13	7 582	2	7 911
10	1986	7	17 409	5	17 586

Table 8.18 Five-year survival rates and standard errors for RCTs comparing radiotherapy alone (RTx) with radiotherapy together with adjuvant chemotherapy (RTx+Chm) following surgery for childhood medulloblastoma.

Study	RTx+Chm		RTx	
	S_5	SE(S_5)	S_5	SE(S_5)
1	0.55	0.026	0.42	0.020
2	0.58	0.058	0.60	0.054
3	0.74	0.083	0.56	0.099
4	0.59	0.060	0.50	0.065
5	0.17	0.217	0.63	0.341
6	0.46	0.114	0.30	0.118

8.7. In Example 8.7, suppose an additional trial came to light which showed an HIV prevalence of 10/10 000 in non-SSA, non-IDU women.
   (a) Does this study conflict with the available evidence?
   (b) How would its inclusion alter the findings?