META-ANALYSIS METHODS FOR COMBINING INFORMATION FROM DIFFERENT SOURCES IN EVALUATING HEALTH INTERVENTIONS

Thesis submitted for the degree of

Doctor of Philosophy

At the University of Leicester

By

Alexander Julian Sutton B.Sc., M.Sc.

Department of Epidemiology and Public Health

University of Leicester

January 2002

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

META-ANALYSIS METHODS FOR COMBINING INFORMATION FROM DIFFERENT SOURCES IN EVALUATING HEALTH INTERVENTIONS

Alexander Julian Sutton B.Sc., M.Sc.

Abstract

This thesis considers the quantitative synthesis of evidence from different study types in order to assess the effectiveness of health interventions. Bayesian MCMC methodology is used extensively, but not exclusively, for the analyses described herein. The thesis commences with consideration of different study designs used in health and related disciplines together with consideration of the validity of these sources. Existing synthesis methods for combining information, first, from a single study design (often referred to as meta-analysis), and then from multiple sources of evidence are then reviewed.

A meta-analysis of the randomised evidence on cholesterol lowering observations is presented. This analysis is then extended to a more generalised synthesis by including data from aetiological cohort studies in the analysis using hierarchical modelling methods. Such models allow for heterogeneity between study types. A second generalised synthesis considers evidence from three sources relating to the use of electronic fetal heart rate monitoring during labour. The particular problem of publication bias, and how it can be addressed in a generalised synthesis framework, where there are potentially differential levels of publication bias for the different sources of evidence, is discussed. Adverse events from interventions are often rare, and hence, difficult to detect and quantify using randomised controlled trials. The use of generalised synthesis to quantify adverse events is illustrated using data relating to adverse events of hormone replacement therapy and breast implants. The sparseness of the event data in these examples presents specific statistical problems which are explored. A sensitivity analysis framework for assessing the robustness of results to under-reported adverse events is outlined. A final example, the use of warfarin to prevent strokes in patients with atrial fibrillation, illustrates how disparate sources of data can be synthesised to construct a net-clinical-benefit model where potential benefits of treatment are weighed up against potential harm due to adverse events. This analysis synthesises clinical event data from randomised controlled trials, observational cohort studies for both benefit and harms as well as quality of life data. The net-clinical-benefit of the treatment is expressed, together with corresponding uncertainty measures, for patients with different underlying risks.

This thesis illustrates that with the increase in computer power and development of software to fit complex models using Bayesian MCMC methodology, it is now possible to think beyond the models currently used to synthesise medical data. It is hoped that such efforts will be seen as tentative first steps in a future where quantitative models are created routinely to summarise the totality of evidence, and inform models to make decisions for future patients.

Acknowledgements

I would like to thank my supervisor David Jones for his help, wisdom, insight, patience and encouragement throughout the duration of my studies. I would like to thank other members of the medical statistics group in the Department of Epidemiology and Public Health, particularly Keith Abrams and Paul Lambert for helpful and thought provoking discussion of my work. I am also pleased to thank Paul Glaziou, William DuMouchel, Teresa Prevost, Russel Wolfinger for taking the time to discuss their work (on which much of this thesis builds) with me; Jo Leonardi-Bee for updating the cholesterol lowering dataset; Andy Vail and Janet Hornbuckle for allowing me to use the EFM meta-analysis dataset, and Jenny Kurinczuk for useful discussions regarding perinatal mortality generally. Finally, thanks and love go to Nicola Cooper for her help, support, patience, and not gloating too badly when beating me to the finishing line.

Contents

Chapter 1 Introduction

Chapter 2 Critical review of current methods for meta-analysis

Chapter 3 Critical review of methods for synthesising disparate sources of evidence

Chapter 4 Combining randomised and observational studies: the example of cholesterol lowering and risk of coronary heart disease and all cause mortality

Chapter 5 Further generalised synthesis methods including further analysis of the cholesterol lowering evidence

Chapter 6 Generalised synthesis of evidence and the threat of dissemination bias: electronic fetal heart rate monitoring (EFM)

Chapter 7 Synthesis of studies reporting rare outcomes incorporating the case studies of adverse effect of hormone replacement therapy and breast implants

Chapter 8 Meta-analysis of composite measures of benefit and harm: a Bayesian exposition of the Net Benefit model

Chapter 9 Discussion, including lines of further work

Chapter 1 Introduction

1.1 Background

Over the last fifty years of the twentieth century there has been an ever-accelerating increase in the volume of medical and related literature published. The lack of systematic summaries of one (very important) part of this literature, namely randomised clinical trials of medical interventions, lead Archie Cochrane to comment in 1979 that:

"It is surely a great criticism of our profession that we have not organised a critical summary, by speciality or sub-speciality, adapted periodically, of all relevant randomised controlled trials" (Cochrane, 1979)

In subsequent years, this has changed. Since the early 1980s, when the first systematic reviews including quantitative summaries, or meta-analyses, were published, the number of systematic review/meta-analyses papers has also increased at an accelerating rate. In addition, the last decade has witnessed the inception of the Cochrane Collaboration, (Chalmers et al. 1997) a worldwide group whose principal aim is to provide accessible, thorough, up to date reviews of as many interventions as their largely voluntary manpower permits. These developments are crucial aspects of the Evidence Based Medicine drive, (Sackett et al. 1996) which now underpins much clinical practice and training throughout the world.

Just as the number of meta-analyses carried out has exploded, so too has the methodological literature concerned with the methods used to undertake such syntheses. A recent (systematic) review of this literature (Sutton et al. 1998) identified well over 500 references providing novel information on some aspect of systematic review methodology. Hence, although it is only 24 years ago that the term 'meta-analysis' was first used in this context, (Glass, 1976) much has been written in the intervening years, and many of the methods have attained some degree of refinement.

The main bulk of this PhD has been written in a climate during which the use of meta-analysis has become routine. In these five years (1996-2001) the output of the Cochrane Collaboration has expanded from a single floppy disk to multiple CD ROMs; specific

meta-analysis software has been written making the majority of statistical analyses routine; (Sterne et al. 2000; Sutton et al. 1998) and several textbooks on the subject have appeared. (Sutton et al. 2000a; Egger et al. 2000) Indeed it would seem that the world is finally catching up and taking stock of the vast array of randomised evidence as Archie Cochrane hoped.

However, it should not be forgotten that the RCT reports are only a fraction of the total medical and related literature. Indeed, for example, aetiologically orientated meta-analyses of observational studies are being produced in nearly the same quantities as those appearing on interventions combining RCTs.

A third distinct fraction of the total medical literature comprises the reports of observational studies assessing interventions. (Stroup et al. 2000) Over the years such studies have received less attention from meta-analysists, even though with the increase in the use of electronic databases and record linkage the routine audit of healthcare has been greatly facilitated. Although, guidelines for their reporting have now been published. (Stroup et al. 2000) There is a simple and good reason why such studies are often ignored by meta-analysts when assessing interventions; this is the fear of introducing bias into the analysis. Potential design weaknesses in observational studies, allowing possible confounding, and hence bias, raises concerns that such studies can give the wrong answers and therefore are not to be relied upon.

There is growing wisdom however that such a "closed door" approach to observational evidence is not always optimal. After completing several years work on this thesis it is heartening to observe a growing interest in the potential value of observational studies of interventions. This is highlighted by the registration in 1999 of a Cochrane Collaboration methods group to develop guidelines and methods for the inclusion of non-randomised evidence into Cochrane reviews (Oxman, 2001) (see also http://www.cochrane.dk/nrsmg/).

1.2 Non-randomised study designs

The various non-randomised study designs in common use in healthcare which may provide evidence, in addition to that from randomised studies, in a meta-analysis of effectiveness are described below.

Non-randomised controlled trial

The distinction between these studies and RCTs is that in these the treatment allocation is not randomised. Patients may have been recruited in a similar manner to an RCT but either for reasons of ethics, feasibility, or simply due to poor study design the treatment allocation was not randomised. For example, some trials allocate patients to treatments on an alternating basis; such studies are sometimes called pseudo-randomised trials.

Historically-controlled trial

In historically controlled trials, either the treatment, but more often the control group(s), may have been created retrospectively by searching through existing patient records. These sorts of studies produce a comparative treatment effect, much like an RCT. However, the lack of a concurrent treatment allocation mechanism may introduce bias through lack of comparability between the groups being compared.

Cohort designs

Cohort studies are very common in epidemiology. (Farmer & Miller 1996) They can be either prospective, or retrospective. Subjects are followed up over a period of time and comparisons of interest, such as mortality rates, can be compared either between subgroups of the whole cohort or other cohorts or data sources. By comparing groups defined by intervention received, treatment comparisons can be made.

Case-control studies

Case-control studies are another very common study design in epidemiology. (Farmer & Miller 1996) Diseased individuals (the cases) are compared with a sample of non-diseased subjects (the controls). Cases and controls can be matched, on known confounding factors, often gender or age. Although not usually a natural design for assessing effectiveness, they have been used with success in several areas, including the assessment of the effectiveness of bicycle safety-helmets. (Thompson et al. 1999)

Before and after studies/Interrupted time series

The name of these studies is largely self-explanatory: outcomes are monitored, often in a given hospital or clinic before and after a new intervention is implemented. The effectiveness of the new intervention is assessed by comparing performance before and after it was introduced. Interrupted time series designs are conceptually very similar; but here the same subjects are assessed before and after the new intervention is introduced.

Case series/ n of 1 studies

The response of individual patients to treatment may be published. This occurs frequently for relatively rare diseases. Although less common in medical studies than the social sciences, studies assessing multiple alternative therapies sequentially on individual patients (*n* of 1 studies) are still used in some areas, such as speech therapy (Enderby & Emerson, 1995), or pain management. These studies are distinct from crossover trials since here patient treatment allocation order is not randomised.

Routine database/audit data

Database/audit data is becoming increasingly accessible in principle, following improvements in data storage and retrieval methods. Hlatky (Hlatky, 1991) notes that descriptive studies and analyses of prognostic factors are established research uses of databases, but using them to compare therapies remains controversial, due to the

potential for bias. Indeed it has been suggested that the problems with databases are even bigger than those for historical control studies. (Byar, 1991) However, large administrative databases are available that could be used to compare therapies in large populations of patients, and some feel if the data are of high enough quality that their use to compare therapies is justified. (Hlatky, 1991)

Other study types

Other study types including surveys, animal studies and earlier phase clinical trials may also provide valuable information regarding the effectiveness of an intervention; the list above is not exhaustive, but fully demonstrates the range of potential non-randomised data relevant to the assessment of interventions.

1.3 The debate over the value of observational evidence in assessing effectiveness

There clearly is not unanimous support for raising the profile of observational studies of interventions, and their use for evaluation purposes remains a contentious issue. As mentioned above, there is concern that biases may invalidate observational studies results. Allocation bias may be particularly serious since patients are not randomly allocated, but chosen for a particular intervention. If the allocation indication is influenced by either practitioner or patient preferences then confounding by indication may exist. (Psaty et al. 1999) If differences are known about they can be adjusted for at the analysis stage, (Psaty et al. 1999) using methods used to adjust for other forms of confounding in observational studies. However it is impossible to be certain all such differences have been identified.

There are instances where carrying out a randomised controlled trial would be: (Black, 1996)

a) unnecessary, when the treatment effect is so large that unknown confounding factors can be ignored;

b) inappropriate, for assessing or preventing rare events; investigating outcomes of interest far into the future, or interventions of which the effectiveness is dependent on the subject's beliefs and preferences. (The analysis of rare events is given considerable attention in this thesis in Chapter 7.)

c) impossible, due to the reluctance of clinicians to participate, ethical objections, political or legal blocking or administrative impracticability problems and the complexity of the intervention.(MRC Health Services and Public Health Research Board, 2001)

d) inadequate due to a lack of generalisability of the trial results. Even when efficacy of an intervention has been proven in trials, this has not always translated into real gains in effectiveness when such interventions were put into routine practice. In such instances the artificial, experimental, nature of a trial has been blamed, possibly due to restrictive patient inclusion criteria, and artificially high levels of care, for not giving a realistic idea of what can be achieved in routine practice. (Sutton and Abrams, 1998) (An example of where observational evidence is used to extrapolate results to patients with characteristics not necessarily present in the RCTs is detailed in Chapter 8.)

There are further instances where timeliness of RCT evidence is problematical: policy decisions are required, but the randomised evidence currently available is not adequate to give a conclusive answer. Areas where observational evidence may be particularly valuable (as identified by the Cochrane Non-randomised studies working group (unpublished)) include public health promotion interventions, assessments of organisational change, and surgery.

In conclusion, Black (Black, 1996) suggests that the complementary roles of randomised and non-randomised studies should be acknowledged in the quest for scientific rigour in evaluation. This is the view held by the author. However, the *how to* combine the complementary strengths of randomised and observational evidence is far from obvious. It is this serious deficiency in the literature regarding methods for quantitatively combining the results from experimental and non-experimental designs (see Chapter 3) that provided the motivation for undertaking this thesis.

1.4 The validity of findings from observational studies and non-randomised trials assessing medical interventions

Recently, the validity of the results of observational studies of interventions has been assessed in several investigations. (Concato et al. 2000; Benson and Hartz, 2000; Briton et al. 1998; Reeves et al. 1998) These empirical investigations have compared the results obtained from observational studies with those from RCTs on the same topic and examined how closely they agree. All four of these investigations reported that, contrary to the belief that non-randomised studies give larger estimates of treatment effect, no obvious trends emerged. In some instances there was close agreement in the estimates produced by the two types of evidence. (Benson and Hartz, 2000) However, Kunz and Oxman (Kunz and Oxman, 1998) compared the results of randomised and non-randomised controlled trials of the same interventions, but came to different conclusions. In many instances the discrepancy between results was large, and on average the non-randomised trials gave a larger treatment effect, sometimes 150% or more larger than the randomised trials. The overestimation was neither consistent nor predictable and relative decreases of 90% were also observed for the non-randomised studies. This led the authors to conclude that failure to use adequately concealed random allocation can cause distortions as large or larger than the size of the effects that are to be detected. Clearly, this is an area where further empirical research is required before a generally interpretable pattern emerges.

1.5 Hierarchies of evidence

In their guidelines for carrying out systematic reviews, Deeks et al. (Deeks et al. 1996) produced an example hierarchy of evidence, which is reproduced below (Figure 1.1). This grades study types according to the reliability of their results.

Although this is a useful starting point for considering the relative merits of evidence from different sources, it is of limited use as study validity not only depends on the type of study, but also how well it was designed, carried out and analysed. Indeed, a poor RCT may be less reliable than a well-conducted observational study. (Deeks et al. 1996) Further, a consistent assessment of quality across different study designs is very

difficult, despite generic scoring criteria having been developed. (Downs and Black, 1998)

The issue of study quality and the magnitude and direction of biases to which observational studies are susceptible to are very important issues when synthesising such information. This is not the main theme of this thesis but is discussed further in Chapter 9.

Figure 1.1 An Example of a Hierarchy of Evidence (*reproduced from Deeks et al.* (Deeks et al. 1996))

I	Well-designed randomised controlled trials
	Other types of trial:
II-1a	Well-designed controlled trial with pseudo-randomisation
II-1b	Well-designed controlled trials with no randomisation
	Cohort studies:
II-2a	Well-designed cohort (prospective study) with concurrent controls
II-2b	Well-designed cohort (prospective study) with historical controls
II-2c	Well-designed cohort (retrospective study) with concurrent controls
II-3	Well-designed case-control (retrospective) study
III	Large differences from comparisons between times and/or places with and without intervention. (in some circumstances these may be equivalent to level II or I)
IV	Opinions of respected authorities based on clinical experience; descriptive studies and reports of expert committees

1.6 Aims of this thesis

This thesis considers methods for synthesising data from different sources, with special emphasis on the role of non-randomised evidence in a meta-analysis of the effectiveness of a medical intervention, and how it can be combined with randomised evidence. The standpoint taken is that non-randomised studies are distinct from RCTs and hence should be treated distinctly. That is to say the examples considered, and the methods developed herein, do not simply 'lump' together and pool randomised and non-randomised studies using standard meta-analysis models.

Bayesian statistical methods have been used extensively, but not exclusively throughout this work. Their implementation in the software package WinBUGS (Speigelhalter et al. 2000a) provides a flexible environment in which, non-"off the peg" statistical models can be implemented with relative ease, as well as providing certain theoretical advantages over classical alternatives. Although the WinBUGS code is usually not included in this thesis, the models used are always explicitly described which should be sufficient to allow the methods to be replicated. When the code is less straightforward, this is given in an appendix.

In all examples the data used were that which were available from study reports, and, hence, aggregated across patients. No attempt to use individual patient data (ipd), either to carry out further analyses on individual studies, or to synthesise at the patient level was made. Consideration is given in the discussion (Chapter 9) to ways in which ipd may be used advantageously.

1.7 Outline of the Thesis

Chapter two briefly reviews the standard meta-analysis methods used for combining studies with the same designs, and highlights problematic issues in doing so.

Chapter three then critically considers the methods that have been developed previously for synthesising information from studies with different designs.

Chapter four reports a synthesis of the evidence relating to the effect on mortality and coronary events of reducing blood cholesterol levels. This topic was chosen because it

is an area of high research output where many studies of different designs are available. The analysis was originally carried out in 1998, and the data were up-to-date at the time, but much of the analysis was updated in 2001, using data available up-to-date as of the end of 2000. This analysis provided the opportunity to implement the current 'state of the art' methods for synthesising evidence, and forms the base for which much of the work described in future chapters develops.

Following a detailed analysis of the randomised evidence in Chapter four, Chapter five considers the non-randomised evidence from aetiological studies on the relationship between blood cholesterol levels and mortality. Bayesian three-level hierarchical models and mixed effect regression models (previously only implemented using frequentist methods) are implemented which illustrate contrasting approaches to how the evidence can be combined. In addition, a multivariate model to combine multiple continuous outcomes simultaneously is described and applied to a meta-analysis of dental interventions. Previously only frequentist formulations of such a model had been described; here advantages of the Bayesian approach are discussed. A random effects model for estimating indirect comparisons is also described and applied to a meta-analysis of treatments for the prevention of pneumonia in HIV infection. Previously only fixed effects models had been used to combine this data, but the random effect model developed here would appear more appropriate.

Chapter six considers the issue of publication bias when combining studies from different sources, and specifically considers the issue of different publication bias mechanisms being important for different study designs. An assessment of publication bias using data on the use of electronic fetal heart rate monitoring is described. The use of fetal monitoring is still a contentious issue with no clear benefits on overall mortality being demonstrated in trials, but benefits are apparent when the non-randomised literature is considered. Much of the data in this example are sparse; the problems associated with synthesising such evidence are considered further in Chapter seven. The analysis considered here raises some interesting issues regarding how to deal with publication bias in a generalised synthesis framework. This includes the use of the Trim and Fill method in combination with the methods for combining data described in Chapter five providing a way of 'adjusting' for publication bias in a generalised

synthesis framework. Limitations of the Trim and Fill method are highlighted and two promising approaches to improving this method are outlined.

Chapter seven considers meta-analysis of rare events, an area where non-randomised evidence may be particularly valuable due to the lack of data often available from RCTs. Before considering generalised synthesis models, the suitability of traditional meta-analysis models for the combination of extremely rare events is questioned. Bayesian "exact" simulation methods, bootstrapping, and data-inflation are all examined as potential alternatives. Data on the use of hormone replacement therapy and risk of breast cancer and coronary heart disease are used to illustrate these methods. The randomised evidence on this subject consists of extremely sparse data, and a previous meta-analysis of it met with a critical reaction. In order to address some of these criticisms, a novel simulation-based sensitivity analysis approach to dealing with uncertainties in the reported data is described. A less technical method than the hierarchical models of Chapter 4 is used to synthesise observational with this randomised evidence, and an intuitive diagram that allows the randomised and non-randomised evidence to be summarised and "weighed up" against each other is presented. Finally, an exact random effect method for combining rare events from observational studies with different designs is described. Previously, only a fixed effect model was developed for combining such data. This new model is then applied to a meta-analysis of the risk of connective tissue diseases from breast implants.

Chapter eight, is still directly concerned with the synthesis of evidence from studies with different designs, and in particular addresses the question of how to assess whether a particular intervention will be beneficial for an individual patient (a question first considered in Chapter 4). However it departs from the use of hierarchical models used thus far. The net-benefit model of Galziou and Irwig (Glasziou and Irwig, 1995) is revisited and implemented using Bayesian methods. This model provides a method of extrapolating a general (aggregate) measures of effect to individual patients with different characteristics that are believed to modify the treatment effect magnitude. The warfarin and atrial fibrillation example originally used to illustrate the method is re-evaluated. In this analysis RCT effectiveness data, observational data on adverse effects, quality of life data and information from multivariate risk equations are all synthesized. This advances the original work by providing a framework for including

uncertainty related to the estimation of all model parameters, and permits both direct probability statements and a credible interval for an individual's net benefit to be expressed.

Chapter nine concludes the thesis, discussing issues raised in previous chapters, and makes recommendations for areas of further work.

Bayesian methods are used in many places in this thesis; while the application of such methods to evidence synthesis and health technology assessment more generally is in its infancy, guidelines (Bayeswatch) have recently been developed to encourage comprehensive reporting when such methods are used. (Spiegelhalter et al. 1998) While such guidelines were followed for the analyses described herein, due to space limitations, it was not possible to report all the recommended information for every analysis. Hence, for the first substantial Bayesian analysis, reported in section 5.1, results are given in their entirety as recommended in Bayeswatch; for subsequent analyses such information is available on request.

Chapter 2 Critical review of current methods for meta-analysis

2.1 Introduction

In this chapter the methods for meta-analysis are reviewed. Many of these methods have been used routinely for several years, while others, such as the methods to address publication bias, are somewhat more recent developments. Due to limitations in space, consideration is not given here to the important pre-synthesis components of a systematic review, including protocol specification, literature searching and data extraction; however these have been reviewed by the author elsewhere. (Sutton et al. 1998) Similarly, since the current literature is considerable, attention is focused on the most mainstream and commonly used methods that are used and extended in this thesis. Notable omissions include methods for vote-taking and combining p-values, analysis of individual patient data and analysis of survival data. More comprehensive reviews of the meta-analysis methodology literature, which include these and other topics, can be found elsewhere. (Sutton et al. 1998; Sutton et al. 2000a)

Since Bayesian methods are used extensively in the latter chapters, this review includes the Bayesian formulations, where they exist, alongside their Classical equivalents, as well as extensions unique to the Bayesian approach.

2.2 Fixed effect models

The simplest meta-analysis methods which produce an overall pooled estimate are fixed effect models. Using a fixed effect model to combine treatment estimates assumes no heterogeneity between the study results; the studies are assumed all to be estimating a single true underlying effect size. Clearly, in many instances this may not be realistic, and hence the need for the more sophisticated methods described later.

2.2.1 General fixed effect model - the inverse variance-weighted method

The general form of the fixed effect model, the inverse variance-weighted method, was first described by Birge (Birge, 1932) and Cochran (Cochran, 1937) in the 1930s, and more recently placed in a formal meta-analysis framework by Whitehead and Whitehead (Whitehead and Whitehead, 1991). Each study estimate is given a weight directly proportional to its precision (that is inversely proportional to its estimated variance). For $i = 1, \ldots, k$ independent studies to be combined, let T_i be the observed effect size with estimated variance v_i, θ_i the true underlying effect size, and σ_i^2, for the ith study. For a fixed effect model all population effect sizes are assumed equal i.e. $\theta_1 = \ldots = \theta_k = \theta$, where θ is the true common underlying effect size. Hence, the model is given by $T_i = \theta + e_i$, where e_i are error terms and are realisations of normally distributed random variables with expected value 0 and variance σ_i^2. It follows that T_i satisfies the distributional relationship $T_i \sim N(\theta, \sigma_i^2)$ A pooled estimate of the treatment effect is given by:

$$\overline{T}. = \frac{\sum_{i=1}^{k} w_i T_i}{\sum_{i=1}^{k} w_i} . \qquad (2.1)$$

The weights that minimise the variance of $\overline{T}.$, and hence are routinely used, are inversely proportional to the variance in each study: (Shadish and Haddock, 1994)

$$w_i = \frac{1}{v_i} . \qquad (2.2)$$

The details of the variance formulae depends on the effect measure being combined. An estimate of the variance of the pooled estimate $\overline{T}.$ is given by the reciprocal of the sum of the weights, i.e.

$$\text{var}(\overline{T}.) = 1\Big/\sum_{i=1}^{k} w_i \qquad (2.3)$$

In the formulae above, the variances of the effect sizes (v_i) used to derive the weightings (w_i) are estimated from the data but are treated if they were the true variance, hence no allowance is made for error in the calculated term v_i i.e. it is assumed that $T_i \sim N\left(\theta, w_i^{-1}\right)$. If w_i were the true inverse variance of T_i, rather than being an estimate, then $\overline{T}.$ would be the maximum likelihood estimate of the true underlying effect size.

If $\overline{T}.$ is assumed to be normally distributed, an approximate 100(1-α)% confidence interval for the population effect, θ, is given by:

$$\overline{T}. - z_{100(1-\alpha/2)}\sqrt{1\Big/\sum_{i=1}^{k} w_i} \le \theta \le \overline{T}. + z_{100(1-\alpha/2)}\sqrt{1\Big/\sum_{i=1}^{k} w_i} \quad , \qquad (2.4)$$

where $z_{100(1-\alpha/2)}$ is the 100(1-α/2) percentage point of a Standardised Normal Distribution. The global null hypothesis that the treatment effect in all studies is equal to 0 is tested by comparing the statistic $U = \left(\sum_{i=1}^{k} T_i w_i\right)^2 \Big/ \sum_{i=1}^{k} w_i$ with the chi-squared distribution with 1 degree of freedom. It should be noted that the variance estimation formula for the standard inverse variance-weighted method can sometimes be biased and too sensitive to the minimum of the estimates of the variances in the *K* studies; an adjusted variance formula is available. (Li et al. 1994)

The above calculations require an estimate of effect size and corresponding variance from each study. The maximum likelihood estimates of these can be used, or, alternatively, if using efficient score and Fisher's information statistics, given by Z_i and V_i respectively (the first and second order derivatives of the log likelihood evaluated at $\theta = 0$ (Whitehead, 1992)), $T_i = \frac{Z_i}{V_i}$ and $w_i = V_i$. Also let $T_i w_i = Z_i$, then

$\bar{T}_{\cdot} = \sum_{i=1}^{k} Z_i \Big/ \sum_{i=1}^{k} V_i$. The approximate distributional result $Z \sim N(\theta V, V)$ can be used when θ is small. The ratio Z/V is an approximate maximum likelihood estimate for θ. The test statistic can be expressed as $U = \left(\sum_{i=1}^{k} Z_i\right)^2 \Big/ \sum_{i=1}^{k} V_i$. It has been termed the 'one-step' estimate because it is obtained on the first step of a Newton-Raphson procedure to maximise the log-likelihood function when the starting value for θ is 0. Although this estimate is asymptotically unbiased under the null hypothesis that θ=0, it becomes increasingly biased the further that θ moves from 0.

2.2.2 Bayesian fixed effect models

Most authors who have considered a Bayesian approach to meta-analysis have implemented random effects models, but fixed effects models are possible. Bayesian methods differ from frequentist ones in that both the data and model parameters are considered to be random quantities, and the likelihood function is thought of as defining the plausibility of the data given values of the model parameters. The model parameters are considered unknown random quantities and prior distributions may be specified for them, which can be based on evidence external to the study, in this case meta-analysis, in question or on subjective *a priori* beliefs. The joint prior probability density function for all the model parameters is then combined with the likelihood function using Bayes' Theorem (Lee, 1989) to obtain the joint posterior probability density function. In this thesis, all posterior densities for all model parameters are derived using MCMC simulation. For example, if the outcome is assumed to be normally distributed, then:

$$\begin{aligned} T_i &\sim N[\theta, v^2 i] \qquad\qquad i = 1 \ldots . k \\ \theta &\sim N[-,-], \end{aligned} \tag{2.5}$$

where θ is the estimate for the underling effect size, v_i^2 is the estimated variance of the effect size (n.b.as for the classical approach the v_i are treated as if they were the true

variances, when in fact they are estimated from the data) and [-,-] indicates a prior distribution to be specified. An important aspect of any Bayesian analysis is the choice of prior distributions. Sutton and Abrams (Sutton and Abrams, 2001) review priors used previously for meta-analyses. This aspect of the analysis is considered further in Chapter 7; as are further Bayesian fixed effect models, which appear appealing for combining data where the outcome is rare

2.2.3 Alternative fixed effect models for combining odds ratios

Other fixed effect methods specific to combining odds ratios have been developed. Under most conditions the estimates obtained from each method should be very similar to one another. However, when the data are sparse, results may differ and some traditional methods may break down altogether. This issue is explored in detail in Chapter 7.

If a comparative binary outcome is being considered, generally it will be possible to construct a 2 by 2 table, for each study, including all the information required for the commonly used outcome measures. Typical 2 by 2 tables for an RCT and a case-control study are presented in Figures 2.1a and 2.1b respectively.

Figure 2.1a - Outcome data from a single RCT

	Failure /Dead	Success /Alive
New Treatment	*a*	*b*
Control	*c*	*d*

Figure 2.1b - Outcome data from a single Case-control study

	Diseased (cases)	Non-diseased (controls)
Exposed	*a*	*b*
Not exposed	*c*	*d*

Mantel-Haenszel method for combining odds ratios

This method was first described by Mantel and Haenszel (Mantel and Haenszel, 1959). The pooled estimate is calculated by:

$$\overline{T}_{MH(OR)} = \frac{\sum_{i=1}^{k} a_i d_i / n_i}{\sum_{i=1}^{k} b_i c_i / n_i}, \qquad (2.6)$$

where a_i, b_i, c_i, and d_i are the four cells of the 2x2 table illustrated in Figure 2.1a and b for the $i = 1 \ldots k$ studies, and n_i is the total number of persons in the ith study.

A variance estimate for the estimated summary odds ratio, $\overline{T}_{MH(OR)}$, is required in order to calculate a confidence interval around this point estimate. The formula commonly used was derived by Robins et al. (Robins et al. 1986a; Robins et al. 1986b) This formula computes a variance estimate for the **log** of $\overline{T}_{MH(OR)}$, as:

$$v_{MH(\ln(OR))} = \frac{\sum_{i=1}^{k} P_i R_i}{2\left(\sum_{i=1}^{k} Ri\right)^2} + \frac{\sum_{i=1}^{k} \left(P_i S_i + Q_i R_i\right)}{2\left(\sum_{i=1}^{k} R_i\right)\left(\sum_{i=1}^{k} S_i\right)} + \frac{\sum_{i=1}^{k} Q_i S_i}{2\left(\sum_{i=1}^{k} S_i\right)^2}, \qquad (2.7)$$

where $P_i = (a_i+d_i)/n_i$, $Q_i = (b_i+c_i)/n_i$, $R_i=a_id_i/n_i$, and $S_i = b_ic_i/n_i$.

A 100(1-α)% confidence interval for the summary odds ratio, θ, is thus given by:

$$\exp\left[\ln\left(\overline{T}_{MH(OR)}\right) - z_{100(1-\alpha/2)}\left(v_{MH(OR)}\right)^{1/2}\right] \leq \theta \leq \exp\left[\ln\left(\overline{T}_{MH(OR)}\right) + z_{100(1-\alpha/2)}\left(v_{MH(OR)}\right)^{1/2}\right], \quad (2.8)$$

where $z_{100(1-\alpha/2)}$ is the 100(1-α/2) percentage point of a Standardised Normal Distribution. Several other variance estimates have been proposed; these are further explored elsewhere (Phillips and Holland, 1987; Robins et al. 1986a; Robins et al. 1986b; Emerson, 1994). Sato (Sato, 1990) has developed a method that works directly on the odds ratio scale (as opposed to ln OR). Simulations have shown (Sato, 1990) that this works as well as the method of Robins et al. given above. If any of the cells of the 2x2 tables are 0 (i.e. there are no events or every person experiences an event) then a continuity correction factor is required, and hence 0.5 is usually added to every cell of the 2x2 table in question.

Peto's method for combining odds ratios

This method was first described by Peto et al. (Peto et al. 1977) and more thoroughly by Yusuf et al. (Yusuf et al. 1985). It can be regarded as a modification of the Mantel-Haenszel method. An advantage over the Mantel-Haenszel method is that it can still be used when cells in individual studies 2 by 2 tables are zero; it is also easy to calculate. Unfortunately Peto's method may produce serious under estimates, (Fleiss, 1994) when the odds ratio is far from unity (i.e. there are large treatment or exposure effects). This is unlikely to be a problem in clinical trials, but could be so in the meta-analysis of epidemiological studies. (Spector and Thompson, 1991)

Defining n_i as the number of patients in the ith trial and n_{ti} as the number in the treatment group of the ith trial, let d_i equal the total number of events from both treatment and control groups, and O_i the number of events in the treatment group. Then E_i, the 'expected' number of events in the treatment group (in the ith trial), can be calculated as $E_i = (n_{ti}/n_i)d_i$. For each study two statistics are calculated: 1) O-E, the

difference between the observed and the number expected under the hypothesis that the treatment is no different from the control; and 2) v, the variance of the difference O-E. For k studies the pooled estimate of the odds ratio is given by (Berlin et al. 1989):

$$\overline{T}_{PETO(OR)} = \exp\left[\sum_{i=1}^{k}(O_i - E_i)\Big/\sum_{i=1}^{k} v_i\right], \tag{2.9}$$

$$\text{where } v_i = E_i\left[(n_i - n_{ti})/n_i\right]\left[(n_i - d_i)/(n_i - 1)\right].$$

An estimate of the approximate variance of the natural log of the estimated pooled odds ratio is provided by:

$$\text{var}(\ln \overline{T}_{PETO(OR)}) = \Big/\left(\sum_{i=1}^{k} v_i\right), \tag{2.10}$$

A 100(1-α)% (non symmetric) confidence interval is thus given by:

$$\exp\left(\frac{\sum_{i=1}^{k}(O_i - E_i) \pm z_{\alpha/2}\sqrt{\sum_{i=1}^{k} v_i}}{\sum_{i=1}^{k} v_i}\right), \tag{2.11}$$

where $z_{\alpha/2}$ is the α/2 percentage point of a Standardised Normal Distribution.

Combining odds ratios via maximum-likelihood techniques

Maximum likelihood estimates are difficult to compute exactly, but they are the most efficient for large sample sizes. Unfortunately, there is no way of knowing how large the sample sizes must be for this property to hold. (Hasselblad and McCrory, 1995) Emerson (Emerson, 1994) reports that Breslow found that unconditional maximum likelihood estimation, which had earlier been investigated by Gart, not consistent for estimating the odds ratio when the number of counts remained bounded.

Conditional maximum likelihood estimates also exist. They use the conditional distribution of the data in each table, given the fixed values for the total counts in the margins. This leads to an estimator that is consistent and asymptotically normal. (Emerson, 1994; Hauck, 1984) It is superior to the unconditional maximum likelihood estimator, and equal or superior to the Mantel-Haenszel estimator in both bias and precision. (Hauck, 1984)

Exact methods of interval estimation

The above methods for interval estimation are all asymptotic; their justification assumes either that the counts are large or that the number of strata is large. Exact methods do exist that are not restrained in this way, and are based on exact distribution theory. Although these methods have long been available in principle, modern computer power (using network algorithms) now makes them routinely available. See Emerson (Emerson, 1994) for a review of this topic.

The relationship between the different classical methods of meta-analysing odds ratios

Using the 2 by 2 table notation outlined in Figure 2.1, the efficient score for the underlying log odds ratio in the *i*th study is $Z_i = b_i - (a_i + b_i)(b_i + d_i)/(a_i + b_i + c_i + d_i)$ and Fisher's information is $V_i = (a_i + b_i)(c_i + d_i)(b_i + d_i)(a_i + c_i)/(a_i + b_i + c_i + d_i)^2(a_i + b_i + c_i + d_i - 1)$ when analysis proceeds using a likelihood which conditions on the total number of successes in the study $(b + d)$. It can be seen that Z_i can be expressed as $O_i - E_i$ of the Peto method in equation (2.9) and is thus based on the above formulae. Hence, estimation could proceed using efficient score statistics and Fisher's information as explained in Section 2.2.1. It should be noted that the Mantel-Haenszel test statistic is the U statistic calculated from the Peto approach. Therefore, the Mantel-Haenszel test statistic is connected with the Peto estimate rather than the Mantel-Haenszel estimate. The Mantel-

Haenszel estimate can be considered as a weighted average of the individual odds ratios, with weights $b_i c_i / n_i$ which approximate the inverse variances of the individual estimates when θ is near 1. (Breslow & Day, 1980) The inverse variance weighed estimate would be optimal, in the sense that it has minimum variance, if the variance of the individual study estimates were known and not estimated from the data.

Bayesian "exact" methods for combining odds ratios

It is possible combine odds ratios using method (2.5), however, more appealing, are specifications which do not require the assumption of normality for pooling odds ratios. Two models are described below that both make the assumption that the number events in each arm of each study are binomially distributed.

A model which only assumes the underlying effect difference between groups is the same between studies (the assumption of the Classical fixed effect approaches) is specified below, using the *a*, *b*, *c*, *d* notation as before

$$a_i \sim Bin[p_{1i},(a_i + b_i)] \qquad c_i \sim Bin[p_{2i},(c_i + d_i)] \qquad i = 1.\ldots.k$$

$$\log it(p_{1i}) = \mu_i \qquad \log it(p_{2i}) = \mu_i + delta \qquad (2.12)$$

$$\mu_i \sim [-,-] \qquad delta \sim [-,-]$$

$$OR = exp(d),$$

Where p_{1i} and p_{2i} are the probabilities of events in the two groups being compared for the *i*th study. μ_i is the estimated ln(odds) of an event in group one, and *delta* is the ln(odds ratio) between groups; priors are required for these parameters.

If it is assumed that the underlying proportion of events in corresponding arms of each of the studies is the same, then the following model can be fitted.

$$a_i \sim Bin[p_1,(a_i+b_i)] \quad c_i \sim Bin[p_2,(c_i+d_i)] \quad i=1.....k$$

$$p_1 \sim Beta[-,-] \qquad p_2 \sim Beta[-,-] \qquad (2.13)$$

$$OR=(p_1\times(1-p_2))/(p_2\times(1-p_1))$$

Note, p_1 and p_2 do not have to be given Beta distribution priors, however these are the conjugate choice (Lee 1989).

The key difference between these models and (2.5) is the assumption that at the lowest level of the model the responses in each arm of a study can be modelled directly. In (2.5) calculation of the log odds ratio, when there are zero or complete responses in either arm of a study, requires a continuity correction factor to be added. It is this assumption of Normality of the log odds ratio, or other transformed measures of binary data, in models such as (2.5) that is frequently not valid. Similar "exact" models have been developed for the risk difference outcome (Carlin, 2000); these are described in Chapter 7.

2.2.4 Discussion of the relative merits of each method

With a number of different approaches to combine odds ratios available, it would be desirable to have guidelines indicating which particular method is most appropriate in which circumstances.

As mentioned previously, the Peto method has come under strong criticism. It has been demonstrated that this method may produce seriously biased odds ratios and corresponding standard errors when there is severe imbalance in the numbers in the two groups being compared. (Greenland and Salvan, 1990) Bias is also possible when the estimated odds ratio is far from unity. (Fleiss, 1993) Fleiss (Fleiss 1981) describes conditions under which the inverse-weighted and the Mantel-Haenszel method are to be

preferred: if the number of studies to be combined is small, but the within-study sample sizes per study are large, the inverse-weighted method should be used. If there are many studies to combine, but the within-study sample size in each study is small, the Mantel-Haenszel method is preferred.

A comparison between the Mantel-Haenszel and (conditional and unconditional) maximum likelihood techniques has been carried out. Generally if the sample sizes of the studies are large (all cells >=5) the methods will give almost identical results, otherwise differences between the methods will be small. As there seem to be no clear benefits to be reaped from the difficult computation of the maximum likelihood method, using the inverse-weighted and Mantel-Haenzel methods when indicated would seem the best strategy in most cases. If, however, samples sizes are small for individual studies exact methods may be preferred. (Greenland and Salvan, 1990)

Another consideration when deciding which method to use is whether any of the trials arms have zero observed events. Using the Mantel-Haenszel estimate, a study with zero total events is completely excluded from the analysis if no continuity correction is used. This is unappealing as a trial with zero events from 200 subjects would then be equally as non-informative as a trial with only 20 subjects. A recent investigation into this problem recommended that a continuity correction (adding 0.5 to each cell) should be used for sparse data in meta-analysis, except in the situation when there is strong evidence suggesting that very little heterogeneity exists among component studies. (Sankey et al. 1996) The issue of continuity correction factors for meta-analysis are considered in detail in Chapter 7. At the time of writing, simulation studies appeared to show that the Peto method outperformed other simple methods, including the Mantel-Haenszel method and the standard inverse variance-weighted method (section 2.2.1) when there were small numbers of events in one or more cells of studies 2 by 2 tables. (Deeks et al. 1999)

None of the comparative assessments cited above have compared the performance of the Bayesian "exact" models against the classical ones.

2.3 Random effect models

Random effects models have been advocated as a more conservative alternative to fixed effect models. This approach assumes the studies are estimating different (underlying) effect sizes, and takes into account the extra variation implied in making this assumption.(Whitehead and Whitehead, 1991) More specifically, these underlying effects are assumed to vary at random, and typically the distribution of such effects is assumed to be normal. Hence this model includes two sources of variation; the between- and within-study variance.

While many consider random effects models always to be preferable for combining data from medical studies, the decision on which model to use can also be made by considering a statistical test for heterogeneity. Several slightly different formulae for a general test are, for the most part, essentially equivalent, being based on χ^2 or F statistics. (Dickersin and Berlin, 1992). The one devised by Cochran (Cochran, 1954), which is widely used, is given below. It tests the hypothesis that the true treatment effects are the same in all the primary studies (H_0: $\theta_1 = \theta_2 = \cdots = \theta_k$, where the θ_i's are the underlying true treatment effects of the corresponding, $i = 1$ to k, studies in the meta-analysis), versus the alternative that at least one of the effect sizes (θ_i) differs from the remainder.

Essentially, this is testing whether it is reasonable to assume that all the studies to be combined are estimating a single underlying population parameter and whether variation in study estimates is likely to be wholly random. This is essentially testing the assumption underlying the fixed effect model. The test statistic is

$$Q = \sum_{i=1}^{k} w_i \left(T_i - \overline{T}.\right)^2, \qquad (2.14)$$

where k is the number of studies being combined, T_i is the treatment effect estimate in the ith study, $\overline{T}.$ is the inverse variance weighed estimate of treatment effect, and w_i is

the weight attached to that study in the meta-analysis. Q is approximately distributed as a χ^2 distribution on $k-1$ degrees of freedom under H_0. Unfortunately, when the number of studies in a meta-analysis is only small or moderate, the low power of the test can make interpretation difficult.

2.3.1 Standard Classical model

The standard random effects model used in meta-analysis was described by DerSimonian and Laird.(DerSimonian and Laird, 1986) The model assumes that the study specific effect sizes come from a random distribution of effect sizes with a fixed mean and variance. Expressed algebraically, where T_i is an estimate of effect size and θ_i is the true effect size in the ith study

$$T_i = \theta_i + e_i, \tag{2.15}$$

where e_i is the error with which T_i estimates θ_i, and

$$\text{Var}(T_i) = \tau^2_\theta + v_i, \tag{2.16}$$

where τ^2_θ is the random effects between study variance and v_i is the variance due to sampling error in the ith study. If the random effects variance was zero the above model would reduce exactly to the fixed effects model.

Formulae can be derived using both a weighted and an un-weighted approach; these can be estimated using four different methods; weighted and un-weighted least squares (WLS, UWLS), and maximum and restricted maximum likelihood (ML & REML); the latter two assume normality of the underlying effect parameters. The likelihood to be maximised is slightly modified using REML (from that of ML), to adjust for the fact that the underlying mean and variance are being estimated from the same data. The REML estimates are the iterative equivalent to the weighted estimators. (DerSimonian and Laird, 1986) The relative merits of each of the above methods have not been widely

investigated, however the WLS approach has become the standard.(Shadish and Haddock, 1994) For derivations and formulae see Shadish and Haddock (Shadish and Haddock, 1994).

2.3.2 Extensions to the Classical model

Including uncertainty induced by estimating the between study variance

Although the random effects model gives wider confidence intervals than that of a corresponding fixed effect analysis, they are still too narrow, because the method assumes the between study variance is known, when in fact it is estimated from the data.(Hardy and Thompson, 1996; Biggerstaff and Tweedie, 1997; Biggerstaff, 1997) Two modifications have addressed this problem. Hardy and Thompson (Hardy and Thompson, 1996) developed an approach using profile likelihood methods, which assumes normality of the data, to calculate appropriate confidence regions. This method still uses the study estimates of each individual study variance as the true underling variance. If a full likelihood method were pursued, allowing for this uncertainty, the confidence intervals for the overall treatment effect would be expected to be even wider. Except when all the trials are small, the additional uncertainty would not be expected to have a great impact on the results and so pursuing a full likelihood approach is unnecessarily sophisticated for most practical purposes.(Hardy and Thompson, 1996) (However, a full likelihood approach for binomial data (which includes the conditional distribution of each 2 by 2 frequency table given its margins), is discussed below(Van Houwelingen et al. 1993) since it offers other advantages.)

Biggerstaff and Tweedie (Biggerstaff and Tweedie, 1997) address the same problem by developing a variance estimator for Q (the between study heterogeneity statistic), that leads to an interval estimation of $\hat{\tau}^2$, utilising an approximating distribution for Q. They also develop asymptotic likelihood methods for the same estimate. This information is then used to give a new method of calculating the weight given to the individual studies which takes into account variation in these point estimates of $\hat{\tau}^2$. These new weights are between the standard fixed and random effects, in terms of effect on down-weighting

the results of large studies and up-weighting those of small. (A past concern has been that when $\hat{\tau}^2$ is large the standard random effects model gives too much weight to the relatively small studies.) These new weights will differ most from those of the standard random effects model when the number of studies to be combined is small; the results are similar to the standard random effects model when 20 or more studies are to be combined.

Exact approach to random effects meta-analysis of binary data

Van Houwelingen developed a likelihood based approach to random effects, for binary data, (Van Houwelingen et al. 1993) which avoids use of approximating Normal distributions and can be used when the assumptions of normality are violated, an assumption which is rarely checked in meta-analyses. Solutions are obtained via the EM algorithm (Dempster et al. 1977). An extension is given to a bivariate random effects model, in which the effects in both groups are supposed random. In this way inference can be made about the relationship between improvement and baseline effect. This is a non-parametric procedure, and has been recommended (Hardy and Thompson, 1996) when the normality assumption is violated.

2.3.3 Bayesian random effect models

General model for normally distributed data

Many of the authors who have considered a Bayesian approach to meta-analysis have implemented a hierarchical model in which various assumptions of Normality have been made. (DuMouchel, 1989; DuMouchel, 1994b; DuMouchel and Harris, 1983; Waternaux and DuMouchel, 1993; DuMouchel, 1994a; Abrams and Sanso, 1998; Verdinelli et al. 1995) This mirrors the Classical approach of DerSimonian and Laird.

$$T_i \sim \mathrm{N}[\theta_i, \sigma^2_i] \qquad \sigma^2_i \sim [\text{-},\text{-}] \qquad i = 1,..,k$$

$$\theta_i \sim \mathrm{N}[\mu, \tau^2] \tag{2.17}$$

$$\mu \sim [\text{-},\text{-}] \qquad \tau^2 \sim [\text{-},\text{-}].$$

It should be noted that full uncertainty in the parameter estimates is (automatically) taken into account in the estimation of posterior parameters, so that unlike the Classical approach, no extension is required to allow for this.

"Exact" model for combining odds ratios

An exact random effect Bayesian models can be constructed by extending (2.12), and has been adopted by a number of authors (Skene and Wakefield, 1990; Rogatko, 1992; Smith et al. 1995b; Smith et al. 1995a; Waclawiw and Liang, 1994; Higgins and Whitehead, 1996; Spiegelhalter et al. 1994).

$$a_i \sim Bin[p_{1i},(a_i+b_i)] \qquad c_i \sim Bin[p_{2i},(c_i+d_i)] \qquad i = 1.....k$$

$$\log it(p_{1i}) = \mu_i \qquad \log it(p_{2i}) = \mu_i + delta_i$$

$$delta_i \sim \mathrm{N}[\phi,\tau^2] \tag{2.18}$$

$$\mu_i \sim [-,-] \qquad \phi \sim [\text{-},\text{-}] \qquad \tau^2 \sim [\text{-},\text{-}],$$

where ϕ represents the overall pooled effect, on a log odds ratio scale, and τ^2 is a measure of the between-study heterogeneity. A similar extension of (2.13) has also been implemented, (Byar, 1980) however this is not pursued here.

An alternative to equations (2.17) or (2.18) is to assume the random effects are *t* rather than normally distributed. (Smith et al. 1995b; Seltzer, 1991) This may be sensible when the number of studies being combined is small, and it is difficult to assess whether the normality assumption of the random effects has been violated since it provides a more robust estimation procedure.

2.3.4 Comparison of random with fixed effect models

The argument over which model is theoretically and/or practically superior has been running for many years with many comments scattered through the literature. Investigations into the differences between results obtained from the two methods have been made. (Berlin et al. 1989; Mengersen et al. 1995) For example, Berlin et al. (Berlin et al. 1989) compared the results of 22 meta-analyses; in three, different conclusions would have been drawn about the treatment effect, (Dickersin and Berlin, 1992) the Peto fixed effect method suggesting a beneficial treatment effect while the random effect method did not. Random effects models have been criticised on grounds that unrealistic/unjustified distributional assumptions have to be made.(Peto, 1987) However it has also been argued that they are consistent with the standard specific aims of generalisation.(Raudenbush, 1994) A further consideration is that random effects models are more sensitive to publication bias because of the greater relative weight given to smaller studies (see Chapter 6).(Greenland, 1994) Perhaps it is wise to conclude that neither fixed nor random effect analyses can be considered ideal.(Thompson, 1993)

In the context of the generalised synthesis of evidence, a random effect approach would often appear to be the obvious choice, since, by definition, data from heterogeneously designed studies is being combined.

2.4 Exploring heterogeneity

Random effect models account for heterogeneity between studies, but they do not provide a method of exploring and potentially of explaining the reasons study results vary. Investigating why study results vary systematically may lead to the identification of associations between study or patient characteristics and the outcome measure, which would not have been possible in single studies. For meta-analyses of RCTs this in turn may lead to clinically important findings and may eventually assist in individualising treatment regimes.(Gelber and Goldhirsch, 1987)

Regression type models can be used to explore reasons why study results may systematically differ. Alternatively, if discrete factors are being explored, simple subgroup methods may suffice, although this special case is not pursued explicitly in this thesis. Both study and patient characteristics can be explored using these methods. It should be stressed, however, that this type of analysis should be treated as exploratory as associations between their characteristics and the outcomes can occur purely by chance, or due to the presence of confounding factors. Further, regression analysis of this type is also susceptible to aggregation bias, which occurs if the relation between patient characteristic study means and outcomes do not directly reflect the relation between individuals' values and individuals' outcomes.(Lau et al. 1998; Lambert et al. 2001a) A further restriction is that the data available for analysis from original study reports may be limited.

2.4.1 Meta-regression

Two types of regression models are possible; one is an extension of the fixed effect model, commonly known as a meta-regression model, and the other an extension of the random effects model, called a mixed model (because it includes both fixed and random terms). The fixed-effect methods are most appropriate when all variation (above that explainable by sampling error) between study outcomes can be considered accountable by the covariates included. A mixed-model is more suitable when the predictive covariates explain only part of the variation/heterogeneity, and a random effect term is used to account for the remainder. However, as with fixed and random models themselves, it has been argued that one should always include a random effect term as there will always be some degree of between study heterogeneity not captured by the covariates. It should be noted that regression models are most useful when the number of studies is large, and cannot be sensibly attempted when very small numbers of studies are being combined. (Raudenbush, 1994)

The fixed standard fixed effect regression model is not outlined explicitly, since it can be considered a special case of the more general mixed effect model, with the heterogeneity term set to zero.

2.4.2 Classical mixed effect model (random-effects regression)

As a starting point take the random effects model outlined previously (2.15): $T_i = \theta_i + e_i$, where T_i is the estimated effect size of the true effect size θ_i for each of the k studies, $i = 1, \ldots, k$. We also assume the e_i are statistically independent, each with a mean of zero and estimation variance v_i. The variance for these estimates of treatment effect, as before (2.15) can be expressed as: $Var(T_i) = \tau^2_\theta + v_i$, where τ^2_θ is the between study, or random effects variance and v_i is the within study variance. Now we extend this to formulate a model for the true effects depending on a set of study characteristics plus error: (Raudenbush, 1994)

$$\theta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + u_i, \qquad (2.19)$$

where

β_0 is the model intercept;

$X_{i1}, \ldots, X_{ip}$ are coded characteristics of studies hypothesised to predict the study effect size;

$\beta_1, \ldots, \beta_p$ are regression coefficients capturing the association between study characteristics and effect sizes;

u_i is the random effect of study i, that is, the deviation of the true effect in study i from the value predicted on the basis of the model. Each random effect, u_i, is assumed independent, with a mean of zero and variance σ^2_i.

Under the fixed effects specification, the study characteristics $X_{i1}, \ldots, X_{ip}$ are presumed to account completely for variation in the true effect sizes. In contrast, the random effects specification assumes that part of the variability in these true effects is unexplainable by the model. This model is a consistent extension of the models presented previously. If the model has no predictors, i.e. $\beta_1 = \ldots\ldots = \beta_p = 0$, then it reduces to that of the random effects model. If the random effects variance is null i.e. $\tau^2_\theta = 0$, then the results will be identical to that of the fixed effects meta-regression model.

Estimation of the parameters of this model can be achieved using an iterative least squares algorithm; see Raudenbush (Raudenbush, 1994) for details. Several other methods can be used to obtain solutions, for example the method of maximum likelihood can be used. This requires a further assumption that each T_i is normally distributed (see section 6.4.1). (Huque and Dubey, 1994) Small biases have been found using the iterative least squares approach (Berkey et al. 1995) which can be eliminated if an empirical Bayes estimate is used. (Berkey et al. 1995) Although not straightforward, it is also possible to derive a moment estimator for τ^2, the between study variance in a mixed model, when only one covariate is included. This is a direct extension of the DerSimonian and Laird weighted least squares random effect model; Thompson et al. provide equations (Thompson et al. 1997) and also discuss further ML and REML estimates.

2.4.3 Alternative formulation of Classical meta-regression models

Probably the most notable alternative to the regression models above is the use of a logistic regression model when the outcome is an odds ratio. Such an approach has advantages similar to those of the exact Bayesian models for combining odds ratios, including the removal of the normality assumption of the individual effect sizes, and the need for continuity correction factors for studies with arms with no/all events.

A fixed effect logistic regression model is simple to implement, and has been used by several authors. (Detsky et al. 1992; L'Abbe et al. 1987; Thompson, 1993; Thompson et al. 1997) Recently, due to advances in computational power and software, it is possible to implement random effect logistic regression models (Thompson et al. 1997) which include a between study variation term.

2.4.4 Bayesian meta-regression models

Covariates can be included in a Bayesian meta-analysis model in a straightforward way. For example, including a study level covariate, denoted x_i in equation (2.18) produces the following model. (Smith et al. 1995b)

$$a_i \sim Bin[p_{1i},(a_i+b_i)] \qquad c_i \sim Bin[p_{2i},(c_i+d_i)] \qquad i = 1.....k$$

$$\log it(p_{1i}) = \mu_i \qquad \log it(p_{2i}) = \mu_i + delta_i + \beta x_i \tag{2.20}$$

$$delta_i \sim N[\phi, \tau^2] \qquad \beta \sim [-,-]$$

$$\phi \sim [-,-] \qquad \tau^2 \sim [-,-],$$

The equivalent extension of model (2.17) for normally distributed outcomes, and for the fixed effect models ((2.12) and (2.13)) are all straightforward.

2.4.5 Modelling patients' underlying risks

Studies may appear heterogeneous because of differences in the baseline risk of the patients. If the overall effectiveness of a new treatment is related to the severity of the disease, this could affect decisions about which patients should be treated. (Thompson et al. 1997). The usual way of investigating baseline risk within trials is to consider the observed risk of events in the control group (or sometimes the average risk in the control and treatment groups).(Brand and Kragt, 1992) This variable could be included in a regression model in an effort to explain heterogeneity between study results. However, this type of analysis can lead to bias, (Senn, 1994) since this measure of baseline risk forms part of the definition of the treatment difference (i.e. the event risk in the control group is used in the calculation of an odds ratio, relative risk etc.). In other words, if by chance, the risk of an event in the control group is low then the estimated treatment effect will be greater (OR will be further from 1). If, on the other hand the event risk in the control group is high, then the OR will be nearer 1. Thus, even if there is no true relationship between baseline risk and treatment effectiveness, one is likely to be observed due to this statistical artefact - regression to the mean. (Thompson et al. 1997) This problem is reduced if studies are large (leading to less random variation in control risk) and if there are a large number of studies. Alternative models have been developed to avoid the problem of "regression to the mean".(McIntosh, 1996; Cook and

Walter, 1997; Thompson et al. 1997) One of these, an extension of the Bayesian random effects model described in equation (2.18), is outlined below

$$a_i \sim Bin[p_{1i},(a_i+b_i)] \qquad c_i \sim Bin[p_{2i},(c_i+d_i)] \qquad i=1.....k$$

$$\log it(p_{1i}) = \mu_i \qquad\qquad \log it(p_{2i}) = \mu_i + delta_i \tag{2.21}$$

$$delta_i = delta_i' + \beta\left(\mu_i - \bar{\mu}\right) \qquad \beta \sim [-,-]$$

$$delta_i' \sim N[\phi, \tau^2]$$

$$\mu_i \sim [-,-] \qquad \phi \sim [-,-] \qquad \tau^2 \sim [-,-],$$

where $\bar{\mu}$ denotes the average of the μ_i trials, and d_i' the treatment effect in trial i adjusted for underlying risk. This model can be extended to include further study level covariates following model (2.20). One potential drawback is that this method only works on the log odds ratio scale; using other scales is possible in principle but currently difficult in practice.

One method of avoiding the use of patient risk directly would be to use a prognostic score based on patient covariates (i.e. predictors of risk) and then relate treatment effects to this score for individual patients. (Thompson et al. 1997) Thompson et al. suggest the prognostic score would best be based on data other than that from the trials which form the meta-analysis for treatment effects. A related approach is considered using the net benefit model; see Section 3.11 and Chapter 8.

Another alternative, if individual patient data are available, is to relate treatment effects to individual patient covariates in an attempt to investigate heterogeneity. This avoids the problems discussed above and would be directly useful to the clinician considering treatment for an individual patient.(Thompson et al. 1997)

2.4.6 Generalisation and extensions to meta-regression models

Stram (Stram, 1996) presents a very general mixed-effects regression model framework. He developed a model from which most other models used in meta-analysis can be viewed as special cases. Explicitly this model builds on the standard random effects model, (Peto, 1987) the mixed model, the model of Begg and Pilote (Begg and Pilote, 1991)- to incorporate single arm studies with two-arm studies (see Section 3.4), and the model of Tori et al.(Tori et al. 1992) for combining surrogate endpoints. The general form of the model is:

$$\mathbf{Y}_i = \mathbf{X}_i\alpha + \mathbf{Z}_i\beta_i + \zeta_i + \mathbf{e}_i\,, \qquad (2.22)$$

where there are $i = 1, 2, \ldots\ldots, K$ independent studies. Y_i is an $(n_i \times 1)$ vector of one or more related estimates of treatments or treatment comparisons of interest; X_i is an $(\mathbf{n_i} \times \mathbf{p})$ matrix of known covariates related to the p vector of unknown fixed effect parameters, α; and Z_i is an $(n_i \times q)$ vector of known covariates related to a $(q \times 1)$ vector of unobserved random effects, β_i, for each study. The two remaining $n_i \times 1$ unobserved random vectors, ζ_i and e_i, specify two types of error in Y_i. The ζ_i specify the sampling errors in Y_i, and e_i specifies other sources of error or heterogeneity between studies and between arms of the same study.

In this model it is assumed that β_i, u_i and e_i are each independent multivariate normal random vectors. One of the new extensions offered by this model is the possibility for random effect covariates.

Multi-level or hierarchical models, which can also implement weighted random effect regression, have been applied to meta-analysis (Lambert and Abrams, 1996). This is an area of current research. Such models have the potential to combine summary data with individual patient data, including both study level and patient level covariates. (Goldstein et al. 2000; Turner et al. 2000) Further extensions of mixed models are discussed in Chapter 3 for the generalised synthesis of evidence.

2.5 Issues/methods in the synthesis of observational studies

Meta-analyses of epidemiological studies, seeking to identify risk factors for disease, are carried out and inform public health initiatives. The use of such a technique in this area is not without dispute, and certainly problems additional to those encountered when carrying out a meta-analysis of RCTs arise. (Spitzer, 1991; Greenland, 1987)

One particular problem is the lack of standardisation in the way in which results of epidemiological studies are reported. For example, exposure groups using different exposure level cut-off points may have been used, and estimates adjusted for different covariates. There is also the concern that biases from the individual studies may propagate through the analysis. Two papers which consider many of the difficulties in the pooling of observational study in detail are Greenland (Greenland, 1987) and Chêne and Thompson (Chene and Thompson, 1996).

Clearly, if evidence is being combined from observational and randomised sources, full consideration needs to be given to the special problems of the synthesis of the observational evidence.

2.6 Publication bias: a threat to the validity of a meta-analysis

In order to avoid drawing unbiased conclusions from a meta-analysis it is important that all, or at the very least, the majority of, the relevant primary studies be identified on a given subject. Unfortunately, even comprehensive searches of the literature (including grey material) and the use of other less formal methods such as personal communication may not produce an unbiased sample of studies. It has long been accepted that research yielding statistically significant results is potentially more likely to be submitted, published or published more rapidly than work with null or non-significant results (Easterbrook et al. 1991), which leads to an over representation of false-positive results in the literature (Begg and Berlin, 1989). The implications of this for meta-analysis are

that, combining only the identified published studies uncritically may lead to an incorrect, usually over optimistic, conclusion.

2.6.1 Detecting publication bias

Several methods are available to assess if publication bias is present in a meta-analytic dataset. These include the visual inspection of a funnel plot and two statistical tests.

Funnel plot

The results from smaller studies will be more widely spread around the mean effect because of larger random error. A plot of sample size versus treatment effect from individual studies in a meta-analysis should thus be shaped like a funnel if there is no publication bias. (Light and Pillemar 1984) If the chance of publication is greater for studies with positive statistically significant results, or larger effect size estimates, or some other less defined mechanism, the shape of the funnel plot may become skewed. When the true outcome effect is small but not zero, small studies reporting a small effect size will not be statistically significant and therefore less likely to be published, while small studies reporting a large effect size may be statistically significant and more likely to be published. Consequently there will be a lack of small studies with small effect estimates in the funnel plot, and the funnel plot will be skewed with a larger effect among smaller studies and a smaller effect among larger studies. (Light and Pillemar 1984) This will result in an overestimation of the treatment effect in a meta-analysis.

A problem with the funnel plot assessment is that the appearance of the plot may change depending the scale which is used to represent the notion of study size, and the outcome measure used. Additionally, a skewed funnel plot may be caused by factors other than publication bias. (Egger et al. 1997) For example, it has been shown that if the quality of studies varies with the study size, a funnel plot may give the visual impression of publication bias when this is really confounded by study quality. (Petticrew et al. 1999) Other possible sources of asymmetry in funnel plots include different intensity of intervention, differences in underlying risk, poor methodological design of small

studies, inadequate analysis, fraud, choice of effect measure, and chance. (Egger et al. 1997) Funnel plots for sparse data are explored in Chapter 6.

Rank correlation test

The rank correlation test (Begg and Mazumdar, 1994) examines the association between effect estimates from the primary studies and their variances, to exploit the fact that publication bias will tend to induce a correlation between the two factors (i.e. smaller studies (with larger variances) will tend to have larger effect size estimates), and constructs the rank-ordered sample on the basis of one of them. The test is a distribution-free method, which involves no modelling assumptions, but it suffers from a lack of power and so the possibility of publication bias cannot be ruled out even when the test is non-significant. Define the standardised effect sizes of the k studies to be combined to be

$$T_i^* = \left(T_i - \overline{T}_\bullet\right) \Big/ \left(\widetilde{v}_i^*\right)^{1/2}, \tag{2.23}$$

where

$$\overline{T}_\bullet = \left(\sum_{j=1}^{k} v_i^{-1} T_j\right) \Big/ \sum_{j=1}^{k} v_i^{-1}, \tag{2.24}$$

and T_i and v_i are the estimated effect size and sampling variance from the ith study,

and

$$\widetilde{v}_i^* = v_i - \left(\sum_{j=1}^{k} v_j^{-1}\right)^{-1}, \tag{2.25}$$

which is the variance of $\left(T_i - \overline{T}_\bullet\right)$.

It is then necessary to evaluate P, the number of all possible pairings in which one factor is ranked in the same order as the other, and Q, the number in which the ordering is reversed. A normalised test statistic is obtained by calculating

$$Z = (P - Q)/\left[k(k-1)(2k+5)/18\right]^{1/2}, \qquad (2.26)$$

which is the normalized Kendall rank correlation test statistic for data that have no ties, though this can be relaxed. (Steichen, 1998) This statistic is compared to the standardised normal distribution. Any effect size scale can be used as long as it is assumed its distribution is asymptotic normal.

This test can be considered complementary to the funnel plot. Begg (Begg, 1994) suggests using a very liberal significance level. Additionally, due to the test having very low power for meta-analyses including only small numbers of studies, more emphasis should then be given to an informal visual inspection of the funnel plot. (Begg and Mazumdar, 1994)

Linear regression test

To test the asymmetry of a funnel plot, Egger et al. (Egger et al. 1997) suggested a method based on a regression analysis of Galbraith's radial plot (Galbraith, 1988). As before, for $i = 1 \ldots k$ studies in the meta-analysis, let T_i and v_i be the estimated effect sizes and sample variances from each study. Define the standardised effect (z-statistic) as $T_i^* = T_i / v_i^{1/2}$, the precision as $s^{-1} = 1/v_i^{1/2}$, and the weight as normal ($w_i = 1/v_i$). To perform the test T^* is fitted to s^{-1} using standard weighted linear regression with weights w and equation $T^* = \alpha + \beta s^{-1}$.

The intercept $\hat{\alpha}$ is used to measure asymmetry; if it is estimated to be significantly different from 0 then it is concluded that there is evidence of publication bias in the meta-analysis dataset. A negative intercept indicates that smaller studies are associated with bigger effects. In their original paper Egger et al. also performed an un-weighted regression (Egger et al. 1997) and reported the most statistically significant of the weighted and un-weighed results. By applying this method, Egger et al. (Egger et al. 1997) observed significant asymmetry in 38 per cent of published meta-analyses in a selection of journals and in 13 per cent of Cochrane reviews. From comparisons (Egger

et al. 1997; Sutton et al. 2000) between the tests it would appear that the linear regression test is more powerful that the rank correlation test, however, a considerable discrepancy in the results of the two tests has been observed. (Sutton et al. 2000)

2.6.2 Assessing the likely impact of publication bias

The development of methods to assess the impact of publication bias on the results of a meta-analysis has been an active area of interest for several years. While it is only possible to give a brief summary of most of them here, a detailed review of the area has been published elsewhere. (Sutton and Song, 1999)

The first method to be developed was called the 'file drawer estimate'. In essence, this method considers the question: "how many new studies averaging a null result are required to bring the overall treatment effect to non-significance? " (Rosenthal, 1979). It was developed by Rosenthal (Rosenthal, 1978; Rosenthal, 1979), as it could be seen as estimating the number of studies filed away by researchers' without being published.

Following this more sophisticated methods were developed using weight functions. Weight functions are used to adjust results where only partial information is available, and the chance of having particular data is related to a feature of the data. (Iyengar and Greenhouse, 1988). Hence, in a meta-analysis setting, weight functions are used to model the selection process and develop estimation procedures that take that selection process into account. (Hedges, 1992) These were first introduced into meta-analysis by Hedges.(Iyengar and Greenhouse, 1988) There are two aspects to such models: a) the effect size model which specifies what the distribution of the effect size estimates would be if there were no selection; and b) the selection model which specifies how this effect size distribution is modified by selection. (Hedges and Vevea, 1996) Usually these models assume that the chance of a study being included in the meta-analysis is related to the statistical significance of its outcome (implying journals are more likely to publish significant results than non-significant ones). In these instances the outcome considered is the observed p-value. Both Classical and Bayesian formulations have been developed.

Recently, Copas (Copas, 1999) has presented a method for adjusting for publication bias based on a method described by Copas and Li. (Copas and Li, 1997) Using this method, the process of study selection is assumed to be described by a separate regression model with residuals which are correlated with study outcome. A random effects meta-analysis model is used, (DerSimonian and Laird, 1986) together with a separate selection equation with a single correlation parameter ρ linking selection to outcome. A likelihood approach is taken but the model cannot be fully identified without strong and unverifiable assumptions, so a sensitivity approach based on an overall probability of study selection is adopted.

Trim and Fill

Finally another new method called 'trim and fill' is considered in more detail here, since it is employed/developed further in Chapter 6. This new development formalises the use of funnel plots, and estimates and adjusts for the numbers and outcomes of missing studies. (Duval and Tweedie, 2000a; Duval and Tweedie, 2000b) An iterative rank-based algorithm estimates how many studies are missing. This number of studies are "trimmed" from the asymmetric outlying part of the funnel (i.e. those with the largest effect size estimates): these can broadly be thought of as studies which have no counterpart on the other side of the funnel plot (i.e. this is the truncation that is picked up by 'eye-balling' a funnel plot). Then the symmetric remainder are used to estimate the 'true centre' of the funnel using standard meta-analysis techniques. The 'Trimmed' studies are then replaced and their 'missing counterparts' imputed or 'Filled': these are mirror images of the 'Trimmed' studies with the mirror axis placed along the adjusted pooled estimate. This last stage is necessary for the variance of the pooled estimate to be calculated correctly.

This approach assumes studies are suppressed and not published under a scenario where it is the magnitude of the effect size, and not the p-value which determines the chance of publication (the size of the studies is not taken into consideration). The key assumption of the method is that it is the most extreme negative studies (i.e. those with the smallest outcome estimates) which have not been published. Three different iterative estimators for the number of missing studies have been derived. Simulation studies (Duval and

Tweedie, 2000a; Duval and Tweedie, 2000b) suggest two of these estimates work well in all but very extreme cases, however no one estimate is always superior to another. The first is

$$R_0 = \gamma^* - 1, \tag{2.27}$$

where γ^* is the rightmost run of ranks associated with positive values of the values of the observed effect sizes minus the current estimate of the global effect size (δ_i). These values change iteratively as the current estimated R_0 extreme studies are trimmed before calculation of the next iteration. The second estimate is

$$L_0 = \left[4T_n - n(n+1)\right] / \left[2n - 1\right], \tag{2.28}$$

where T_n is the sum of the ranks of the absolute values of the δ_i for positive δ_i only (the Wilcoxon statistic for the dataset). Like, R_0, L_0 is calculated iteratively until convergence is achieved.

A test based on this approach also appears powerful compared to those described previously, (Begg and Mazumdar, 1994; Egger et al. 1997) if there are more than 5 or 6 missing studies. This method is much simpler to compute than those using selection modelling, which was a motivating reason for its development.

2.7 Study quality: a further threat

The concept of judging research quality in synthesis dates back to Glass in 1976 (Glass, 1976). The primary concern is that combining study results of poor quality may lead to biased, and therefore misleading, pooled estimates being produced. Sophisticated analyses will not eliminate the limitations of poor data,(Thacker, 1988) "... in some respects, the quantitative methods used to pool the results from several studies in a meta-analysis are arguably of less importance than the qualitative methods used to determine which studies should be aggregated." (Naylor, 1988) However, assessment of

quality is controversial; Greenland (Greenland, 1994) has indicated that quality assessment is the most insidious form of bias in the conduct of meta-analysis.

There are at least three approaches for assessing research quality. The first system (Wortman, 1983) applies the validity framework developed by Cook and Campbell (Cook and Campbell 1979), and focuses on non-randomised studies often found in the social science literature (Wortman, 1994). The second is via a quality scoring system, the first of which was developed by Chalmers et al. (Chalmers et al. 1981; Sacks et al. 1987) for assessing RCTs exclusively, although checklists were available before this (Moher et al. 1995). The objective of these scales is to provide an overall index of quality (a comparison with the validity framework approach can be found in Wortman (Wortman, 1994)). Since these first attempts, many different scales and checklists have been developed; for a review of these for RCTs see, Moher et al. (Moher et al. 1995) and for those for observational studies see Deeks et al (Deeks et al. 1996). Recently, a scale which could assess the quality of both randomised and observational studies has been developed, (Downs and Black, 1998) which obviously has implications for quality assessments made in the generalised synthesis of evidence, since studies with different designs can be assessed using the same instrument. Finally, using individual markers of quality e.g. randomisation procedure, can be considered as a third alternative. (Moher et al. 1996)

An appealing feature about using a scale is that it provides an overall quantitative estimate of quality. However the validity of many of the present scales has been criticised. It has been suggested (Moher et al. 1998a) that most scales have been developed in an arbitrary fashion with no attention to accepted methodological standards. Additionally, many scales are not truly measuring quality but focus on extraneous factors more related to the adequacy of reporting or generalisability. (Moher et al. 1998a)

Unfortunately, no clear association between study quality and study results exists consistently across all trials. However, an empirical study has shown (Schulz et al. 1995) that over a large number of RCTs, encompassing many subject areas, inadequate methodological approaches, particularly those representing poor allocation concealment,

are associated with bias. Dickersin and Berlin (Dickersin and Berlin, 1992) review of meta-analyses of both RCTs and observational studies which have addressed this issue.

While a detailed review of the specific scales and checklists and their pros and cons is beyond the scope of this chapter, it is important to note that large differences in results can be observed by using different scales, at least those for RCTs, which is where empirical investigations have been focused. (Moher et al. 1996; Moher et al. 1998a; Juni et al. 1999)

2.7.1 Incorporating study quality into a meta-analysis

Once a formal assessment of study quality has been made, using a measurement scale, or individual quality markers, a decision has to be made how to use this information. Incorporating study quality into a meta-analysis can be considered a special case of exploring heterogeneity, i.e. to what extent does variation in measures of quality between studies explain variation in estimates of treatment or exposure effects?

A plot of study effect against quality score can be examined, or a cumulative meta-analysis, ordering by quality score can be carried out. (Detsky et al. 1992) Alternatively, a regression analysis can be carried out either including indicator variables for individual components of quality, or a quality score based on a scale.

Rather than weight each study by a measure derived from its precision, as is normally done in a meta-analysis, each of the individual study estimates could be weighted by a variable which measures the perceived quality of the study.(Detsky et al. 1992) In doing this it should be noted that although actual estimates are affected only by the relative weights used, the width of the confidence intervals is affected by the absolute weights used. (Detsky et al. 1992) To avoid this problem each study's score can be divided by the mean score, to leave the width of confidence intervals unchanged.

A further possibility is to multiply the precision of the study by its quality score, and to use this product as the weighting for each study. (Moher et al. 1998b; Berard and

Bravo, 1998) In such a way both the size and quality of the study are incorporated into the weighting calculation. This could be done using either a fixed or random effects model. (Berard and Bravo, 1998) For a fixed effect model, the new weights become,

$$w_i' = (QS_i)(w_i) \tag{2.29}$$

and similarly for the random effects model,

$$w_i' = (QS_i)(w_i^*), \tag{2.30}$$

where QS_i is the quality score allocated to the ith trial in the analysis and w_i and w_i^* are the original weights given to the ith trial in fixed and random effects analysis respectively, as defined by equation 2.2 and the reciprocal of 2.15. The pooled estimate can now be calculated as before using the new weights. The variance of the pooled estimate can be calculated using: (Berard and Bravo, 1998)

$$\text{var}(\overline{T}.) = \frac{\sum_{i=1}^{k} QS_i w_i'}{\left(\sum_{i=1}^{k} w_i'\right)^2}. \tag{2.31}$$

Caution has been expressed at incorporating or using quality score in the weighting given to studies because, while weighting study estimates by the precision has desirable statistical properties, quality scores are not direct measures of precision and this approach lacks statistical or empirical justification.(Detsky et al. 1992)

Finally, another approach is to exclude the studies of poor(est) quality altogether. This can be viewed as an extreme form of weighting - giving the poorest studies zero weight, (Light, 1987) but sensitivity analysis of the effect of cut-off on the results is essential.

There would appear to be little consensus concerning the optimum way of dealing with study quality in meta-analysis (although there is broad agreement that a quality assessment should always be carried out). There is growing support for using individual

indicators rather than an overall quality score, following a comparative assessment. (Moher et al. 1998b; Juni et al. 1999) Perhaps the best path to take is to consider the methods outlined below as part of a sensitivity analysis (see below), and assess the influence adjustment for quality has on the results of an analysis unmodified by study quality. The key problems are that: a) the influence of factors affecting validity will differ depending on the question and context of the trial; and b) studies do not uniformly report sufficient details of the methods used in their design, conduct and analysis, to allow these factors to be measured in the same way in each study. If studies of different designs are being considered, as is the case for much of this thesis, a further problem exists. For studies of different designs, the factors influencing internal validity are likely to differ and so standard scores will not be generically relevant.

While not being a central topic of this thesis, the issue of study quality is discussed further in Chapter 9.

2.8 Sensitivity analyses

Sensitivity analysis provides an approach to testing the robustness of the results of a meta-analysis to key decisions and assumptions that were made in the process of conducting it. (Oxman 1996) While such a process is relatively straightforward if only a small number of specific issues are to be explored, it becomes more cumbersome if there is uncertainty regarding many aspects of the analysis. Chapter 7 re-analyses a meta-analysis where there is uncertainty regarding the data from several studies, and considers a multi-dimensional simulation approach to addressing it.

2.9 The limitations of current meta-analysis practice. Is taking a weighted average always appropriate?

The methods predominantly used at present for pooling study results are fixed (Fleiss, 1993) and random effect models.(DerSimonian and Laird, 1986) Put simply, these methods take weighted averages of the estimates obtained from each of the studies being

combined. Both types of analysis produce an overall pooled estimate, together with a confidence interval, which is considered as the "bottom line" when assessing the effectiveness/efficacy of an intervention.

Naylor (Naylor, 1989) discusses ways in which apparently similar RCTs may differ; these are summarised below.

- Differences in inclusion and exclusion criteria
- Other pertinent differences in baseline states of available patients despite identical selection criteria
- Variability in control or treatment interventions (e.g. doses, timing, and brand)
- Broader variability in management (e.g. pharmacological co-interventions, responses to intermediate outcomes including cross-overs, different settings for patient care)
- Differences in outcome measures, such as follow-up times, use of cause-specific mortality, etc.
- Variation in analysis, especially in handling withdrawals, drop-outs, and cross-overs
- Variation in quality of design and execution, with bias of imprecision in individual estimates of treatment effect.

When treatment estimates from studies that differ in such ways as these are combined, what interpretation can be given to the pooled estimate? The answer to such a question is that it probably depends on the degree to which these factors do really vary, and to what extent these factors affect the effectiveness of the intervention; but the potential for producing results with little interpretability is very real. When considering multiple sources of evidence, clearly the potential for qualitative differences between studies is much greater still.

Adjustment for the factors outlined above is possible using the meta-regression methods outlined in Section 2.4.1. Covariates can either relate to study characteristics, such as a quality score; or a summary measure of individual values, such as mean age of patients. (Lau et al. 1998)

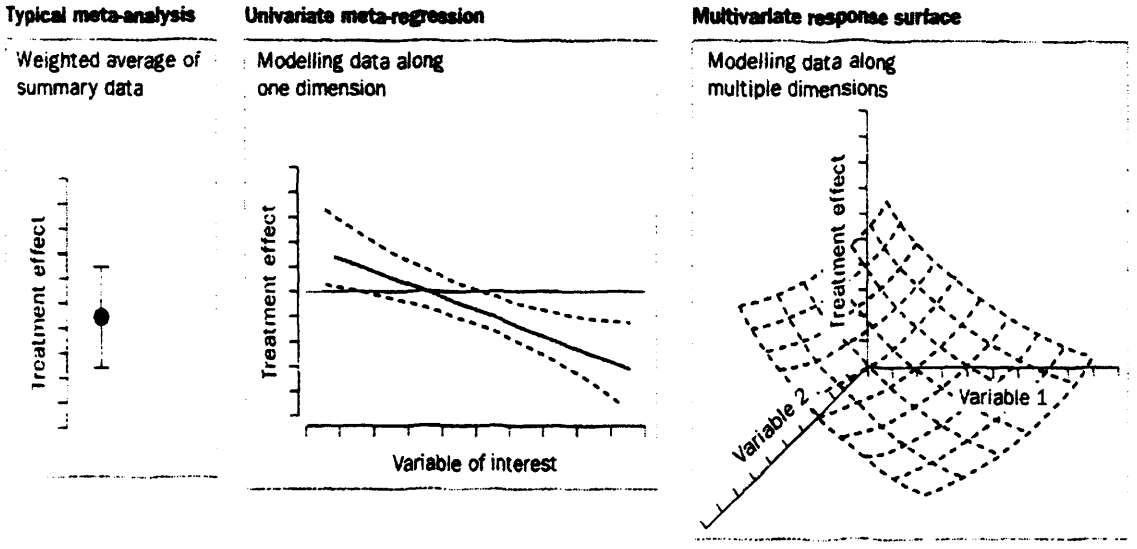
Concerned with such issues Rubin considered a "new" approach to meta-analysis which does not rely simply on study averaging, and can be viewed as a very detailed and complete multiple-(meta)-regression model. (Rubin, 1992) In this he aimed to estimate the effect of an ideal study, rather than simply take an average of those that had actually been carried out, providing insights into the underlying science, using an extension of meta-regression ideas. This method involved building and extrapolating a response surface, by considering factors related to each of the studies being combined. A distinction is made between factors of scientific interest (such as gender or age of subjects), and scientifically uninteresting design variables (e.g. the sample sizes of the studies or year the studies were completed). The response surface of interest is the effect of treatment as a function of these two kinds of factors, and it expresses the typical treatment effect as a function of these. For the design variables - these can then be set at values for a perfect study, so the answer can be expressed as a function of only the factors of interest. If one considers multiple sources of evidence, some with greater potential for bias than others then the specification (and hence adjustment for) these design variables is going to be even more crucial than if one considers only a single design type. This approach has been implemented in an econometrics setting, (Vanhonacker, 1996) but not in a health research setting. One serious drawback of the method is the requirement for a large number of studies to estimate a response surface successfully. Indeed after recent research into the power of meta-regression, (Lambert et al. 2001a) orders of magnitude more studies would be required to fit such models than are typically available in medical research.

It should be noted that the ideal study profile described by Rubin has aims in common with the original motivation for cross-design synthesis, (see Section 3.7)(Droitcour et al. 1993) where results from a theoretically ideal study, (a study which may actually be impossible to implement in practice) could be produced, in someway maximising the strengths and minimising the weaknesses of particular designs. Indeed, it has been questioned whether a meta-analysis is simply an extension of clinical trials seeking to confirm a single answer or whether it is it a unique discipline aiming to explore multiple answers? (Anello and Fleiss, 1995)

Currently, meta-analyses are primarily carried out to pool treatment effects from individual trials in an attempt to get a more accurate, and precise estimate of the true average effect of the intervention. All inferences regarding such an analysis pertain to the effect size typically seen in an 'average' subject in the trials. In some, possibly many, situations this may be adequate, the effect of the intervention on individuals may not vary predictably to any great degree. However, in some situations, there may be differential treatment effects across patients with different characteristics. For example, there may be a certain amount of risk involved due to a side effect of a treatment. Identifying patients for whom the potential benefit of the treatment outweighs the potential harm, possibly through calculating their risk for the disease, would be desirable. (Schmid et al. 1998) This aim is considered further in Chapter 8. Additionally, with ever growing pressure to demonstrate cost-effectiveness, when new expensive drugs show only moderate benefit over cheaper existing treatments, knowledge of subgroups of patients in which the marginal benefit is greatest is desirable. Indeed it has been stated that knowing how best to treat the individual should be the ultimate goal of both clinical trials and meta-analyses. (Lau et al. 1998)

Meta-regression and the ideas of Rubin, data availability permitting, provide one way to investigate intervention effects beyond the overall mean. This idea is illustrated clearly in Figure 2.1, reproduced from Lau et al. (Lau et al. 1998) who also question whether meta-analysis should be producing a single or multiple answers.

Figure 2.1 Summing-up evidence in single and multiple dimensions. ***Reprinted from Lau et al. (Lau et al. 1998) with permission from Elsevier Science.***



The first panel of this figure shows the result obtained from a simple fixed or random effect meta-analysis; panel two a regression line showing how the mean treatment effect varies with respect to a single covariate of interest; and panel three a response surface, in this case with an effect response surface described for two (continuous) covariates. If the two covariates represent patient characteristics, say for example age and blood pressure, then a treatment estimate for any combination of those values can be calculated. Clearly it is possible to extend this to many covariates, data permitting.

Lau et al (Lau et al. 1998) went on to state that

> "Large trials, while more precise than smaller trials, may miss important treatment variation and may not be any more generalisable than smaller studies unless their inclusion criteria and recruitment capture broad populations and different settings."

This idea of extrapolation to specific patients, or individualisation of treatment regimes clearly is often difficult to establish in practice. Mega-trials with tens of thousands of patients are set up just obtain an accurate estimate of the average intervention effect for

what are usually modest treatment benefits; what hope is there of ever saying, for a patient with characteristics x_1-x_n, suffering from y, drug z at dose d is the best treatment?

Subgroup analyses in trials are carried out, with the intention of investigating the influence of factors other than treatment factors on the treatment effect, (Schneider, 1989) and indeed with the advent of multi-centre trials these do have considerable numbers of patients in them, however they very often still lack power to detect differences in treatment effects and there are issues with over-interpretation of patient subgroups.

Meta-analysis has the advantage of increased power over individual trials for finding differences in effectiveness in subgroups of patients.(Yusuf et al. 1991) However, if summary data are being used (i.e. which it is in the majority of meta-analyses), covariates are limited to those aggregated at the patient level (i.e. if patient age is being investigated, only the average age for patients in the study are available and data on individual ages of patients are not available). Unfortunately, it has recently been demonstrated that meta-regression using aggregated patient level characteristics has much reduced power compared to individual patient data analyses.(Lambert et al. 2001a) Such a finding suggests that extension of meta-regression analyses of summary data, as suggested above, will not be feasible and individual patient data will be required to fit models with multiple covariates. (Whitehead et al. 2001; Higgins et al. 2001) The issue of obtaining and analysing individual patient data is not considered in detail in this thesis.

2.10 Limitations of meta-analysing solely randomised controlled trial data

RCTs have been established as the gold standard study design for establishing the efficacy of a new medical intervention. This is primarily due to the fact that such a design minimises the potential of biases which could lead to misleading results being

produced. However, this does not imply that RCTs answer all questions pertaining to the use of an intervention in every context.

Section 1.3 considered instances where the observational evidence would be desirable in addition to, or instead of, the randomised evidence. One of the issues raised there was the potential lack of generalisability of RCTs due to restrictive inclusion criteria. For example, is it wise to extrapolate results to the treatment of very elderly persons, if the elderly had been excluded from all the RCTs? A related issue is that, while an RCT may produce a good estimate of the efficacy of an intervention, its effectiveness, or effect in routine practice may be somewhat different. The difference may be due to factors such as more intense care being provided in a trial etc. A side issue is that it could be argued that the effect of care may vary with the size of a trial, smaller trials being more artificial and mega-trials not deviating as noticeably from routine practice. This could be one reason why recently mega-trials produced different results than previous smaller trials. (Anonymous, 1995b)

2.11 Summary

This chapter has reviewed all the methods commonly used currently to perform meta-analysis. While these methods are largely geared to estimating "average" effects on "average" patients, methods are less developed for the ultimately desirable aim of predicting treatment effects for individual patients. Although methods which aim to do this have been outlined, they are currently hampered due to restrictions on the quantity of data available, especially randomised data. Indeed, the sheer amount of data required to identify the treatment interactions necessary to begin to individualise treatment regimes means examining observational data may be the only feasible way to proceed towards this end in general. Hence, on these grounds a strong case for considering evidence from studies other than RCTs in a synthesis examining treatment effectiveness can be made. The crucial issue that remains is how can these extra sources of evidence be advantageously included in a synthesis? The next chapter reviews methods that have been developed to do so.

Chapter 3 Critical review of methods for synthesising disparate sources of evidence

3.1 Introduction

Although currently either meta-analyses are limited to including data from a single study type, or data from different study types are combined ignoring the different sources, several methods have been developed specifically to synthesise disparate sources of information. These are briefly outlined and discussed below. Broadly, as this chapter progresses, the methods described consider the synthesis of increasingly disparate sources of evidence.

3.2 Combining different randomised designs

Perhaps the most logical place to start this chapter is to consider the synthesis of results from the common randomised designs – the single period standard RCT, the crossover trial and the cluster-randomised trial. No new methods are required to pool crossover trials, provided there is no treatment carry-over effect in the second period. If there is, then the second period results should be excluded from the analysis. (Fortin et al. 1995) Similarly, provided a cluster randomised trial has been analysed correctly, taking into account the inter-class correlation, then these can also be combined with standard RCTs using regular methods outlined in Chapter 2.

3.3 Combining matched and unmatched data

Duffy et al. (Duffy et al. 1989) present a method for combining matched and unmatched data from RCTs, though the same methodology is directly applicable to case-control studies, including the situation of several controls per case. The motivating example for this methodology was to provide an estimate for the effect of photocoagulation on the rate of visual deterioration. In this instance a matched study was one where each patient had one of their eyes (selected at random) treated while the other one remains untreated. The methodology is an extension of the Mantel-Haenszel

procedure (Mantel and Haenszel, 1959; Sutton et al. 1998). To combine the results from the matched studies, each matched pair within a study is treated as a stratum. By doing this, stratification by study is performed automatically.

Moreno et al. (Moreno et al. 1996) describe the use of (logistic) regression methods for combining matched and unmatched case-control studies, using individual patient data. The logistic regression model proposed combines conditional logistic regression likelihood function for the matched cases and controls and an unconditional logistic regression likelihood function for the unmatched study.

3.4 Combining studies with historical controls (single-arm studies)

Begg and Pilote (Begg and Pilote, 1991) present a random effects model to estimate an overall treatment effect when some comparative studies (e.g. RCTs) are to be combined with non-comparative, historically controlled "single arm" studies, (Section 1.2) where the historical controls are ignored.

The model differs from the standard random effects meta-analysis model for two reasons. Firstly, an estimate for the event rate for each study arm is found (rather than a difference or a ratio between treatments). Secondly, in this model the treatment effects are fixed but a random effect baseline term is included. In this way, uncontrolled (non-comparative) studies can now be included in the analysis, thus creating the potential to combine extra information compared with the standard model, which may be of particular value when a dominant proportion of information on a treatment exists from the uncontrolled studies. A test for systematic bias in the uncontrolled studies and an extension to include a random effects term for the treatment effect as well as the baseline effect are also discussed. Li and Begg provide an extension to this (Li and Begg, 1994), by presenting a more general theory removing the need for distributional assumptions and they use empirical Bayes estimators for the variance terms. A fully Bayesian formulation of this model has also been recently applied; (Stevenson, 1998) this model is outlined below.

Consider two treatments to be compared, where there are n comparative trials, yielding data summaries (x_i, y_i), $i = 1, \ldots, n$, where x_i is the observed effect of treatment A and y_i is the observed effect of treatment B. Additionally, there are k uncontrolled studies of treatment A with observed effects u_i, and m uncontrolled studies of treatment B with observed effects v_i. All outcomes are assumed to be normally distributed with known variances. Hence, algebraically:

$$
\begin{aligned}
x_i &\sim N\left(\theta_i, s_i^2(x)\right) && i = 1, \ldots, n \\
y_i &\sim N\left(\theta_i + \delta, s_i^2(y)\right) && i = 1, \ldots, n \\
u_i &\sim N\left(\theta_i, s_i^2(u)\right) && i = (n+1), \ldots, (n+k) \\
v_i &\sim N\left(\theta_i + \delta, s_i^2(v)\right) && \mathrm{i} = (n+k+1), \ldots, (n+k+m) \\
\theta_i &\sim N\left(\mu, \sigma^2\right) && \mu \sim [-,-] \\
\delta &\sim [-,-] && \sigma^2 \sim [-,-]
\end{aligned}
\qquad (3.1)
$$

where θ_i is the true effect of treatment A in study i, and δ is the additional benefit of treatment B over treatment A (assumed constant across studies). An extension which allows the treatment effect, δ, to vary between studies is possible.

An important, and desirable, aspect of this model is that the relative contribution of the uncontrolled studies is directly related to the degree of homogeneity between the studies, as evidenced by the closeness of the estimated baseline effects (i.e. the uncontrolled studies are given more weight the more homogeneous the results are). Models combining multiple and single arm studies are considered further in Chapter 5.

3.5 Combining studies containing multiple and/or arms administered different interventions

Different comparisons are made in the different RCTs investigating similar questions as a result of allocation of different treatments to various arms in each study. For example, suppose two new treatments (Treatment A and Treatment B) are developed at similar times. Initially, trials comparing each treatment to placebo (Treatment C) are carried out. Hence trials of Treatment A v Treatment C, and Treatment B v Treatment C exist. Later, in an attempt to establish which of the new treatments is superior, trials of Treatment A v Treatment B are carried out. In a meta-analyses of Treatment A, both Treatment A v Treatment C and Treatment A v Treatment B trials provide information, but taking account of the different comparisons being made is essential in combining the two types of trial. Further, trials with three arms (Treatment A, Treatment B and Treatment C) may also exist.

Several extensions of the standard fixed and random effect meta-analysis models have been described to deal with these situations. These include the approach of Gleser and Olkin (Gleser and Olkin, 1994) who describe a model capable of combining studies in which more than one type of treatment has been compared to a control group. Hence, it could combine studies comparing Treatment A v Treatment C and Treatment B v Treatment C, but not Treatment A v Treatment B, because no control group (C) is included in such studies.

Berkey et al. have developed models where multiple comparisons (e.g. A vs B and A vs C etc.) can be combined simultaneously. (Berkey et al. 1996; Berkey et al. 1998) This framework is also capable of combining studies comparing multiple treatments where any study may consider only a subset of treatments. A Bayesian formulation of this model is developed in Section 5.3.

Perhaps, most flexible of all is the model of DuMouchel, (DuMouchel, 1998) which allows many extensions of the standard random effect model and mixed models, including the combination of studies with heterogeneous designs. It has the capacity to model multiple outcomes from trials, provided they are all reported on the same scale.

Different study designs are accounted for because subjects receiving different treatment regimes are treated distinctly, allowing variable numbers of study arms and even multiple treatment cross-over combinations to be modelled appropriately. So, for example, if one is considering a binary outcome, the log odds in each group could be modelled (for each treatment if cross-over combinations exist), rather than a comparative effect such as the log odds ratio; this is conceptually very similar to the modelling approach of Begg et al. (Section 3.4) which allows the inclusion of single arm trials in a meta-analysis. By constructing contrasts of parameters in the model comparative effects can be calculated after parameters in the model have been estimated. This method has a drawback, which is not explicitly stated, that has to be weighed against the added flexibility the approach allows. Namely, it "breaks" the randomisation of patients since it models each group of patients from each study individually (although random study effect terms are included, hence some acknowledgement of results from the same study being 'linked' is made). This model is explained in greater detail in Chapter 5, where it is implemented using a Bayesian formulation.

A model that does preserve the randomisation, allowing direct comparison of two active treatments when they are both only compared in trials to a control/placebo is described by Bucher et al. (Bucher et al. 1997) This extends only a fixed effect model however, and does not allow combination of trials with more than two arms. A Bayesian approach that does allow trials with three arms and evaluates all two-way comparisons of three treatments, (A, B and C) has been described by Higgins and is outlined below. (Higgins and Whitehead, 1996)

$$T_{ABi} \sim N(\theta_{ABi}, v_{ABi}) \qquad T_{ACi} \sim N(\theta_{ACi}, v_{ACi})$$

$$T_{BCi} \sim N(\theta_{BCi}, v_{BCi}) \qquad \begin{pmatrix} T_{ABi} \\ T_{ACi} \end{pmatrix} \sim N\left(\begin{pmatrix} \theta_{ABi} \\ \theta_{ACi} \end{pmatrix}, \begin{pmatrix} v_{ABi} & d \\ d & v_{ACi} \end{pmatrix}\right)$$

$$\theta_{ABi} \sim N(\mu_{AB}, \tau^2) \qquad \theta_{ACi} \sim N(\mu_{AC}, \tau^2) \qquad (3.2)$$

$$\theta_{BCi} \sim N(\mu_{BC}, \tau^2) \qquad \begin{pmatrix}\theta_{ABi}\\ \theta_{ACi}\end{pmatrix} \sim N\left(\begin{pmatrix}\mu_{AB}\\ \mu_{AC}\end{pmatrix}, \begin{pmatrix}\tau^2 & \tau^2/2\\ \tau^2/2 & \tau^2\end{pmatrix}\right)$$

$$\mu_{BC} = \mu_{AC} - \mu_{AB}$$

$$\mu_{AB} \sim [-,-] \qquad \mu_{Ac} \sim [-,-]$$

$$\tau^2 \sim [-,-]$$

This model can be considered an extension of equation 2.17 (the random effect model for Normally distributed data), and the notation used here is a simple extension of that for 2.17, where the suffixed two letters indicate the two-way comparisons. The covariance between the treatment effects, *d*, is assumed known. Note, this model does require the assumption that the between study variance is the same among all studies being combined irrespective of which treatments they compare. A version of this model for combining binary data on the log odds ratio scale, which is a direct extension of equation 2.18 is given elsewhere. (Higgins, 1997)

In section 5.4 a variation of this model is fitted to a meta-analysis dataset on sulphamethoxazole-trimethoprim or dapsone/pyrimethamine as prophylaxis against Pneumocystis carinii in HIV infected patients. This data was previously analysed by Bucher et al. using their fixed effect approach.

3.6 The Confidence profile method

Eddy (Eddy, 1989) first presented the confidence profile method in 1989. It has been put forward as a general Bayesian method for assessing health technologies. It has been described as:

> "*a set of quantitative techniques for interpreting and displaying the results of individual experiments; exploring the effects of biases that affect the internal validity of experiments; adjusting experiments for*

factors that affect their comparability or applicability to specific questions (external validity); and combining evidence from multiple sources" (Eddy et al. 1992)

An analysis which combined information from RCTs, case-control studies, and simple observational studies to assess the value of mammography screening in women under the age of 50 years using this method, has been cited as one of the first examples of bringing together evidence from randomised and non randomised studies (Eddy et al. 1988). Specifically, the early results (five to seven years) of four controlled studies were combined using this method together with a long-term study (18 years) whose results were adjusted for incomplete compliance and differences in the screening protocol.

This method can accommodate the following problems which may occur when trying to synthesise evidence: multiple pieces of evidence, different experimental designs, different types of outcomes, different measures of effect, biases to internal validity, biases to comparability and external validity, indirect evidence, mixed comparisons, gaps in experimental evidence. (Eddy et al 1992) Despite often being described as a Bayesian method, it can be formulated under classical conditions where maximum likelihood estimates and covariances for the parameters in a problem can be derived. (Eddy et al 1992)

A key feature of this method is that it models biases explicitly. Hence, the result of an analysis by this method, a posterior distribution for the parameter of interest, incorporates all the uncertainty the assessor chooses to describe about any of the parameters used in the analysis. Biases that may be considered include misclassification rates, measurement error probabilities, and contamination rates. These biases may not be known precisely, but some evidence may exist from which estimates can be obtained. However the difficulty in the specification of this area of the model is a serious drawback of the method. As with more standard Bayesian analyses it is possible to incorporate subjective judgements into the model in a structured way via prior distributions, though this is not a requirement of the method.

Hasselblad and Mc Crory (Hasselblad and McCrory, 1995) claim that the method eliminates the need for sensitivity analysis, stating that every parameter, if properly

modelled, contains all the information about biases and uncertainty. Thus the final answer includes uncertainty about all of the parameters in the model. This is a curious statement as it is generally advocated that the robustness of parameter estimates should be tested over a range of prior distributions, especially since certain parameters including estimates of bias will not be known precisely.

A piece of software, which is available commercially, called FAST*PRO (Eddy and Hasselblad, 1992) has been developed to carry out the confidence profile analysis, but mistakes in this have been noted. (Egger et al. 1998) This is a brief overview of a method which is complex (Eddy et al 1992; Eddy, 1989; Eddy et al. 1990b; Eddy et al. 1990a) and has not gained widespread acceptance, only being used occasionally. This can be partly attributed to the fact that when it appeared the approach was vastly different from any other mainstream methods. The problems with model specification, alluded to above, are also certainly contributing factors. However, it is probably true that the method was "ahead of its time". Recently, the approach was re-visited (Spiegelhalter et al. 2000b) using the WinBUGS software (Speigelhalter et al 2000a) which uses a similar graphical modelling framework, and to which it translated very naturally. Implementation in this WinBUGS software has several advantages over FAST*PRO including relaxing of some of the modelling assumptions including symmetry of posterior distributions. (Speigelhalter et al. 2000b) The method has also been cited as an influencing factor in the cross-design synthesis approach described below.

Although the Confidence Profile method is not considered further explicitly in this thesis, many of the models developed are in the same spirit, and although the graphical representation of models are not given in favour of algebraic specifications, equivalent representations could be produced. In particular, the net benefit model described in Chapter 8 is very close to the modelling ideas first expressed under the Confidence Profile framework with the notable omission of bias parameters in favour of a sensitivity analysis approach.

3.7 Cross-design synthesis

The initial work on cross-design synthesis was carried out by the Program Evaluation and Methodology Division of the U.S. General Accounting Office (GAO). (General Accounting Office., 1992) Their idea was to create methodology for a new form of meta-analysis that aims at capturing the strengths of multiple-study designs while minimising their weaknesses. Its purpose is to provide better answers to difficult research questions, (Droitcour et al. 1993) and, "relates to improving cost-effectiveness of health services, against a backdrop of concern about perceived spiralling health care expenditure." (Lancet, 1992) Its motivation comes from many of the ideas discussed previously in this thesis, including: 1) the low generalisability of RCTs; 2) limitations of meta-analysis from moving from judging whether a treatment is, in principle, efficacious, to deciding how to manage a particular patient; and 3) the questionable reliability of subgroup analyses of (individual) RCTs.(Lancet, 1992)

Droitcour et al. (Droitcour et al. 1993) state that definitive answers about the effects of various treatments in medical practice can be provided only by a body of research that meets two key criteria: (a) scientific rigour in comparing treatment outcomes and (b) generalisability to the conditions of medical practice. Randomised controlled trials are designed to provide unbiased comparisons of outcomes following treatment, but often fall short of meeting the generalisability criterion. Conversely, statistical analyses of databases are uniquely suited to covering outcomes across the full range of patients, but they rarely provide convincing evidence of unbiased comparison (Droitcour et al. 1993). Thus, most RCTs and most database analyses probably fail to meet at least one of the two criteria for providing valid answers to questions about a treatment's effect in medical practice. However, if the strengths of complimentary study designs can be combined both criteria can be met. It is interesting to note that Droitcour et al. (General Accounting Office., 1992) comment that although their work reviews methods for assessing, adjusting, and combining study results, its greatest emphasis is placed on methods for assessing study weaknesses.

Droitcour et al. (Droitcour et al. 1993) acknowledge the previous work of Rubin (Rubin, 1992) (see section 2.9) and Eddy (Eddy et al. 1992; Eddy et al. 1990b; Eddy et al. 1990a; Eddy, 1989) (see section 3.6) who separately explored ways of synthesising results from studies with a diversity of designs. They say cross-design synthesis builds on these directions but with two key differences: 1) it focuses on combining results from studies with complementary designs; 2) it uses a two-pronged approach to study assessment. The first prong consists of an overall quality assessment of each study. The second prong is a focused assessment of the potential biases that derive from the primary weakness(es) inherent in a study's design. This second prong is the heart of the strategy of cross-design synthesis; its findings are used only to a) adjust the results of an individual study, and b) identify each study's most appropriate contribution to a synthesis model.

As mentioned above, RCTs and database analyses have complimentary strengths, but one cannot assume that in combining their study results, their strengths will be preserved while their weaknesses counteract each other. For this reason Droitcour et al. (Droitcour et al. 1993) devised the following three stage strategy for minimising weaknesses of study designs:

- Focused assessment of the key study biases that may derive from characteristic design weaknesses to provide the information needed to compensate for specific weaknesses (Droitcour et al. 1993; General Accounting Office., 1992)

- Individual adjustment of each study's results to "correct for" identified biases

- Development of a synthesis framework and an appropriate model for combining results (within and across designs) in light of all assessment information

Droitcour et al. (Droitcour et al. 1993) comment that despite secondary adjustments, there is a possibility that the weaknesses of each design may continue to bias study results. This may be because some patient groups may have been totally excluded from randomised studies, which is a problem that cannot be fixed by standardising individual studies' results to correct for over- or under representation. Similarly, focused

assessment of a database analysis may not detect every imbalance in the comparison groups.

The solution put forward to this problem is to devise a framework for organising, analysing, and combining results from different categories of study designs. Although this general approach has its roots in meta-analysis, the framework derives directly from the work of Hlatky. (Hlatky, 1991)

Once this has been done the investigator must decide whether to: a) present results from each stratum separately; b) present only estimates from certain strata (e.g., strata that contain only those studies deemed to be of high quality); or c) combine estimates across strata using adaptations of the various methods of meta-analysis.

Droitcour et al. (Droitcour et al. 1993) point out three major strengths of cross-design synthesis, namely: 1) It can draw upon different kinds of studies that, in combination, can tell more about how medical treatments work than any single type of study can; 2) It can be applied to existing results in several areas because diverse study designs are increasingly being used to evaluate treatment effectiveness; and 3) It has the ability to produce the generalisable information needed to support credible medical practice guidelines.

A limitation, the authors point out, is the necessity of relying on investigator judgement for many decisions. Until refinements of this strategy are developed, GAO believes it is best applied by those knowledgeable about both a specific medical treatment and evaluation methods in general. (Chelimsky et al. 1993) This is a crucial point; while the method is conceptually sensible and appealing, the assessments and adjustments required will be difficult to make with the instruments currently developed for such purposes. This echoes the concerns made about dealing with study quality in a generalised synthesis framework discussed in Chapter 1, section 2.7 and Chapter 9; a problem that remains largely unresolved in this thesis.

An anonymous editorial in the Lancet was cautious about this new methodology arguing risk with cross design synthesis is that the more expensive, time-consuming, and reliable component - RCTs - will increasingly be replaced by database analyses.

(Lancet, 1992) Chelimsky et al. (Chelimsky et al. 1993) disagreed with this, commenting that RCTs were a necessary part of cross-design synthesis.

3.8 Bayesian hierarchical models

Several authors have considered the flexible approach offered by the use of hierarchical models to combine data; these are considered below and utilised throughout this thesis.

3.8.1 Bayesian three-level hierarchical model for the general synthesis of evidence

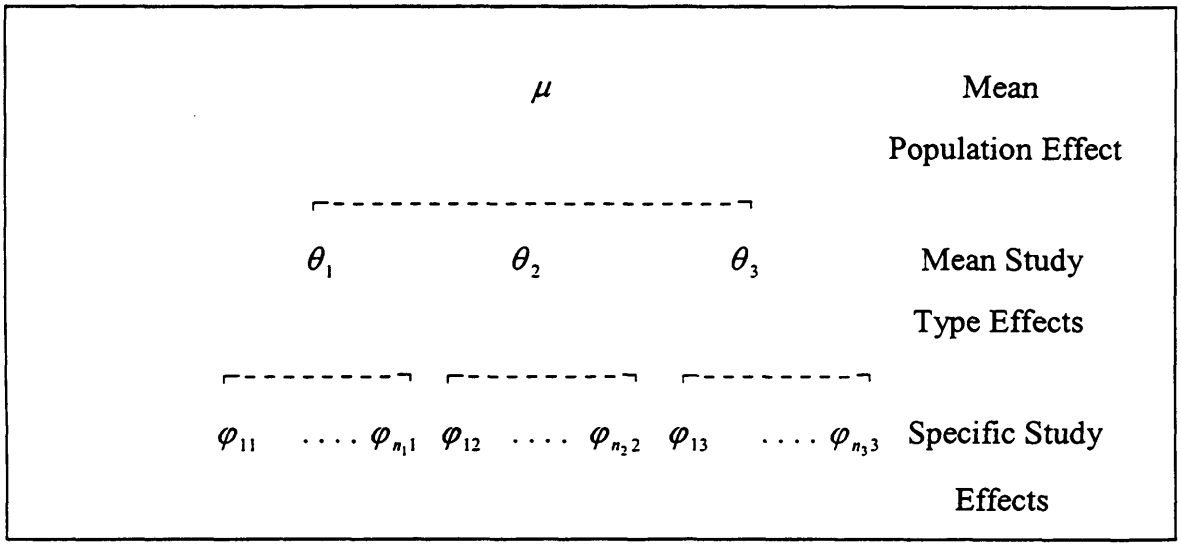
Prevost et al. (Prevost et al. 2000) developed a model to include studies with disparate designs into a single synthesis, acknowledging that when considering the effectiveness of an intervention evidence from non-randomised studies may be relevant in addition to the RCTs. Although this work follows in the cross design synthesis spirit of Droitcour et al. (General Accounting Office., 1992), the methodology used is somewhat different and is more specifically operationalised. More generally, this approach presents a framework for sensitivity analysis to include evidence from different study types.

The hierarchical nature of the model specifically allows for the quantitative within and between sources heterogeneity, whilst the Bayesian approach can accommodate a priori beliefs regarding qualitative differences between the various sources of evidence. These prior distributions may represent subjective beliefs elicited from experts, or other data-based evidence, which though pertinent to the issue in question is not of a form that can be directly incorporated, such as data from animal experiments. (Abrams et al. 1997) It is important to note however, that the specification of informative priors is far from automatic. See Sutton and Abrams (Sutton and Abrams, 2001) for an example of this approach in practice. (A further alternative is to use non-randomised studies to derive a prior for the pooled treatment effect for a meta-analysis of RCTs, this is also pursued elsewhere (Sutton and Abrams, 2001))

This model can be viewed as an extension of the standard random effects model for meta-analysis, (DerSimonian and Laird, 1986) but with an extra level of variation to

allow for variability in effect sizes between different sources of evidence. Figure 3.1 outlines the generic structure of the model. μ is the overall pooled effect across all sources of evidence, the θis are the pooled effects within study types (i = 1,2, . . I), where study types may indicate RCTs, cohort studies, case-control studies etc., and φ_{jis} (i = 1,2, . I; j = 1, . . . , n_i) are the individual study-specific estimates. Extensions including incorporation of prior constraints, prior beliefs, and study level covariates are also feasible.

Figure 3.1 3 level model for synthesising studies of different designs



The full model specification is as follows

$$
\begin{aligned}
&T_{ij} \sim \mathrm{N}[\varphi_{ij}, v_{ij}], \qquad (i = 1, \ldots . I; j = 1, \ldots, n_i) \\
&\varphi_{ij} = \theta_i + \upsilon_i z_{ij}, \qquad z_{ij} \sim \mathrm{N}(0,1), \\
&\theta_i = \mu + \tau \varepsilon_i, \qquad \varepsilon_i \sim \mathrm{N}(0,1), \\
&\upsilon_i \sim [-,-], \qquad \tau \sim [-,-], \qquad \mu \sim [-,-],
\end{aligned}
\tag{3.3}
$$

where φ_{ij} is the true intervention effect in the jth study of type i, θ_i is the true effect in the ith type of study and v_i^2 is the variance between studies of type i. μ is the overall mean effect of the populations and τ^2 represents the between study type variance. The

random effects, z_{ij}, reflect the differences in the true intervention effect in the individual studies from the overall study type effect. The random variable, ε_i, allows variation of the mean effect size between study types. This model is further considered, developed and applied to the evidence on cholesterol levels and mortality in Chapter 5.

3.8.2 Grouped random effects models for Bayesian meta-analysis

Larose and Dey (Larose and Dey, 1997a) observe that in meta-analysis results from dissimilar studies are often inappropriately combined. They present a method which addresses this problem using a Bayesian hierarchical model framework, similar to that of Prevost et al. (Prevost et al. 2000), although no third level is included in their model. Their illustrative example considers 15 comparative studies of progabide, a new anti-epileptic drug, all of which use a crossover design. A distinction is made between double-blind (closed) studies and those in which the investigator had knowledge of the treatment regime (open studies). A random effect model is used which calculates an overall mean plus a group-specific random effect; in the example this is for the open and closed studies though other design feature could be dealt with in this way. This model, which considers the proportion of patients who improve on the active treatment as outcome, is outlined below

$$\begin{aligned} r_{ij} &\sim Bin[p_{ij}, n_{ij}] \qquad i = 1, \ldots\ldots, I \quad j = 1, \ldots, k_i \\ \log it(p_{ij}) &= \mu + \lambda_{ij} \qquad \lambda_{ij} \sim \mathrm{N}[\theta_i, \tau_i^2] \end{aligned} \tag{3.4}$$

where j indexes the number of studies in each of the i distinct categories of study (e.g. i = 2 for open and closed studies in the original example). The effect in the ith study grouping is estimated by $(\mu + \theta_i)$ with between study variance τ_i^2. Hence, each group's effect size estimate has a specific between-study variance term. It is assumed that there is exchangeability between studies within each group. The authors' main purpose was not to produce a single overall pooled estimate, but to develop a framework in which heterogeneity between study types could be quantified. (i.e. using a traditional fixed or random effect model) a conclusion of no treatment effect would be made, however using the grouped random effects model demonstrates that the open studies support the

efficacy of progabide while the closed studies support the reverse hypothesis. The authors comment that the model could easily incorporate covariate information through the use of linear structure in the overall mean. (Larose and Dey, 1997b) This model is not considered in detail in this thesis, however, possible extensions are discussed in Chapter 9.

3.9 Exposure risk assessments

When considering the risk associated with the exposure to a certain agent, evidence may be available from several types of observational study (only rarely is randomised evidence available on exposures to potentially harmful substances). Although risk assessment is not a main focus of this thesis, a brief account of the methods developed in this area is given below as they provide examples of the application of generalised synthesis models..

3.9.1 Combining the results of cancer studies in humans and other species

DuMouchel and Harris (DuMouchel and Harris, 1983) propose a class of Bayesian statistical methods for interspecies extrapolation of dose-response functions. In their original analysis, their motivation for considering information from non-human species stemmed from the fact that there was an abundance of precise data available from animals concerning the assessment of cancer risks from environmental agents, but little accurate information on direct effects in humans. A formal distinction is made between conventional measurement error within each dose-response experiment and a novel error of uncertain relevance between experiments. Dose-response data from many substances and species is used to estimate the inter-experimental error. From the data the estimated error of interspecies extrapolation, and prior biological information on the relations between species or between substances, posterior densities of human dose-response are calculated.

3.9.2 Stratified ordinal regression: a tool for combining information from disparate toxicological studies

Cox and Piegorsch discuss the development of methodology for combining studies on acute inhalation assessment. (Cox and Piegorsch, 1994; Carroll et al. 1994) Their goal was to develop methodology for data combination that incorporates the range of endpoint severity, exposure concentrations, and exposure durations. The method is based on severity modelling, wherein concentration, duration, and response are integrated to determine potential risks to humans after acute inhalation exposure to some environmental toxin.

3.9.3 Combining epidemiological and biochemical evidence

Tweedie and Mengersen (Tweedie and Mengersen, 1992) investigate the relationship between lung cancer and passive smoking. Previously, two approaches had been taken for investigating this: 1) the biochemical approach, using cotinine in the main as a marker; and 2) the epidemiological approach. The paper uses both sorts of studies in one meta-analysis. The authors comment on using the now-standard 'Wald adjustment' (Wald et al. 1986) for differential misclassification, this estimates the effect of differential bias introduced by the misclassification of smokers and non-smokers.

3.9.4 Combining sparse outcomes from observational studies with different designs

Epidemiological studies often examine the risks related to rare outcomes. If cohort studies are used to examine rare outcomes, then they have to be large even to observe a small number of events. Case-control studies are often used in such situations due to greater ease of implementation. Austin et al. (Austin et al. 1997) developed a fixed effect exact procedure to combine these results across different study designs using 'exact' methods. This methodology is developed further in Section 7.9 where a Bayesian random effect model is described which would appear a superior method for combining results from studies with different designs.

3.9.5 Combining case-control and prospective studies

Müller et al (Müller et al. 1999) consider methods for combining case-control studies with prospectively collected confirmatory information concerning effects of risk factors that have been identified based on the case-control studies for the purposes of developing risk predictions. A Bayesian hierarchical model is utilised, but the method assumes individual patient data is available.

3.10 Combining heterogeneously reported outcomes

An issue which has not received much attention until recently is the need for methods to combine heterogeneously-reported information. Section 3.5 has already considered the case when different comparisons are being made within studies. A further problem exists when the results of studies are reported on scales of different types. This is a common occurrence when results studies with different designs are to be combined.

Abrams et al. (Abrams et al. 2000) consider continuous outcomes, where in some instances, the difference from baseline is reported, while in others only baseline and follow-up measures are reported. For the latter situation the correlation is often not known, and hence the standard deviation for the mean change cannot be calculated and hence included in a meta analysis of difference. A Bayesian meta-analysis model is used which places an informative prior derived from pertinent background information on the correlation between baseline and follow up measurements in the individual studies where no difference estimate is given allowing a synthesis of all studies, irrespective of reporting method, to proceed.

Further methodology has been developed by Dominici et al. (Dominici et al. 1999), who combine information on various treatments for headaches. In this situation outcomes were reported as continuous treatment effects for individual treatments, as differences between treatments, and as 2 x 2 contingency tables for dichotomised responses in the different studies. A hierarchical Bayesian grouped random-effects

model is described which introduces latent auxiliary variables to create a common scale for combining the information.

3.11 Net benefit: Generalising RCT results using additional information from observational studies

Glasziou and Irwig (Glasziou and Irwig, 1995) consider generalising randomised trial results using additional information. They employ 'Net Benefit' as defined by:

$$\text{Net Benefit} = (\text{Risk Level} \times \text{Risk Reduction}) - \text{Harm} \qquad (3.5)$$

This model suggests potential benefit increases with risk, but that harm will remain relatively fixed. Thus at low levels of risk, the benefits will not outweigh the harm and we should refrain from intervening, but at higher levels, the benefit will outweigh the harm. Estimating all elements on the right hand side of equation (3.5) for population subgroups generally requires several sources of data. The authors suggest that the estimate of relative risk reduction should come from (a meta-analysis of) randomised trials, the adverse event rates may come from both randomised trials and other epidemiological studies; risk level will usually come from multivariate risk equations derived from large cohort studies. This model is considered in Chapter 8, where a fully stochastic model, which takes into account uncertainty from all evidence sources, is used to implement the method.

3.12 Propensity scores

Although propensity scores (Rosenbaum and Rubin 1983) are not used to combine studies of different designs, they can be used to compare and consolidate the results from RCTs and routine treatment databases. (General Accounting Office 1994) Propensity scores aim to control for all known confounding factors in an observational study, by regressing all potential cofounders on the treatment regimen actually followed. This function of confounders is each individual's propensity score (and is a

measure of the propensity of each individual to be allocated to the treatment they received), which can be used to adjust for all confounders in an analysis which compares outcomes. Patients can then be sub-classified on the basis of their scores and treatment comparisons made within these subgroups. The results of such analyses can be compared with RCTs to explore whether similar effects are evident in practice from the trial situation, and whether such effects are consistent across patient subgroups.

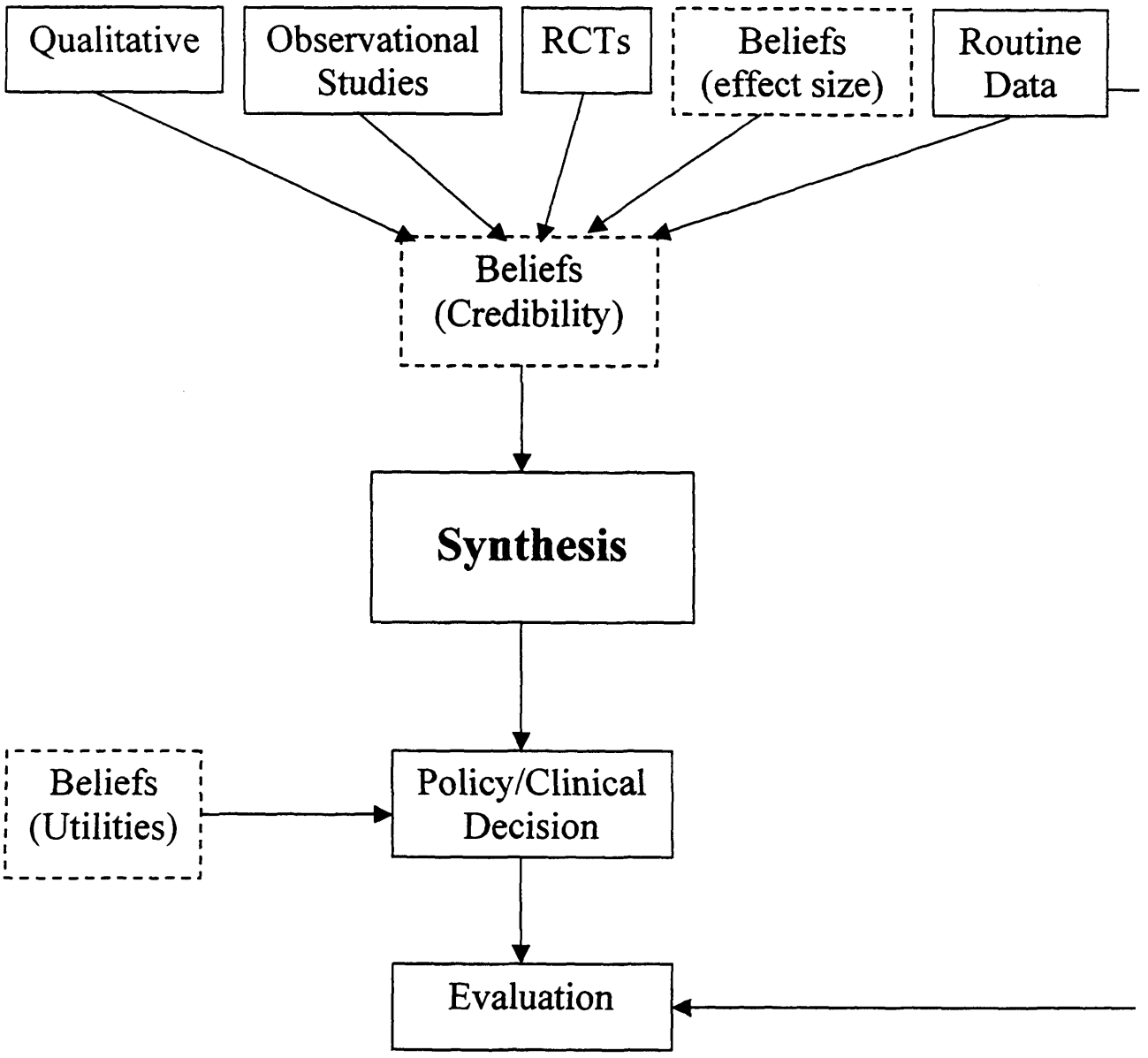
It is interesting to note that this approach has much in common with cross-design synthesis. (Section 3.7) Propensity scores are not considered further in this thesis, although their potential for use in the generalised synthesis of evidence is acknowledged.

3.13 Synthesis of evidence and the clinical/policy decision making process

Meta-analysis is only part of the process of implementing new clinical/policy procedures; decisions still have to be made once treatment effect estimates have been produced.

Figure 3.3 illustrates the process of synthesis of medical evidence. Solid boxes relate to evidence, while dashed boxes relate to peoples beliefs. The various sources of evidence are displayed across the top of the figure. These lead into the synthesis box, as do the synthesisers' beliefs about the credibility of the various forms of evidence. This box influences which evidence is included, the form of analysis, and hence the relative contribution of the different sources of evidence. Following synthesis and the production of effect estimates, policy/clinical decisions are made based on these estimates. At this stage beliefs may also be included in the decision making process, possibly including consideration of economic issues derived through cost-benefit analyses. Following this, the intervention may be recommended for routine practice and then further routine evaluations of its effectiveness are carried out which feed back into the synthesis sources.

Figure 3.3 Synthesis of medical evidence



It is important to note that currently the meta-analyst is only concerned with the synthesis stage of the model, and others then decide on policy. Midgette et al., (Midgette et al. 1994) in their assessment of the effectiveness of intravenous streptokinase on short term survival after suspected acute myocardial infarction, do provide a rare example of a combined decision analysis and meta-analysis. They only include point estimates from the meta-analysis in the decision making process, so that the uncertainty surrounding parameter values is not taken into consideration. Inclusion of the posterior distributions (assuming a Bayesian analysis has been carried out) for parameters in the decision model would be preferable. The subject of decision analysis and how it interfaces with meta-analysis is not the focus of this work but one needs to give consideration to which results are most informative to the decision making part of the process when carrying out the synthesis. Additionally, there is a need to further develop methods for synthesising economic assessment data, (Jefferson et al. 1996) which may be available from a number of sources since this is a crucial aspect of decision modelling. Finally, the whole research process, including the design and monitoring of primary studies, as well as secondary analyses such as meta-analysis and the generalised synthesis of evidence, can be viewed in a decision theoretic framework, which potentially offers a coherent structure to the whole research process. (Claxton, 1999) Such an approach can inform in which areas further research would be most informative. Further discussions of the need to include economic data in a generalised synthesis framework are included in Section 8.6.

3.14 Summary

The preceding sections provide a largely non-technical description of the various approaches to combining evidence from disparate sources. Firstly, simple extensions of the standard meta-analysis methods were considered which accommodate crossover trials, matched with unmatched data. Next, methods which allow indirect comparisons to be made, including the combination of single arm, historical control, studies with RCTs, and studies comparing different combinations of treatments.

More general methods for combining different sources of data were then discussed. The first of these, the confidence profile approach has been developed with a large number of examples, however it does have drawbacks, not least due to the necessity of explicitly defining all biases present. Cross-design synthesis considers the combination of RCT data with database data. The method is conceptually appealing, however there are a lack of practical examples. The specific use of Bayesian hierarchical models to combine distinct sources of data, which offers methods which are practical to implement are reviewed. These are used extensively in Chapters 5 and 6.

Several methods which focus on the combination of observational evidence exclusively are also included. DuMouchel's relatively early work explores the idea of extrapolation from related but not directly relevant studies, and hence considers a broader array of evidence that could be accommodated in a traditional meta-analysis. Using a Bayesian framework, shrinkage between studies is exploited. Tweedie and Mengersen's application is noteworthy for combining information from studies using two different approaches to address the same problem; the example also flags up the issue of adjustment of individual study estimates for known biases before synthesis takes place. This is conceptually similar to the adjustments for publication bias in individual study types before synthesis described in Chapter 6. Austin et al describe a logical extension of the fixed effect model to allow exact estimation in the meta-analysis of rare outcomes from observational studies with different designs.

Two methods that rely heavily on the synthesis of evidence, but use the different sources of evidence to estimate different parameters in the model are described. Glasziou and Irwig consider a wholly different approach using risk functions, for which meta-analysis provides part of the information. Their method then explores net-benefits in subgroups of patients as a way of providing treatment effect estimates for different individuals. Rosenbaum and Rubin develop propensity scores, which allow adjustment of databases, looking at effectiveness of interventions, for multiple confounders. The results of such analyses can be compared with RCTs to explore whether similar effects are evident in practice from the trial situation, and whether such effects are consistent across patient subgroups. Both these methods consider the idea of using synthesis to go beyond producing overall averages and provide an array of estimates for different

patient groups discussed at length in section 2.9. This is an appealing idea, and a recurring theme throughout this thesis.

Finally, brief consideration if given to how evidence synthesis can be used within the decision making process. There is clearly scope for decision models that utilise evidence syntheses of effectiveness data, combined with economic data. Such an approach is appealing, as the whole decision process can be made rational, transparent and objective.

Thus, many different approaches have been suggested for the generalised synthesis of evidence. Clearly much groundwork has been laid, and ideas described, but practical applications of such methods are relatively scarce. (Sutton et al. 1998) One possible reason why such methods have been slow to develop is because the potential permutations of the types of evidence available, and the specific problems encountered are huge. By definition, such analyses are going to be larger and more complex than traditional meta-analyses. If a general methodology were to be produced, it would have to be very broad to accommodate all the different sorts of evidence and all their associated problems, yet it would also have to be detailed to make practical implementation possible. The development of a single framework in which all these methods could contribute is an appealing idea, however far beyond the scope of this thesis and probably several years away if, in fact, it is a desirable or achievable target at all. However, the thesis does provide extensions to many of the methods described above, as well as to the more standard meta-analysis methods described in the previous section.

Chapter 4 Combining randomised and observational studies: the example of cholesterol lowering and risk of coronary heart disease and all cause mortality

4.1 Introduction

The relationship between total serum cholesterol and risk of coronary heart disease (CHD) or all cause mortality was chosen as the first topic to explore using generalised synthesis of evidence. Several methods which synthesise observational and randomised data are reported in Chapter 5, however, initially, individual meta-analyses of individual data types are described in this chapter. Such analyses provide the reader with the necessary background for the generalised synthesis.

The topic of cholesterol lowering was chosen for several reasons. Firstly, it was apparent, through reading previous reviews (see sections 4.3 and 4.4), that the potential literature was substantial, with evidence being available from several different types of study. Additionally, it has been a topic of much interest due to the potentially high clinical impact measures to reduce CHD have, and some controversy (see below and section 4.4), which invites an analysis using novel methods.

Using the results from observational studies, it has been known for some time that people with low serum cholesterol levels have lower incidences of CHD than people with elevated levels.(Calvert, 1994) Cholesterol lowering interventions were first suggested in the 1950s as providing potential benefit by reducing the risk of death/further cardiovascular event in patients surviving one such event. Since then a large number of trials have been carried out investigating the effect of cholesterol lowering and future risk of several major outcomes such as CHD events and total mortality, using a variety of drug, diet and even surgery interventions. Some of the earlier secondary intervention trials (i.e. in patients who already had CHD) found reduced risks of CHD events in the study arms where cholesterol had been lowered, but overall mortality in many of these studies showed no similar reduction. (Muldoon et al. 1990) Various theories have been put forward to explain this inconsistency. It has been

noted that in several trials that risk of suicides and violent deaths has been inflated in patients receiving cholesterol lowering treatments, leading to the suggestion that lowering cholesterol may alter moods and behavioural patterns. (Muldoon et al. 1990; Cummings and Psaty, 1994) There has also been the suggestion that interventions which lower cholesterol may modestly increase the risk of certain cancers, although this is still under debate (Macmahon, 1994; Davey Smith and Pekkenen, 1992) Recently, however, with the introduction of newer drugs, called statins, which are capable of lowering cholesterol levels by greater amounts than could be achieved previously, considerable reductions in the risk of CHD mortality demonstrate benefits which almost certainly outweigh the risks as a secondary intervention in persons with elevated cholesterol levels. Less clear is the benefit of cholesterol lowering as a primary intervention when considering overall mortality, (Anonymous1998) and for people who already have moderate or low cholesterol levels. (Rubins, 1995)

The issues of identifying: 1) which people would benefit from a cholesterol lowering intervention and 2) the optimal intervention strategy for such persons identified are addressed in the analysis which follows.

4.2 Literature identification methods

Originally, this analysis was conducted in 1998 when I started work on this thesis, however, it was updated in 2001 and hence the evidence it contains should be up-to-date as of the end of 2000.

After brief pilot searches, it became apparent that many meta-analyses of the cholesterol lowering RCTs had been carried out previously; these are reviewed in section 4.4. Since many of these cross-referenced each other, it was relatively straightforward to identify what is suspected to be the majority of the previous meta-analyses.

Through scrutinising each of the reference lists from these meta-analyses, RCTs investigating cholesterol lowering interventions were identified. Since the randomised evidence in the area of cholesterol lowering appeared to be so well documented, it was

decided to rely on these previous meta-analysis as the source of evidence to compile a comprehensive list of RCTs by. In addition, Medline was searched for any recent trials reported after the meta-analyses had been carried out (1997- 2000). As a final check, a researcher with an active interest in the area and author of a previous meta-analysis, [1] was contacted to enquire of any additional trials, of which he knew none. It is acknowledged this is a rather non-standard approach to literature identification, however, with over 40 meta-analyses carried out previously (see section 4.4), it would appear that cholesterol lowering RCTs is an exceptionally well documented area of medical research, and hence for the purposes of this thesis such practice could be justified.

Scoping searches were carried out to assess the breadth of the potentially relevant non-RCT evidence. Figure 4.1 documents the different types of evidence found. These range from non-randomised trials, through aetiological observational studies, to experimental studies on animals.

These scoping searches identified a particular paper including a meta-analysis of cholesterol lowering RCTs and aetiological observational studies relating to cholesterol levels and rates of CHD (Law et al. 1994a) The paper did not qualitatively combine observational with randomised studies, however it did compare the meta-analyses of each type of evidence. This paper stated that over 60 cohort studies investigating the relationship between cholesterol level and adverse outcomes existed (in 1994), however, it only considered, and referenced, the ten largest.

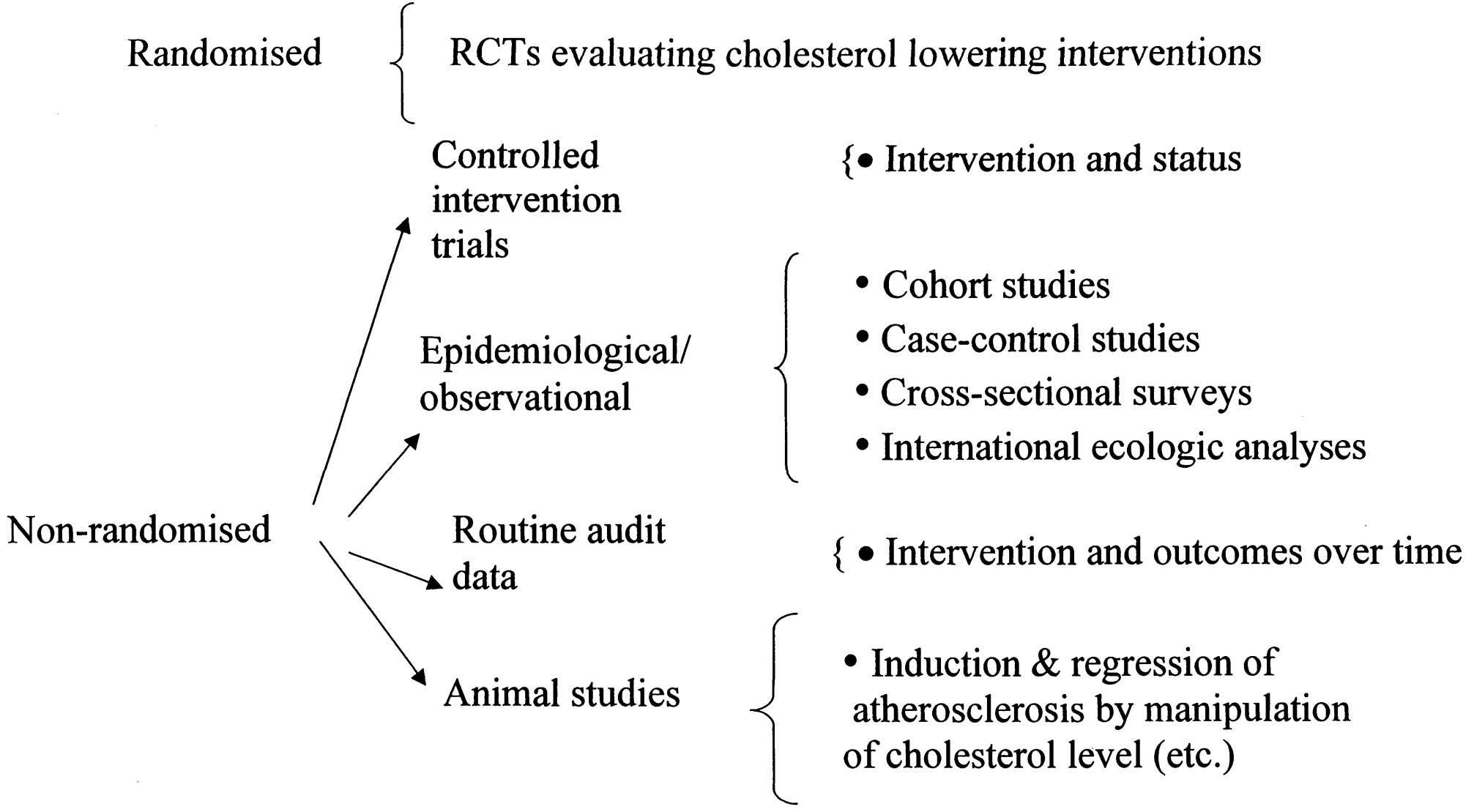
For the generalised synthesis described in this chapter, it was decided to restrict attention to these 10 aetiological observational studies and any reported more recently than 1994 that were of comparable magnitude to these 10. Although this ignores many of the potentially relevant sources of evidence as identified in Figure 4.1, such restriction was necessary to make the analysis feasible.

[1] Fujian Song, Centre for Reviews and Dissemination, University of York

SPECIAL NOTE

THIS ITEM IS BOUND IN SUCH A MANNER AND WHILE EVERY EFFORT HAS BEEN MADE TO REPRODUCE THE CENTRES, FORCE WOULD RESULT IN DAMAGE

Fig 4.1 Summary of all the evidence for potential use in generalised synthesis of evidence of cholesterol lowering



4.3 Overview of literature/studies identified

The literature search located 43 previous meta-analyses of the RCTs (up to 1997) and a list of these is given in the meta-analysis study reference list in Appendix A.I. From these and additional searches 64 distinct RCTs assessing a variety of cholesterol lowering interventions were identified up to September 2001, references to which are given in Appendix A.II. In addition, a cluster dynamic cohort cross-over study was also identified (see Appendix A.II.)

No aetiological observational studies larger than the 10 identified by Law et al. (Law et al. 1994a) were identified, however one update report from one of the original 10 was located. Citations to these 10 observational studies are provided in Appendix A.III.

To make referencing of the different sources of evidence clear, the following convention is adopted. Because several types of study are used, some of which include many individual studies, and often multiple references are associated with each of these, separate reference lists are given for each study type in Appendix A. A letter prefixes the study number; this represents the type of study. Hence, all RCTs are prefixed with an R, crossover studies with an X, cohort studies with a C, and previous meta-analyses with an M. In order to distinguish an individual paper describing a study, a small case letter is added when referring to an individual paper. For example R31 reefers to RCT number 31, and R31b refers to the second paper on the list that describes trial R31.

4.4 Critical review of the previous meta-analyses of the randomised evidence

At least 43 previous meta-analyses have been carried out on the cholesterol lowering intervention trials. All of these trials are randomised except for one cluster crossover dynamic cohort designed study (X1) which was sometimes included (see Appendix A.I). Very few of these have used the same set of trials in their analyses. The date

each analysis was carried out can partly explain the differences in trial selection (i.e. the more recent meta-analyses include more recently published studies). Additionally, many variations in hypotheses being tested and the inclusion criteria, have led to what would initially appear similar meta-analyses combining different studies, assessing different endpoints, examining different covariates, and using different models to combine the data.

Before going on to conduct any further meta-analyses on the RCTs, it is constructive to review previous attempts, and to assess the methods used critically. Examining these papers assisted in the identification of potential differences in study design, and patient characteristics, which are investigated further by including them as covariates in meta-regression models (see section 4.9). A number of narrative reviews were also identified at the literature-searching phase. The description of the RCTs provided in these is often more detailed than those of the meta-analyses. These were helpful in ascertaining the qualitative differences between the RCTs. Where these have been used they are cited in the text.

It was also possible to extract data on a number of the RCTs from the meta-analyses directly, though differences in variables considered between the meta-analyses, and inconsistencies in the data used between meta-analyses, meant the original reports often needed to be consulted also. Additionally, some of the meta-analyses used data updated since the primary study reports were published; this was extracted from the meta-analyses where possible.

A brief synopsis of some of the most influential meta-analyses is given below. One of the first published reviews of cholesterol-lowering trials was by Peto, Yusuf, and Collins [M2] in 1985 which was largely inconclusive due to the modest amount of evidence available at that time. In 1990 Muldoon et al. [M11] carried out a meta-analysis containing six RCT's. This review found no conclusive evidence that overall survival rates were reduced through cholesterol lowering, but further investigation was warranted. This analysis was later criticised for improper data extraction (Oglesby and Hennekens, 1992), and for including trials that were not analysed by the intention to treat method. (Oglesby and Hennekens, 1992) It was updated in 1992 by Davey-Smith and Pekkanen. [M17] The same year Ravnskov produced a meta-

analysis incorporating 22 trials. Antman and Lau in their influential paper (Antman et al. 1992) used cholesterol lowering as one of their examples to demonstrate, using cumulative meta-analysis, how slowly treatments are implemented in routine practice, after they are shown to be effective. In 1994 Davey Smith et al [M25] analysed 35 RCTs in their meta-analysis finding associations between effectiveness and baseline risk. Shortly afterwards, Law, Wald and Thompson [M28 & M29] produced meta-analyses which compared RCT results with other types of studies, finding they largely agreed. This is the only meta-analysis to consider non-randomised evidence (although, as mentioned previously, it did not pool the randomised and non-randomised together). Recently, meta-analyses considering specific interventions, such as diet [M38], or statin drugs [M43] have been carried out.

Ten of these previous meta-analyses carried out between 1990 and 1997 were examined in detail. The selection of these ten was somewhat arbitrary, however, criteria considered during selection included: a) being substantially different from previous analyses; b) being thorough and of good quality; c) using sophisticated analysis methods such as meta-regression. Table 4.1 describes which RCTs were included in each of these 10 meta-analyses. In this table both trials and meta-analyses are ordered by publication year; this makes it possible to distinguish between trials that were deliberately excluded (or missed by each meta-analyst), and those trials for which the results would not have been available. Clearly, the exact times the trial results become available, and the precise times the meta-analyses were carried out would be very difficult ascertain, but the shaded part of Table 4.1 suggests results that clearly were published after the meta-analysis and hence the results of these RCTs would not have been available. In the meta-analysis by Pekkanen and Davey-Smith [M17] results from trial R27 must have been accessed from researchers prior to them being formally published. It can be seen that even after taking publication date into account, there is still considerable variation in the trials included in each of the 10 meta-analyses. It should be noted that Holme [M40] explicitly stated that he included exactly the same trials as Davey-Smith et al. [M25]. Table 4.2 summarises other important characteristics of these ten meta-analyses.

Table 4.1 Summary of RCTs included in 10 meta-analysis from 1990 to 1997

Study Number	Muldoon M11 (1990)	Davey Smith & Pekkanen M17 (1992)	Ravnskov M18 (1992)	Davey Smith et al. M25 (1993)	Law et al. M28 (1994)	Gould M36 (1995)	Holme M40 1996	Rembold M41 (1996)	Marchioli M42 (1996)	Hebert et al. M43 (1997)
R12 (1961)			x	X		X	x			
R2 (1962)			x	X			x			
R5 (1963)			x	X			x			
R22 (1965)			x	X	x	X	x		x	
R8 (1965)			x	X	x	X	x		x	
R7 (1966)			x	X	x		x			
R17 (1968)			x	X	x	X	x	x	x	
R10 (1968)				X	x	X	x		x	
R16 (1969)	X	x	x	X	x	X	x		x	
R5 (1969)				X			x			
R41 (1970)			x					x		
R36 (1971)					x					
R11 (1971)			x	X	x	X	x	x	x	
R13 (1972)				X		X	x		x	
R35 (1973)				X	x		x			
R15 (1975)			x	X	x	X	x	x	x	
R6 (1977)			x	X	x	X	x	x	x	
R20 (1978)	x	x	x	X	x	X	x		x	
R21 (1978)			x	X	x	X	x		x	
R31 (1978)	x	x	x	X	x	X	x	x		
R43 (1978)								x	x	x
R60 (1979)										
R4 (1981)				X	x		x		x	
R37 (1982)			x			X				
R39 (1983)			x							
R23 (1984)				X	x	X	x	x	x	
R28 (1984)	x	x	x	X	x	X	x	x		
R38 (1985)			x							
R40 (1986)						X				
R26 (1987)				X	x	X	x	x	x	
R29 (1987)	x	x	x	X	x	X	x	x		
R24 (1989)	x	x	x	X	x		x			
R9 (1989)				X	x	X	x		x	
R32 (1990)				X			x	x	x	
R33 (1990)				X			x	x	x	
R34 (1990)				X	x	X	x	x	x	
R25 (1990)				X	x	X	x	x	x	
R57 (1991)										x
R30 (1991)		x		X	x		x			x
R19 (1991)				X	x	X	x	x	x	x
R1 (1992)				X			x	x	x	
R51 (1992)								x		
R14 (1992)				X	x	X	x	x	x	
R27 (1993)		x		X	x	X	x		x	
R52 (1993)								x	x	x
R42 (1993)								x	x	x
R55 (1993)									x	x
R44 (1993)								x		x
R46 (1994)								x	x	
R47 (1994)								x		
R48 (1994)								x	x	x
R49 (1994)								x		x
R53 (1994)									x	
R54 (1994)									x	x
R45 (1995)								x		x
R50 (1995)								x	x	x
R56 (1995)										
R58 (1995)										x
R59 (1996)										x
R61 (1997)										
R62 (1998)										
R63 (1998)										
R64 (1999)										
X1		x								

SPECIAL NOTE

THIS ITEM IS BOUND IN SUCH A MANNER AND WHILE EVERY EFFORT HAS BEEN MADE TO REPRODUCE THE CENTRES, FORCE WOULD RESULT IN DAMAGE

Table 4.2 Summary of meta-analyses of cholesterol lowering treatments

Meta Analysis	Year Published	Studies included	Question meta-analysis was trying to answer / endpoint(s)	Description of inclusion criteria	Type of analysis	Pooled estimate(s) (95% CI)	Conclusions/ recommendations
Muldoon M11	1990	6 RCTs	To determine the effects of lowering cholesterol concentrations on total and cause specific mortality in randomised primary prevention trials	Criteria: 1) randomised clinical primary prevention trial of serum cholesterol reduction. 2) it included a treatment group that received instructions for a diet or was given drugs, or both, to reduce cholesterol, and had a control group: all these studies had intended to test the hypothesis that lowering cholesterol prevents coronary heart disease. 3) it resulted in a lowering of serum cholesterol in the intervention group, relative to the control group. 4) It reported both total mortality and cause specific mortality. (Cummings (Cummings and Psaty, 1994) comments limited to men and only)	Peto method (fixed effect)	Total mortality 1.07 (0.94-1.21) CHD mortality 0.85 (0.69-1.05) Cancer mortality 1.43(1.08-1.90) Mortality not related to illness 1.76 (1.19-2.58)	Chol reduction fails to improve total mortality. Association between reduction of cholesterol and deaths not related to illness warrants further investigation.
Davey Smith & Pekkanen M17	1992	7 RCTs 1 Cluster Cross-Over	Does lowering cholesterol reduce mortality (from all causes and specific causes) Drug and diet trials analysed separately.	Primary intervention studies of the primary prevention of heart disease	Fixed effect?	Total mortality Diet 0.95 (0.82-1.09) Drug 1.16 (0.98 -1.38) CHD mortality Diet 0.71 (0.55-0.90) Drug 0.72 (0.55-0.94) (cancer,injury and other	(General discussion only given)

						non CHD given)	
Ravnskov M18	1992	22 RCTs	To see if the claim that lowering cholesterol levels prevents coronary heart disease is true or if it is based on citation of supportive trials only	Criteria: designed and successful lowering of cholesterol concentrations aimed at preventing coronary heart disease, and total mortality or incidence of coronary heart disease reported as end points. (open and blind trials both included) Trials using angiography were excluded.	Peto method (fixed effects)	Total mortality 1.02 (0.97 -1.07) CHD mortality 0.94(0.88-1.00) non-fatal CHD 0.90(0.84-0.96)	Lowering serum cholesterol concentrations does not reduce mortality and is unlikely to prevent CHD.
Davey Smith et al. M25	1993	35 RCTs	To investigate the level of risk of death from coronary heart disease above which cholesterol lowering treatment produces net benefits	All randomised controlled single factor trials of cholesterol lowering treatment with at least six months of follow up in which at least one death occurred. One study (r18) proved impossible to ascertain in which arm of the study the solitary death occurred. (Published and unpublished RCTs)	Random effects, meta-regression and subgroup analyses	Total mortality, CHD mortality, and other cause mortality given stratified by risk of death (& regression line), and also for diet and drug trials separately	Cholesterol lowering drugs seem to produce mortality benefits in only a small proportion of patients at very high risk of death from CHD
Law et al. M28	1994	28 RCTs 10 Cohort studies 3 Internatio nal surveys (only the RCTs were pooled)	To estimate by how much and how quickly a given reduction in serum cholesterol concentration will reduce the risk of ischaemic heart disease	Cohort studies of serum cholesterol concentration and ischaemic heart disease (10 biggest; all >350 ischaemic heart disease events) International studies (not really stated) RCTs - intervention drug, diet or ileal bypass surgery ; outcome ischaemic heart disease events (deaths and non-fatal infarcts). Recorded at least one death and documented a reduction in serum cholesterol concentration of at least 1%.	?? (not specified)	Results given for trials of drug, diet, men without ischaemic heart disease, and with IHD. All these are stratified by follow-up	Results from cohort studies, international studies and RCTs are consistent. There is a benefit from having low cholesterol in relation to IHD.

Gould M36	1995	35 RCTs	1) what is the relation of cholesterol lowering to benefit (or harm) 2) what are the effects of specific types of lipid-lowering regimens on clinical outcomes?(CHD mortality, non-CHD mortality, total mortality)	All published RCTs relating cholesterol reduction to changes in mortality or coronary morbidity and had a duration of ≥ 2 years. (Included angiographic trials, hormones, surgery or multifactorial interventions.	Regression - modelling degree of cholesterol lowering achieved, and effects of type of intervention	Slope predicts 13-14% reduction in CHD mortality for every 10%-point reduction in serum cholesterol. Non-CHD mortality not significantly related to cholesterol reduction. But total mortality is (results given)	Results suggest chol lowering is beneficial but specific adverse effects of Fibrates and hormones increase the risk of CHD (hormones only), non-CHD, and tot mortality.
Holme M40	1996	35 RCTs (same as Davey Smith et al. (Smith et al. 1993))	Review relationship between log odds ratio of total mortality in active vs. control group and degree of cholesterol reduction, adjusting for appropriate covariates such as type of trial (single of multifactorial), risk level in the control group, and type of treatment.	See Davey Smith et al. (Smith et al. 1993)	Odds rations calculated by Peto method. Fixed effect regression, using absolute cholesterol reduction as predictor.	Regression models given for total mortality.	Cholesterol reduction was of borderline significance.
Rembold M41	1996	33 RCTs	"NNT of the prevention of MI and death by antidyslipidemic therapy". Outcomes - Relative risk reduction, NNT	Randomised and involved standard antidyslipidemic therapy (diet, pharmaceuticals, and surgery). For dietary trials (r42-r45) were excluded because reduction in total cholesterol was small(<4%). (primary, secondary and tertiary risk	Cumulative meta-analysis (by publication date), fixed and random effects. Subgroup analysis on HMG-CoA reductase inhibitors or niacin)	Results for all major outcomes reported separately of primary and secondary trials. Further subgroup results are also given	Results support clinical benefit of treating dyslipidemia, both in persons with and without known atherosclerosis

				included). Non-standard treatments (triiodothyronine, garlic, or walnuts) were excluded	("studies with treatments typically employed by physicians in the 1990s")		
Marchioli et al. M42	1996	37 RCTs (34 plus 3 had multiple arms)	"Nonfatal events were not easily available from all published articles, so they were not assessed" (all cause mortality) (odds ratios)	1) designed lowering of blood cholesterol concentrations in the **secondary prevention** of cardiovascular disease (CVD); 2) randomised design; 3) recorded incidence of total mortality; 4) at least 6 months follow up (open and blind trials were accepted) (multifactorial trials and studies that reported data that were gathered beyond planned trial duration were excluded)	Fixed effects, random effects?, regression(fixed) (level of CHD risk, baseline blood total cholesterol level, percentage of achieved reduction of the cholesterol concentration, CHF, Previous MI, interactions)	Many outcomes, and subgroups examined - too numerous to describe.	The effect of cholesterol-lowering interventions at least in the secondary prevention of coronary heart disease can be considered as established, but the transferability of such results to patients remains an unanswered question.
Hebert et al. M43	1997	16 RCTs	To examine whether cholesterol lowering with statin drugs reduces the risks of stroke and total mortality	1) Published English-language articles; 2) Statin drugs alone were used to reduce lipid levels rather than multifactorial interventions including another type of cholesterol-lowering drugs 3) inclusion of data on deaths and/or strokes	fixed effect, Peto, (odds ratios)	Risk of stroke reduced by 29%(14-41%) in statin group and 22%(12-31%) for total mortality. (non significant increases in non CVD and cancer outcomes)	Clear benefit from using statins on stroke and total mortality outcomes

Further to Table 4.2, differences in the type of analyses undertaken are a key issue here, and so considered in more detail. Until relatively recently the only method used to pool the cholesterol RCTs was simple fixed effects analyses. The first random effect model was used in 1993. [M25] This was also the first of the meta-analyses that used regression methods to explore study level covariates. Death rates from coronary heart disease per 1000 person years observed in the control group of each study were examined. The analysis was stratified by this variable, as well as including it as a covariate in a regression model. Unfortunately, the later analysis is flawed (see section 2.4.5 where this is discussed and recently developed valid analyses methods are described) (Senn, 1994). Despite this, the subgroup analysis carried out is valid, and a genuine strong relationship between baseline risk and treatment effect was observed.

In addition to the meta-analysis of Law et al., [M28] Thompson produced and discussed further the analyses based on the same data in a separate paper.(Thompson, 1993) A regression model was used to look at the effect of absolute cholesterol reduction on outcome. Using logistic regression with a log odds ratio outcome alleviates the need to make assumption about normality of the data. (Thompson and Sharp, 1999) A strong negative association was found between ischaemic events and degree of cholesterol reduction. An odds ratio of 0.82 (0.78-0.87) per 0.6 mmol/l decrease in cholesterol was reported, where 0.6mmol/l was the average reduction observed in the trials being combined. Thompson comments that a linear relationship between log odds ratios and absolute reduction in cholesterol, constrained to go through the origin, is assumed. He goes on to say the rationale for this choice of scales comes from large observational prospective studies where it is much clearer that absolute differences in serum cholesterol correspond to proportionate changes in the risk of ischaemic heart disease. (Thompson, 1993) So, in an informal way, information from non-randomised studies was included in this analysis to inform scale choice and the specification of the model.

The effect of treatment duration was also investigated. If treatment duration was simply included as a covariate, the longer trials would include information on events both soon after and a long time after randomisation and so any true effect of duration

would be diluted. The studies were thus subdivided by time periods since randomisation commenced using supplementary information provided by the original trialists. The results form this suggested that there is an increasing benefit of reducing cholesterol with increasing duration of treatment, and that after this and the extent of cholesterol lowering are taken into account there are no obvious systematic differences between drug and dietary trials, or between primary and secondary trials.

Gould et al [M36] also addressed the issue of whether the risk reduction for specified outcomes is related to the actual degree of cholesterol reduction, or the particular means of reducing cholesterol. A regression analysis was implemented including a term for the degree of cholesterol reduction achieved, as well a terms for specific types of intervention (diet, fibrates, hormones and other) in a similar manner to Thompson's analysis described above. An association between cholesterol reduction and CHD and mortality outcomes was observed. In addition, differential effects of fibrates and hormonal therapies was observed.

Holme [M40] investigates: 1) the relationship between total mortality and cholesterol reduction, adjusted for single/multifactor trials; 2) the impact of control of risk level in the control group on this relationship; 3) whether fibrate trials have experienced excess risk as compared to other drug intervention trials. The method of analysis used was (unconditional) logistic regression (compared to Gould who used conditional logistic regression). Holme comments that the initial cholesterol level was not used in adjustments because it did not correlate with the log odds ratio in the data. Single and multifactor trials were split by a dummy variable. Baseline risk was calculated using total number of deaths, as opposed to deaths from CHD used by Davey Smith et al. [M25]

Rembold's analysis [M41] is noteworthy because it uses an expanded set of trials, when compared to the previous studies. Another note of interest is that the number needed to treat scale is used to report outcomes, rather than the odds ratio commonly used. The actual statistical model used is elementary however, with no exploration of covariates.

Marchioli et al [M42] investigated several covariates using logistic regression, namely: 1) level of CHD risk; 2) baseline total cholesterol level; 3) percentage reduction in cholesterol concentration; 4) congestive heart failure; 5) previous MI; and 6) intervention type. This analysis used the majority of the known RCTs and was lengthy and thorough.

Finally, Herbert et al. [M43] produced a meta-analysis combining the statin trials only. Simple fixed effect analyses were carried out on a number of endpoints. This analysis suggests that statins are effective at reducing overall mortality.

Having considered these previous meta-analyses (and others), it would seem that although many are thorough, the analyses carried out are not as sophisticated as is possible currently due to development of meta-analysis methods. For instance, although a number carried out regression analyses to explore covariates, none of them used a mixed effect model, despite it being clear that residual heterogeneity existed. Including baseline risk as a covariate using a standard model is flawed, and although methods are now available to produce a valid analysis this has not been done. A re-analysis using all these RCTs but alleviating the shortcomings identified above is presented in Section 4.9.

Further practical problems identified through examination of the previous analyses are outlined below.

Trials with multiple treatment arms

There seems little consensus on how studies with multiple treatment arms should be included in a meta-analysis. In the previous analyses, various schemes were used, including pooling all treatment arms together, excluding certain arms, and including different arms as separate trials. This latter approach is problematic as estimates from different arms of the same trial are not independent (as they incorporate the same control group). Pooling treatment arms is also problematic for obvious reasons. The RCTs with multiple treatment arms are R10, R14, R15, R38, and R58. Consideration is given to this issue in Chapter 5.

Multi-factor intervention regimes

A problem similar to that of multiple treatment arms arises in multi-factor intervention trials, especially if the effect of individual intervention types is being investigated. Some trials used combinations of interventions to reduce the risk of CHD, while, more specifically, others used multiple interventions to lower cholesterol levels. A related problem is that some trials used drugs only if response to dietary intervention was not satisfactory. These factors make the strict categorisation of studies impossible. For this reason an indicator variable for single- and multi-factor trials is explored (see Section 4.5).

Non-standard designs

Trial X1 used a very unusual design. It is unclear if this study merits the same weighting as a standard RCT.

4.4.1 Meta-meta-analysis?

After this review of meta-analyses was written, an analysis of cholesterol reduction meta-analyses was published. (Katerndahl and Lawler, 1999) In this curious paper, the authors, acknowledging that previous meta-analyses had produced conflicting results, produced a meta-meta-analysis. This analysis took the results of previous meta-analyses and pooled them in a further meta-analysis, including covariates relating to methodological aspects of the previous meta-analyses.

Although this may sound appealing, it is seriously flawed for a number of reasons. Firstly, since all the meta-analyses are based on the same finite source of data – the cholesterol RCTs - they are clearly not independent. Essentially studies are being included multiple times; the actual number depending on how many of the meta-analyses included them initially. Since the earlier trials have potential to be included in more meta-analyses than the more recent ones, then they have the potential to me more influential in the analysis because of this. It would appear that the results of such an approach are unreliable due to this problem.

Although these criticisms were acknowledged by the first author of the paper[2] , this highlights the problems which exist in trying to reconcile a divergent secondary evaluation literature.

4.5 The (completed) randomised trials

For the meta-analyses and generalised syntheses described in this chapter and Chapter 5, 64 completed randomised trials for which cholesterol lowering was at least part of the total intervention were identified as described above. Trials had to report clinical endpoints to be included. Other trials which reported only angiographic endpoints are known to exist. It is acknowledged that such studies reporting angiographic endpoints do add to the total evidence; for example Rubins (Rubins, 1995) describes the effect of lowering cholesterol in people with already low cholesterol (such data are not found in any of the other RCTs). (However, restrictions needed to be made to keep the size of this investigation manageable.) Several important qualitative differences exist between the trials. Perhaps most significantly, different interventions, singularly and in combination, were administered. Drug, diet and surgery interventions were all used to lower cholesterol levels. Even within these three broad categories, interventions varied considerably; for instance, the natures of diets administered, or advised, varied. Similarly, many different drugs have been used; generally the more recent trials used drugs which lower cholesterol levels by greater amounts than was possible using older drugs. Table 4.3 displays the interventions used in the experimental and control groups in all the RCTs.

As well as the interventions administered to the treatment groups, it is worth examining those given to the control groups, as this was very often more than a simple placebo, primarily due to ethical reasons. For example, in several of the

[2] Personal communication

SPECIAL NOTE

THIS ITEM IS BOUND IN SUCH A MANNER AND WHILE EVERY EFFORT HAS BEEN MADE TO REPRODUCE THE CENTRES, FORCE WOULD RESULT IN DAMAGE

Table 4.3 Interventions used in the 60 RCTs cholesterol lowering RCTs

	Drugs										Diets			Surgery	Control group
					Fibrates		Statins								
	Nicotinic Acid	Estrogen	Probucol	Dextroth -yroxine	Gemfib -rozil	Clofibrate	Lovastatin	Simvistatin	Pravastatin	Fluvastatin	Diet -not specified	Low-fat diet	Specific		
Trial Number															
R1													Strict		diet
R2		x													placebo
R3		x													placebo
R4			X												placebo
R5						x									placebo
R6						x									usual
R7											x				usual
R8												x			usual
R9												x			no low fat diet
R10	x	x		x											placebo
R11						x									placebo
R12		x													lactose
R13						x									placebo
R14 (2 active groups: diet and cholestyramine)											x			x	usual
R15						x									placebo
R16											x	x(vegetarian)			usual
R17													Soya bean oil		usual
R18						x									placebo

R19							x								usual
R20															placebo
R21											x				usual
R22													Olive+ Corn oil		usual
R23											x				placebo
R24											x				usual
R25														x	control
R26											x				placebo
R27					x										placebo
R28															placebo
R29					x										placebo
R30							x								placebo
R31						x									olive oil
R32									-						usual
R33 (drugs given varied)							x				x				diet
R34 (2 active groups Nacin-Colest & Lova- Colesr							x				x				diet
R35															placebo
R36						x									placebo
R37											x				usual
R38(drugs only added if necessary)			X			x					x				usual
R39											x				usual
R40						x (if necessary)					x (advice)				usual
R41											x				usual
R42							x				x				diet
R43								x			x				diet
R44									x		x				placebo
R45									x		x				placebo
R46 (Drugs added only if	x				x				x		x				diet

necessary)															
R47 (Interventins individualised drugs given only if necessary)			X		x		x					x	low chlesterol + high carbohyd rate		usual care
R48									x		x				diet
R49							x				x				diet
R50									x		x				diet
R51											x				usual
R52							x				x				diet
R53													alpha-linolenic acid-rich		usual
R54							x								placebo
R55									x		x				diet
R56 (3 groups diet(control), lova+cholesty,lova+probucol			x				x								diet
R57									x (4 active arms got different doses)						placebo
R58									x						placebo
R59									x						placebo
R60											x (advice)				usual
R61								x			x				diet
R62									x		x				diet
R63							x				x				diet
R64										x					placebo

"statin" trials dietary interventions have also been administered to both treatment and control groups.

4.5.1 Outcome variables

Prior to the synthesis of studies, it is also necessary to decide on definitions for outcome variables. The extraction of a common outcome variable proved to be a non-trivial exercise. This was largely due to several different outcomes having been reported for the RCTs and requirement that compatibility with outcomes from other study designs to be maintained. Considering previous meta-analyses of the RCTs alone, potential candidates for outcome variables were those based on all cause mortality, coronary heart disease mortality, coronary heart disease events (fatal or non-fatal episodes), ischaemic heart disease mortality, and ischaemic heart disease events. After examination of the cohort study reports (see Section 4.6) it appeared that the outcome most often reported was coronary heart disease mortality, and this appeared to be the most compatible outcome for generalised synthesis. In addition, total mortality endpoints were extracted for the RCTs.

For the RCTs, extracting the specific numbers for both arms of the trials was sufficient to enable the calculation of an odds ratio (see Appendix A.V. for the extracted outcome data).

4.5.2 Covariates considered here

Characteristics by which the RCTs could potentially differ are discussed below. Data was extracted from all the trials on these characteristics and used in the analyses which follow. See Appendix A.V. for the data for these characteristics.

Intervention type

As previously mentioned, there are three distinct cholesterol reduction intervention types, namely, diet, drug, and surgery. Previously the drug trials have been broken down into smaller subgroups, since different types of drugs lower cholesterol levels by considerably different amounts, and have potentially different side effects. A

sensible dissection of this group is into fibrates, hormones, statins, and other. Additionally, it should be pointed out that different dose levels are used in different trials, but the data are insufficient to allow investigation of dose levels within drug types.

Population type / Baseline risk of CHD in the control group

Reports indicate whether the majority of the subjects had pre-existing CHD. Additionally, some trials' subjects were all diabetic. Hence a three level variable identifying primary, secondary, or diabetic subjects was created.

Although many reports state clearly whether they are investigating treatments for primary or secondary prevention, i.e. whether the predominant patient group have previous history of heart disease, Davey Smith et al. [M25] demonstrated, for 35 of the RCTs considered here (R1-R35), the event rate in the control arms varied greatly, with some primary intervention trials reporting greater levels of risk than secondary prevention trials. For this reason the control group event rate has been reported as CHD deaths per 1000 person years. This was calculated using the same formula as Davey Smith et al., [M25] reproduced below.

$$CHD\ deaths\ per\ 1000\ person\ years\ in\ control\ groups = \left(\frac{Coronary\ heart\ disease\ death}{years\ of\ follow\ up \times \left(\begin{matrix} number\ alive\ at\ end\ of\ trial + \\ 0.5(number\ dying\ during\ study) \end{matrix} \right)} \right) \times 1000 \qquad (4.1)$$

Duration of study

Clearly, trial follow up is potentially an important factor, and one in which the trials vary quite considerably (Thompson (Thompson, 1993) noted a dilution effect over time). Hence, the comparative estimate of treatment effect could be time dependent. Although this is not directly estimable, creating a covariate for duration of study does allow some exploration of its impact.

Cholesterol levels/changes in levels

For the majority of RCTs baseline and intervention altered cholesterol levels are given. There are potentially several possible measures of cholesterol level including total serum blood, low density lipid (LDL), and high density lipid (HDL) cholesterol levels. Previous meta-analyses appear to have concentrated on total cholesterol exclusively as is done here. Baseline cholesterol measurements are reported for the vast majority of the trials, with some reports of measurement at the end of the trial, and often at intervals during the trial as well. Several analyses have looked at change in total cholesterol level over the duration of the trial, in both absolute and percentage terms. Investigating absolute change is preferred here because modelling produces more clinically meaningful measurement, and also it makes later comparison with the observation studies possible.

Percentage female patients

If males and females respond differently to the intervention the relative composition of the trial population could have an effect on the trial outcome, and the percentage female patients is thus extracted.

Age of patients

Age may affect the effectiveness of cholesterol lowering interventions. Unfortunately, data on patient age appears limited in many trial reports. There are also different reporting methods; some reports record the mean age at entry, while others only give the age range. Hence, the decision was taken not to investigate this potential effect modifier.

4.6 The aetiological cohort studies

The cohort studies generally assess serum cholesterol levels in individuals and then follow them up over a period of time. Outcomes such as CHD events, CHD deaths, and all causes of mortality are generally reported in various combinations for each study. Results are commonly expressed as incidence per 100 000 person year observations for quintiles of the population defined by cholesterol levels recorded at the onset of the study, although other divisions of the cohort are used in some instances. The incidence estimates for each group are often adjusted for confounding risk factors, such as smoking and age, although there is little uniformity between studies in the way this is done.

As mentioned previously, many cohort studies of this type are known to exist, but here attention is restricted to the ten largest as reported by Law et al. (Law et al. 1994a) (with the data for study C2 updated to include longer follow up [C2b]). The citations for these studies are given in Appendix A.III.

Table 4.4 provides descriptive data for each of these ten studies. Only studies C5 and C8 included women, and these results are reported separately. CHD mortality is used as the main outcome as this appears to maximise the potential use of the data available from both the RCTs and cohort studies. Two covariates, the age range and follow-up lengths of the studies are also reported in the table. It should be noted that there are tens of thousands of patients and several thousand deaths in the largest of these studies, making them considerably larger than the largest RCTs, although the largest RCT [R39] did randomise over 50,000 subjects.

4.7 The relationship between the RCTs and the cohort studies

Clearly a fundamental issue regarding the combination of the RCTs and the cohort studies is the compatibility of the data. The RCTs produce a comparative measure of effect between groups that are administered a cholesterol lowering intervention and those that are not. The cohort data, as it is usually presented, gives incidence rates for

persons grouped into quintiles for cholesterol levels for that particular cohort with no intervention having been given to them.

In order to make the two sorts of data combinable, some reconfiguration of reported results for at least one type of study is necessary. A sensible procedure was adopted by

Table 4.4 Descriptive statistics for the 10 Cohort studies considered

Study	Number of subjects	Deaths from IHD	Deaths from CHD	Total deaths	Age range (at start of study)	Average follow up
C1	21515	538		1543	35-64	13
C2	7735		640	1257	40-59	14.8
C3	6897	*[1]	*[1]	495	51-59	7 (7.1)
C4	361 662		6327		35-57	12
C5	7000		878		45-64	15
C6	17718		1676	4155	40-69	18
C7	7961		371	2072	46-65	19
C8	46140		6626	15744	17-74	21 (18-20)
C9	9902		1098	3473	40-65	23
C10	8274				40-59	9
Women						
C5	8262		490		45-64	15
C8	46570		3607	11348		21

Note:

[1]*Study 3 reported the following outcomes: non fatal MI - 234; coronary death 171; and total cardiovascular deaths 204;

Law et al., [M28] when informally comparing the randomised and observational evidence. Here a log-linear model was fitted to the incidence rate/average cholesterol

level (quintile) data. Law et al. report that this model fitted the data well, and was not significantly improved upon by more complex models (such as including a quadratic term). [M28] Interpretation of such a model is straightforward: a constant absolute difference in serum cholesterol concentration, from any point on the cholesterol distribution, is associated with a constant percentage difference in the incidence of ischaemic heart disease (the outcome they considered). Hence, a change in risk for any given change in cholesterol level can be calculated. In this way changes in risk associated with decreases in cholesterol observed in the trials can yield comparable estimates.

Law et al.'s original analysis was actually more sophisticated than this because it adjusted the results for age at death, and for two forms of bias incurred in observational studies, namely regression dilution bias and surrogate dilution effect. (Law et al. 1994b) Unfortunately, the results of these original analyses cannot be used here as their outcome of interest was ischaemic heart disease events, but, their basic approach is repeated for CHD mortality, but due to time restrictions, adjustments for the two forms of bias were not made.

4.7.1 Achieving compatible incidence data from the cohort studies and the RCTs

Of the ten cohort studies considered, only seven had data on CHD mortality (Table 4.5), so it was necessary to exclude the others (studies C1, C3, and C10) from the analysis. For these remaining seven, information was available to various degrees of completeness. For C2a, death rates had to be read off a graph with a fairly crude scale (C2a p.410). For study four, the required data were available in tabular form (C4a Table 2) allowing accurate extraction. For study C5, figures had to be read off a very crude graph (C5a Fig.3) to get adjusted mortality rates, and similarly the quintile figures were hard to establish accurately, though attempts were made at dividing histograms of cholesterol distributions (C5a Fig.1) into quintiles as a cross-check. It was impossible to extract the required data for study C6 because results were presented for different age ranges, rather than different cholesterol levels (C6a). Two reports were available for study seven, the first (C7a) provided the information required in tabular form (C7a Table 1); an updated report describing an increased length of follow up (C7b) could not be used as it did not provide the required data.

Study C8 reported results in a different way from the rest, reporting relative risks of death from CHD for quintiles (for total population and for males and females separately), relative to the first (lowest cholesterol) quintile, and reported a relative risk per 0.4 mmol/l increment in cholesterol. Unfortunately it was not possible to convert this data into a form compatible with the remaining studies. Data for study C9 had to be read off a crude graph (C9a Fig. 1) to get incidence rates for deciles of cholesterol, adding measurement error.

Hence, to summarise, from the initial ten studies, three had no information on deaths from CHD, and two did not provide detailed enough data, or data in a form that was comparable with the other studies. This left five studies (numbers C2, C4, C5, C7, and C9) from which data was extracted, for inclusion in the analysis. The decision to use data on males only was taken to keep the analysis simple, however, the possibility of including data on female patients is noted.

4.7.2 Calculating a (pseudo) effect size for each cohort study compatible with those from the RCTs

The method Law et al. [R28] used previously to informally compare the results of the observational studies and RCTs has been briefly described above; this was essentially replicated below, but using CHD mortality as the endpoint variable. Incidence rates for CHD mortality for quintiles (or other proportions) of cholesterol for the five studies are plotted on individual graphs in Figure 4.2.

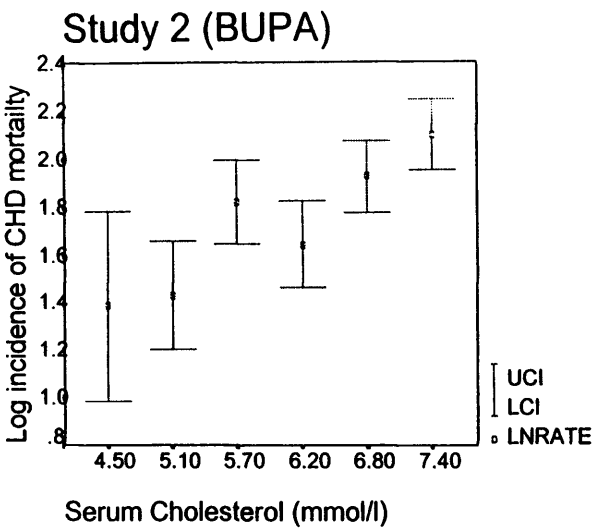
It is important to note that in some study reports data were adjusted for age, but not in others. By necessity, this implies that results for studies C2 and C4 are based on raw incidence rates, while studies C5, C7, and C9 are age-adjusted estimates.

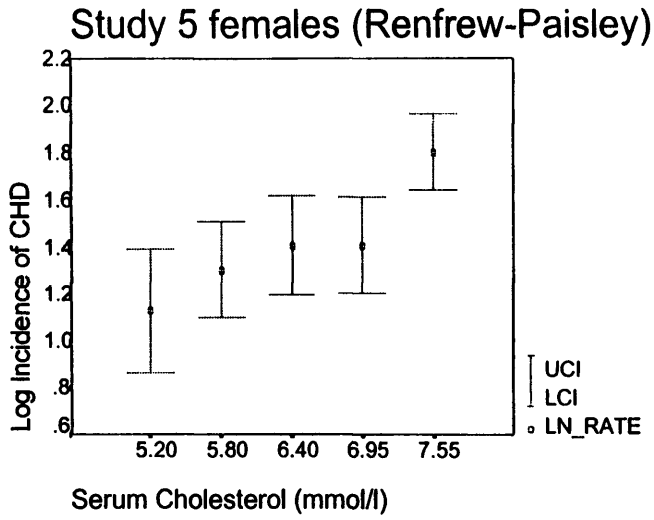
Weighted regression was used to fit a linear relationship for each of these plots, weighting being based on the inverse of the standard error of the incidence rate for each cholesterol group. [M28] All slope coefficients were clearly positive (See Table 4.5 for parameter estimates). The magnitude of such a log-linear relationship provides, for a constant absolute difference in serum cholesterol concentration, a constant percentage difference in the incidence of CHD mortality. Law et al. (M28)

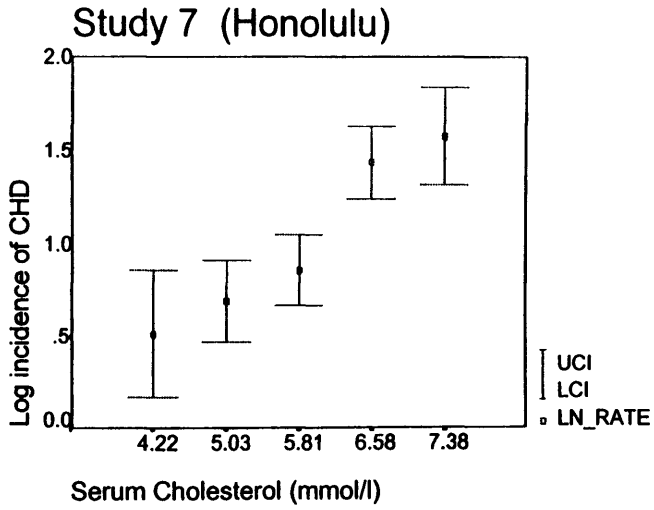
SPECIAL NOTE

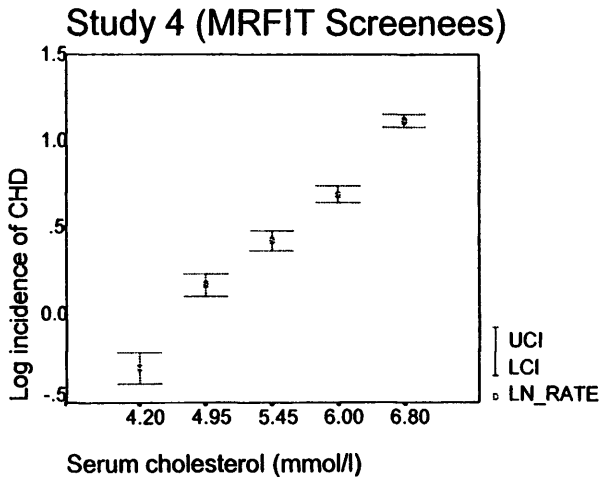
THIS ITEM IS BOUND IN SUCH A MANNER AND WHILE EVERY EFFORT HAS BEEN MADE TO REPRODUCE THE CENTRES, FORCE WOULD RESULT IN DAMAGE

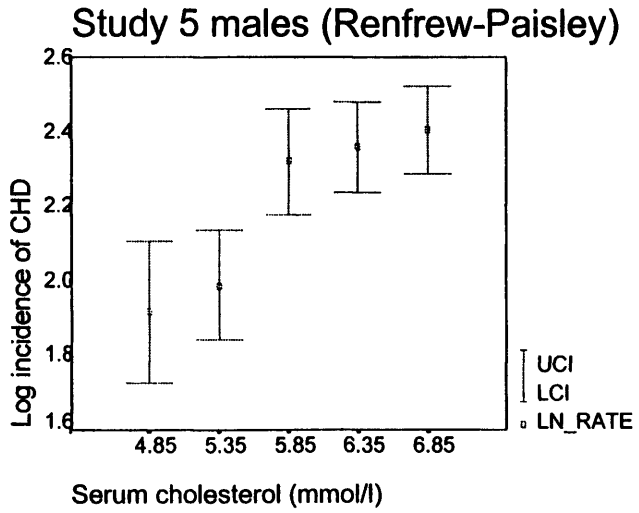
Figure 4.2 Incidence of CHD mortality per 1000 person years, with 95% confidence intervals, for fifths (or other given fraction) of distribution of serum cholesterol concentration

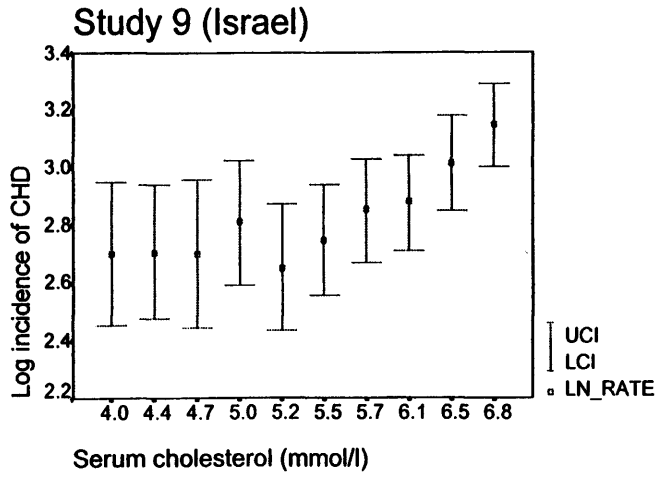












described the estimated percentage decrease in ischaemic heart disease per 0.6 mmol/l decrease in serum cholesterol, as the average reduction in serum cholesterol in the RCTs they considered was 0.6 mmol/l. In the set of RCTs considered here, for those studies for which the average reduction in serum cholesterol level is known, the (weighted) average reduction of serum cholesterol between the control and treatment group was 0.87 mmol/l. Hence, in our set of trials the average cholesterol reduction is nearly 50% larger than that observed in the trials examined by Law et al. (M28). This can be explained due to the current set of trials including more recent studies, in particular those which use "statins", which are known to reduce cholesterol levels by amounts greater than was previously possible.

The percentage decrease in incidence of CHD mortality related to a 0.87 mmol/l decrease in serum cholesterol, using the regression equations, is calculated for each study (Table 4.6). These estimates are then converted into relative risks by the formula:

Relative risk for 0.87 decrease in cholesterol = 1- (% decrease in CHD mortality per 0.87 mmol/l decrease in cholesterol)/100 (4.2)

For example, in study C2 the percentage decrease in CHD mortality per 0.87 mmol/l decrease in cholesterol is 0.193 leading to a relative risk of 1-0.193 = 0.81.

Table 4.5 Coefficients for the weighted regression line:

ln(incidence rate) = α + β(cholesterol level)

Study	α (S.E.)	β (S.E.)
C2	0.255(0.327)	0.247(0.052)
C4	-2.507(0.084)	0.535(0.014)
C5 males	0.592(0.334)	0.274(0.057)
C7	-1.177(0.360)	0.378(0.060)
C9	1.946(0.170)	0.163(0.030)

Table 4.6 Calculation of outcomes for 5 cohort studies

Study	% decrease in CHD mortality per 0.87 mmol/l decrease in cholesterol	Relative Risk (95% confidence interval)
C2	19.3	0.81 (0.74-0.88)
C4	37.2	0.63 (0.61-0.64)
C5 males	21.2	0.79 (0.71-0.87)
C7	28.0	0.72 (0.65 - 0.80)
C9	13.2	0.87 (0.82- 0.91)

Relative risks for the RCTs can be calculated directly and so in principle, the RCTs and the cohort studies can be combined, using the relative risk scale. Indeed, on first inspection it would seem sensible to base analysis on this outcome scale. Unfortunately, there is a drawback with this. It was noted earlier that baseline risk in the control group was one of the potential covariates which would require examination in the RCT model.

The method described by Thompson et al. (Thompson et al. 1997) for doing so is, however, only possible in practice if one is combining studies using the odds ratio scale.

The calculation of odds ratios for the RCTs is direct. Sinclair and Bracken (Sinclair and Bracken, 1994) report how to convert the estimates from the cohort studies into odds ratios this can be done, for comparative studies, using the formula below:

$$RR = \frac{OR}{1 + I_c(OR - 1)}, \tag{4.3}$$

where I_c is usually the incidence (i.e. number of deaths from CHD/ total number of people) of events in the base comparison group. Rearranging (4.3) gives:

$$OR = \frac{RR(1 - I_c)}{1 - RR(I_c)} \tag{4.4}$$

Hence, the conversion is dependent on knowing the incidence rate in the base comparison group. Unfortunately, since the relative risks used here were derived from regression slopes and not a direct comparison between two groups there is no base comparison/control group as such. To get round this issue, the average incidence within each cohort study was calculated and used in equation (4.3).This assumes that the average incidence in the cohort studies is comparable with the incidence rate in the control arms of the RCTs. Each study's estimate is displayed in Table 4.7. It is interesting to note that the incidence rates in the cohort studies are much lower than for the vast majority of trials. Odds ratios based on these incidence rates are also provided in Table 4.8. It can be seen that the odds ratios and their confidence intervals differ only very slightly from the corresponding relative risks. This is to be expected because the incidence rates are relatively low, and for "rare" diseases the relative risk approximates the odds ratio. (Greenland, 1987)

Table 4.7 Calculations required to derive odds ratios for the cohort studies

Study Number	Average incidence rate	% decrease in CHD mortality per 0.6 mmol/l decrease in cholesterol	Relative Risk (95% confidence interval)	Odds Ratio (converted using average incidence rate) (95% confidence interval)
C2	640/7735 = 0.083	14 (8-19)	0.81(0.74-0.88)	0.80(0.72-0.87)
C4	6327/316099 = 0.020	27 (26-29)	0.63(0.61-0.64)	0.63(0.61-0.64)
C5 – males	878/7137 = 0.123	15 (9-21)	0.79(0.71-0.87)	0.77(0.68-0.85)
C7	371/7961 = 0.047	20 (14-26)	0.72(0.65-0.80)	0.71(0.64-0.79)
C9	1098/10059 = 0.109	9 (6-12)	0.87(0.82-0.91)	0.86(0.80-0.90)

4.7.3 Issues related to the comparison of the RCT results with the cohort studies

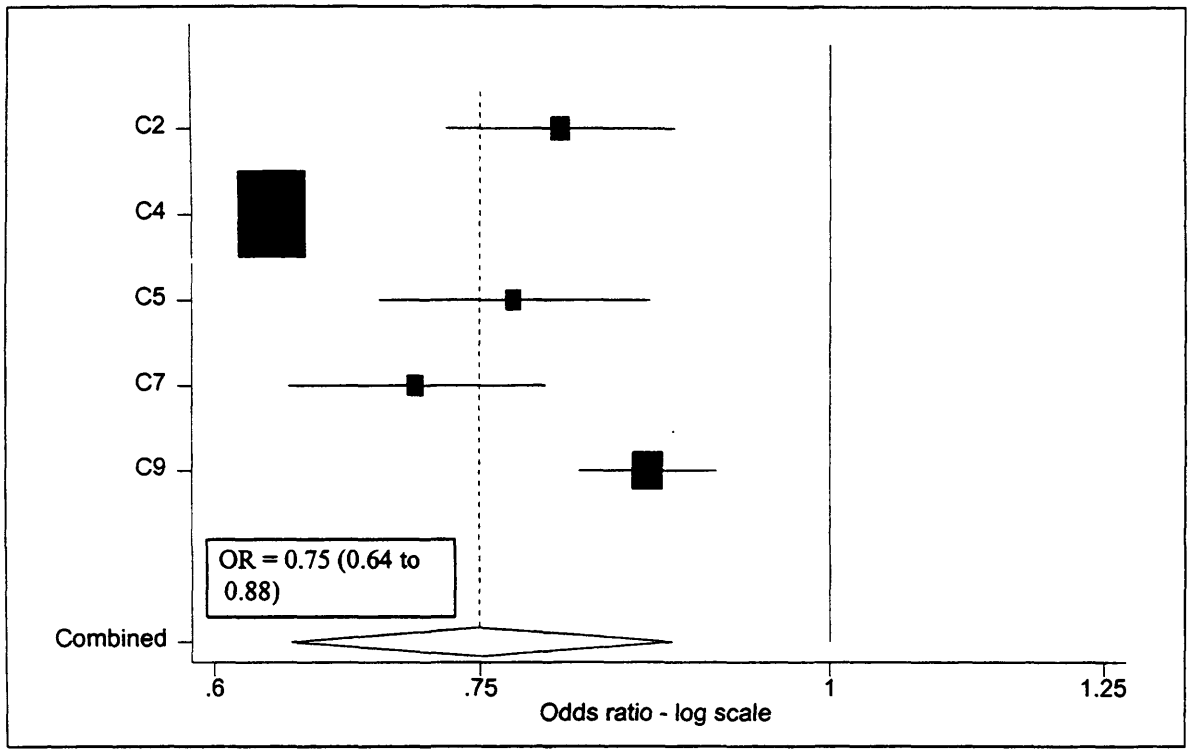
The previous section demonstrated a way of producing results on comparable scales for the cohort and randomised studies. A strong assumption made in combining this data is that the risk reduction associated with reducing cholesterol using an intervention to a particular level is equivalent to the odds ratio between persons having naturally occurring cholesterol levels which differ by the amount reduced in the RCT. Hence, it assumes the beneficial effect, when lowering cholesterol levels, occurs quickly (theoretically instantaneously). (It also assumes no detrimental risks occur when administering interventions, such as adverse events due to drug side effects.) These issues need careful consideration, although Law et al. [R28] did note that the results from the cohort studies, RCTs and the international studies agreed remarkably well. It is

also worth noting that such assumptions are only required when aetiological studies are combined with intervention studies. In the more common scenario of combining randomised and non-randomised studies, such as the electronic fetal monitoring example considered in Chapter 6, no such concerns exist.

4.8 Combining the cohort studies separately

Before proceeding to the synthesis of the randomised and non-randomised evidence, data from each study type is explored and pooled separately. Concentrating initially on the cohort studies, the dataset used for the analysis is provided in Appendix A.VI. Odds ratios from the five cohort studies, are combined using fixed and random effect (classical) models. The female data for study C5 is omitted, because it did not seem sensible to combine it with data on males only from the other studies. The pooled result for the fixed effect model is OR = 0.68 (0.66 to 0.69), and random effects OR = 0.75(0.64 to 0.88) for the reduction in odds of CHD mortality for an absolute difference of 0.87 mmol/l in serum cholesterol level. Hence, there is considerable difference between the point estimates and corresponding 95% confidence intervals. A forest plot of the studies, together with the results from the random effect analysis is shown in Figures 4.3.

Figure 4.3 Forest plot of random effects meta-analysis combining the 5 cohort studies



Clearly study C4, is by far the biggest and as it has an extreme result, it has a large influence on the analyses. The test for heterogeneity is highly significant (Q=116.6 $p<0.001$). Table 4.8 provides the weightings for each of the studies in each analysis.

Table 4.8 Weightings given to each study in the fixed and random effect analysis

Study	Weights (%)		Study Estimate - OR (95% CI)
	Fixed	Random	
C2	434.0 (5.3)	30.4 (19.8)	0.86 (0.81-0.92)
C4	5917.2 (72.3)	32.5 (21.1)	0.73 (0.72-0.73)
C5	307.8 (3.8)	29.5 (19.2)	0.85 (0.79-0.91)
C7	330.6 (4.0)	29.7 (19.3)	0.80 (0.74-0.85)
C9	1189.1 (14.5)	31.8 (20.7)	0.91 (0.88-0.94)

It is very instructive to see how the relative weightings differ between the two analyses. The between study variance in the random effect analysis is estimated to be 0.031, and in this analysis the studies are given almost equal weighting. This contrasts starkly with the fixed effects analysis, where study C4 gets 72% of the weight.

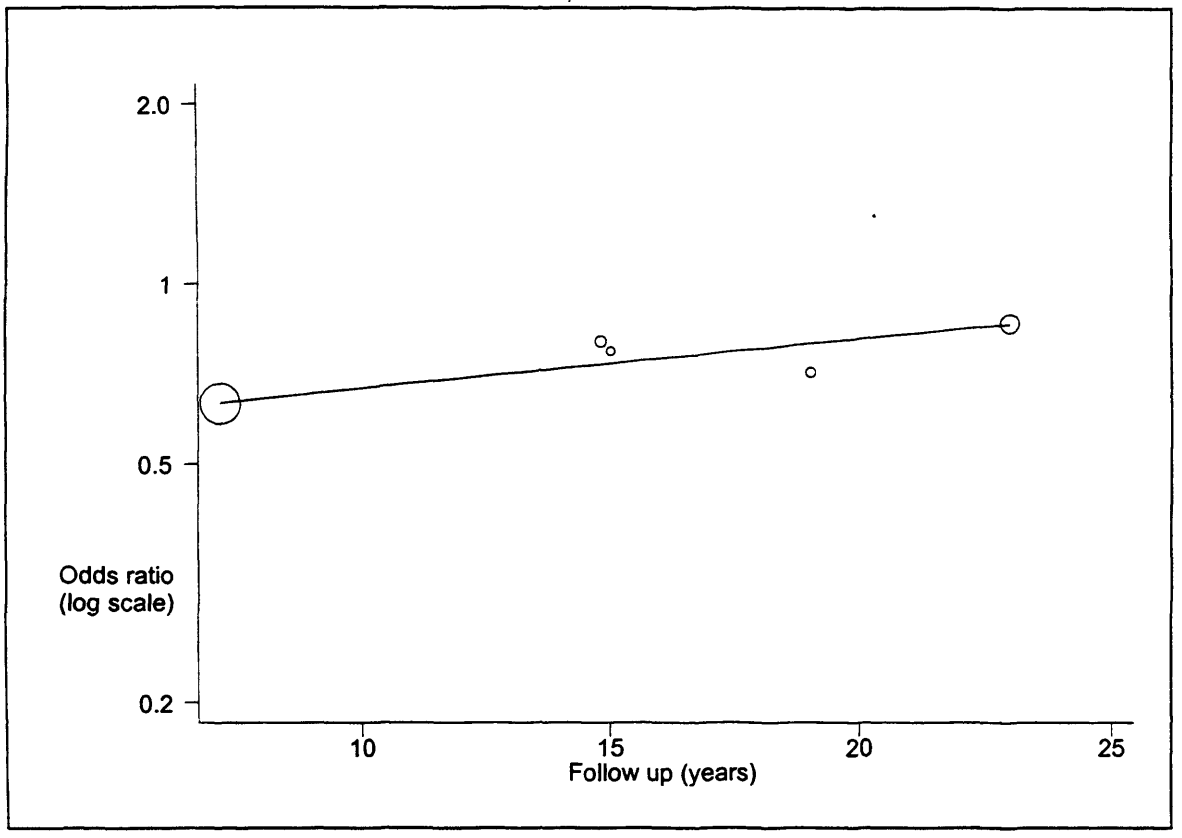
If study C4 is removed from the analysis, the point estimates for the fixed effects model is 0.81 (0.78,0.85), and random effect model 0.79 (0.73,0.86). Hence, the point estimates then both suggest a diminished effect, and are in broader agreement. It is interesting to note that all odds ratios are in the range 0.73 to 0.91, and what may appear from a clinical perspective are similar results, due to the large size of many of the studies, are deemed to be statistically heterogeneous.

It is difficult to say why the result for study C4 (MRFIT screenees) is lower than others. Possible explanations are that it had the shortest follow up time, and the youngest patients of the five studies examined. Interestingly, the unadjusted and adjusted estimates produced by Law et al. (M28) for the outcome ischaemic heart disease events for this study were not as extreme as the unadjusted CHD mortality one derived here.

Although exploring between study heterogeneity is difficult here, due to the small number of studies, the effect of length of study follow-up was explored using a mixed-effect regression model. (Thompson et al. 1997) The resulting regression line is provided in Figure 4.4. The size of the plotting symbol is proportional to the precision of the study estimate.

The equation of the line on figure 4.4 is: $\ln(OR) = -0.60 + 0.02$(follow-up years). The covariate is highly statistically significant (<0.01), and hence the possibility of an association cannot be ruled out. However, due to the small number of data points, this result should be treated with extreme caution.

Figure 4.4 Plot of five studies together with regression line suggesting an association between outcome and length of follow-up



4.8.1 Summary of the cohort study meta-analysis

There is heterogeneity between the five studies included in this analysis. All studies suggest low cholesterol levels are associated with lower death rates from CHD. The difference between the combined results obtained with fixed and random effect models is quite large. A mixed regression model suggests the possibility of an association between outcome and follow-up, with studies with shorter follow-up producing greater effect sizes.

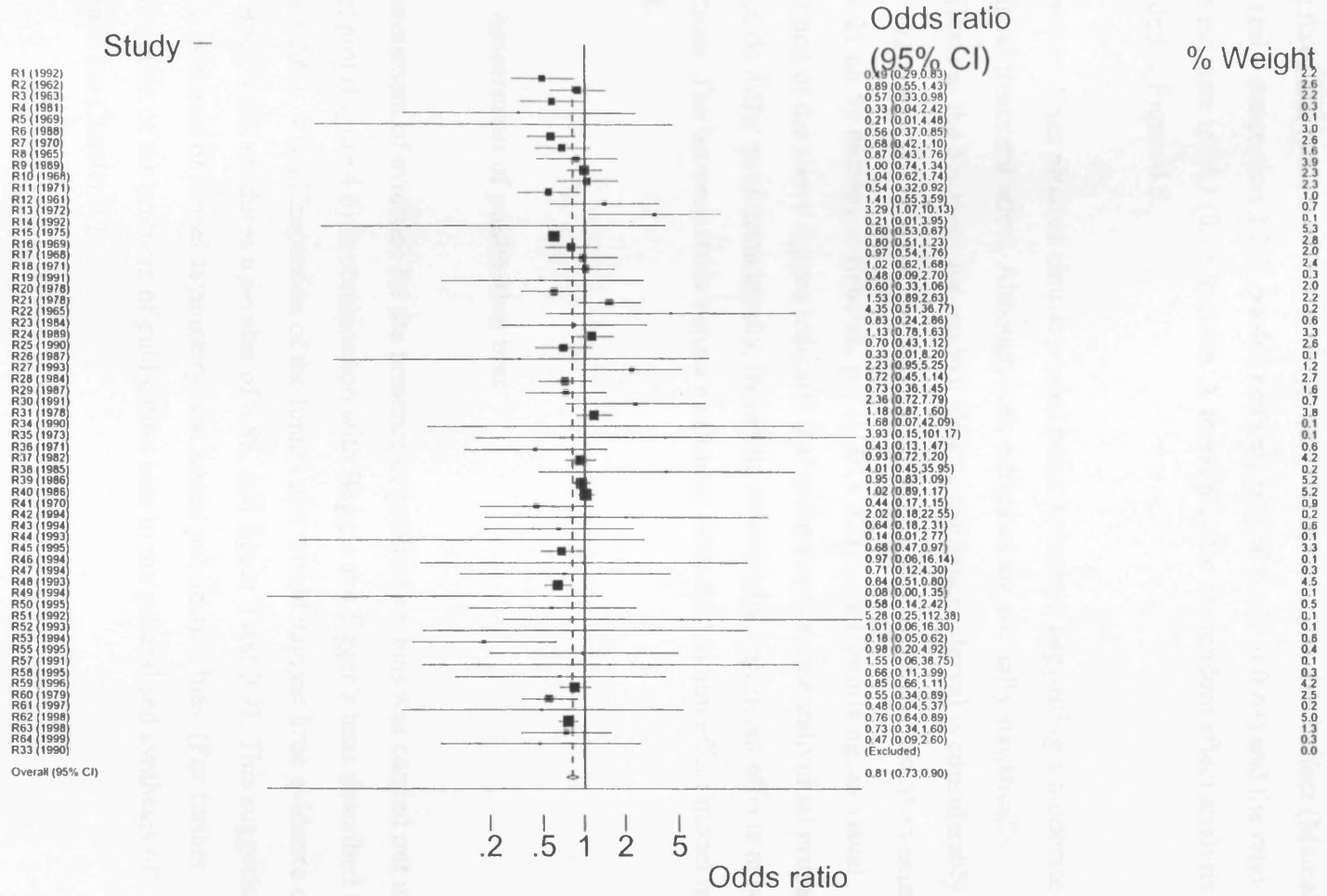
4.9 Combining the RCTs separately

Previous meta-analyses of the cholesterol lowering RCTs (only) are updated here, including more recent studies and using more sophisticated statistical methodology than used previously.

Of the 64 RCTs initially identified, 60 reported CHD mortality (the outcome chosen for compatibility with the cohort studies) and were included in the analysis (studies R32, R46, R54, and R56 are excluded). Although CHD mortality is the primary outcome of interest, data were also available on all cause mortality for all 60 studies (plus an additional study (R54)), and analyses on both endpoints were carried out. Only the CHD mortality analysis is presented below because in most instances both analyses produced similar findings. This is to be expected as CHD mortality makes up a large component (approximately 50% for the control groups) of all cause mortality in the trials. However, where differences between results of the CHD and total mortality analyses were considerable this was noted in the text and results from both analyses displayed. Consideration is given to a model which can include both outcomes in Section 5.2. Although not a primary aim of the analysis, examining both endpoints may provide insight into why previous meta-analyses have found significant beneficial treatment effects when considering CHD mortality, but no such effect for all cause mortality.

The trials varied in size very considerably, from trial R35 with 52 patients to trial R39 with 57,460 patients. Section 4.5.2 described potential covariates which may explain heterogeneity between trials; data on all of these could be extracted for the majority of trials, but inevitably there are missing values. These are clearly noted in the analyses which follow (and in the dataset in Appendix A.V).

Figure 4.5 Random effect analysis of 60 RCTs estimating the effect of cholesterol reduction on CHD mortality



4.9.1 Overall pooled results

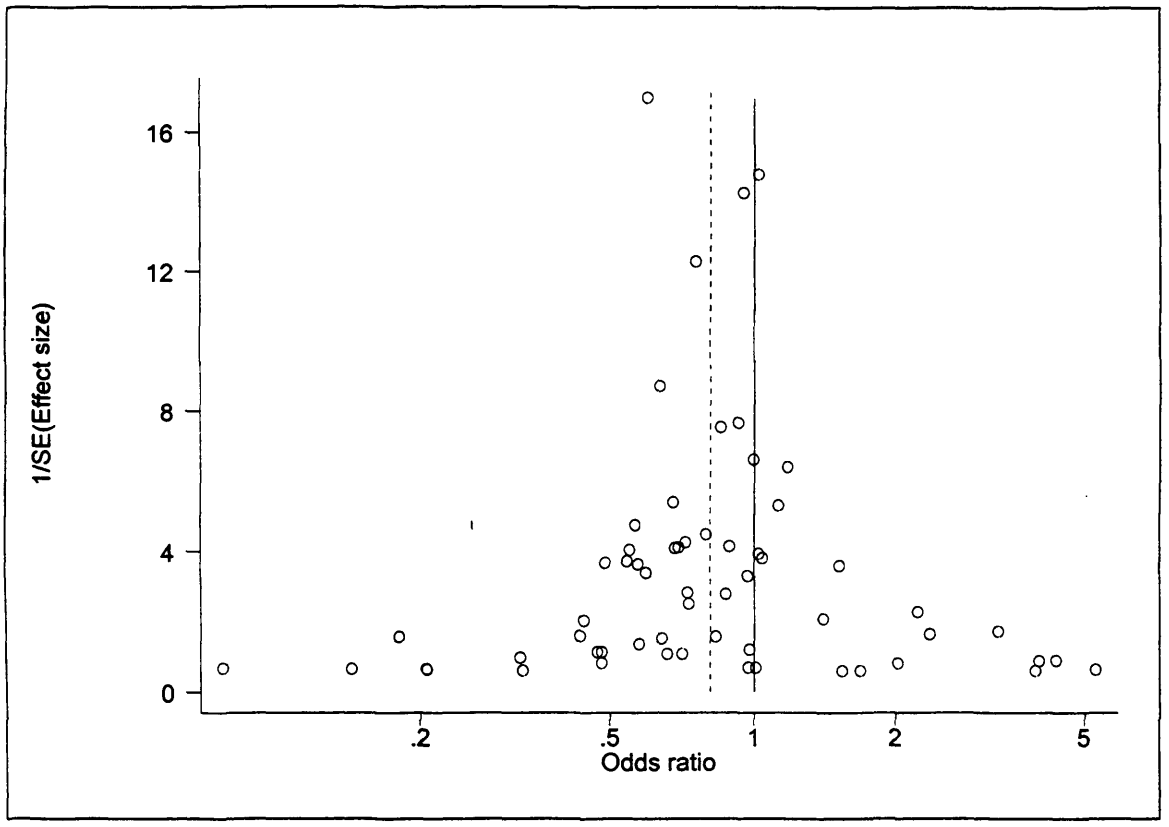
Treatment effects from each trial were calculated on the (log) odds ratio scale, and this scale is used for synthesis throughout the analysis. As a starting point, analyses using fixed and random effects models were carried out. The fixed effect (Mantel-Haenszel – see section 2.2.3) pooled estimate is 0.80 (0.76 to 0.84) and the random effect estimate is 0.81 (0.73 to 0.90). A forest plot for the random effect analysis is provided in Figure 4.5.

The two analyses produce similar pooled point estimates, suggesting a moderate beneficial treatment effect. Although both estimates are formally statistically significant at the 5% level, the random effect confidence interval is considerably wider than the fixed effect one. Considerable heterogeneity between studies exists (Q = 124.21 on 59 degrees of freedom, $p < 0.001$). This is not surprising, as visual inspection of the above figures indicates that estimates from the individual primary studies do differ quite dramatically, including statistically significant effects in both directions. The between study variance estimate from the random effect model is 0.048.

4.9.2 Assessment of publication bias

An assessment of evidence for the presence of publication bias was carried out using a funnel plot (Figure 4.6) in combination with Begg's and Egger's tests described in Section 2.6.1. Visual inspection of the funnel plot would suggest little evidence of bias. Begg's test produces a p-value of 0.88, and Egger's test 0.91. This suggests there is little evidence of funnel asymmetry, and hence publication bias. (For further consideration of the problem of publication bias in the generalised synthesis of evidence see Chapter 6.)

Figure 4.6 Funnel plot of CHD mortality odds ratio estimates from the 56 RCTs



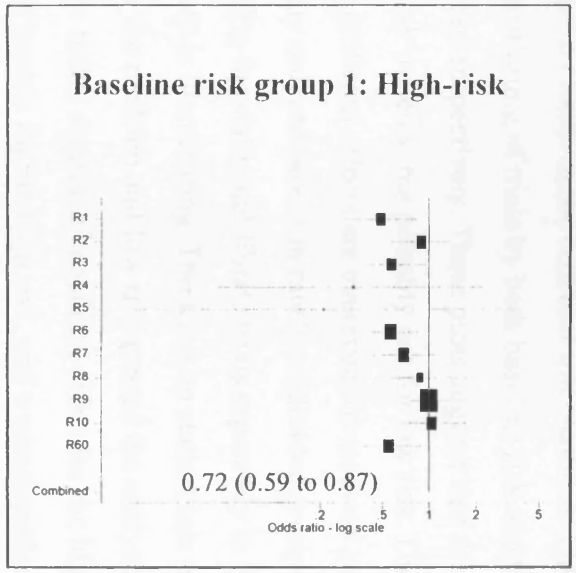
4.9.3 Subgroup analyses

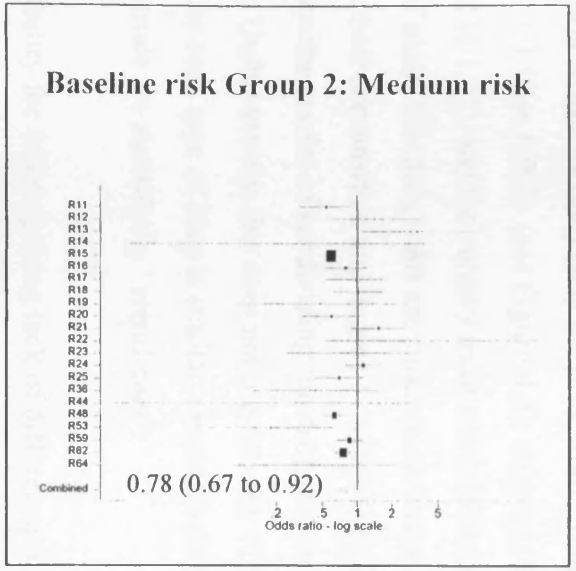
The RCTs can be partitioned and combined in several "natural" subgroups; the results of which are displayed below. If results differ between subgroups, this could suggest factors which have an influence on the treatment effect and explain heterogeneity between trials. Random effects estimates are presented on all plots since residual heterogeneity exists within subgroups in many instances.

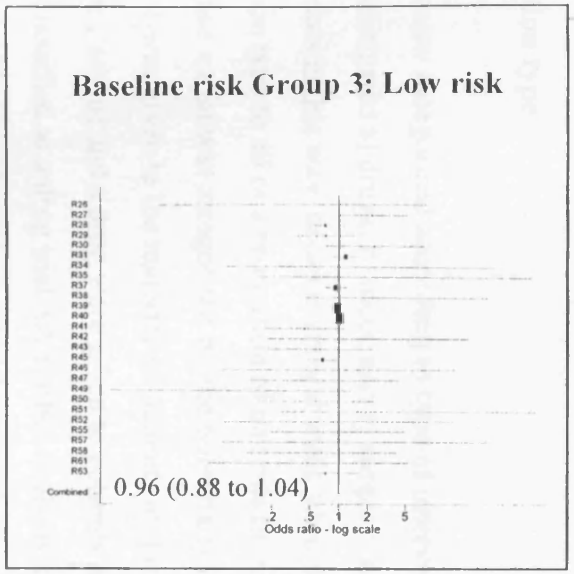
Baseline Risk

The first subgroups considered are defined on the average baseline risk of patients in the control group - as defined by equation 4.1. The RCTs have been split into three groups: group 1(high risk) consists of trials in which there were greater than 50 CHD deaths per 1000 person-years in the control group; group 2 (medium risk) between 50 and 10 CHD deaths per 1000 p-y; and group 3 (low risk), less than 10 CHD deaths per 1000 p-y. These were the group definitions originally used by Sheldon et al. [M25] The result obtained here is similar to that found previously, [M25] with apparently different effects in the three groups (Figure 4.7). For high risk patients, the pooled odds ratio suggests a (clear) benefit of cholesterol lowering (OR = 0.72(0.59 to 0.87)). For patients at moderate risk, the benefit is still substantial and statistically significant, but only marginally so (0.78(0.67 to 0.92)). Interestingly, in patients defined as at low risk, there would seem to be little or no beneficial intervention effect (0.96(0.88 to 1.04)), and the possibility that the treatment may actually be detrimental cannot be ruled out. The heterogeneity test produces p-values of 0.17, <0.001, and 0.45, for high, medium, and low groups respectively - implying that heterogeneity still exists between studies and baseline risk does not account for all between study variation, though this analysis would suggest it did account for a proportion of it.

Figure 4.7 Subgroup analyses of CHD mortality endpoint by baseline risk of CHD mortality in the control group







Intervention type

Trials are now categorised according to type of intervention used, initially as trials which investigated a) drugs, b) diets, and c) surgery. A problem arises when trying to classify trials in this way, because several trials administered more than one intervention type to all or a proportion of patients in the treatment group. When this was the case a trial was categorised by the potentially most invasive treatment, provided it was given to the majority of patients in the active group. For example, if both dietary advice and a drug were given to patients in the treatment arm, then the trial was classified as a drug trial. (A further analysis investigating if there are systematic differences between trials administering single or multiple interventions is reported later in this section) Only the drug trials produced a "significant" treatment effect (0.79(0.70 to 0.90)) (see Figure 4.8), although the diet trial result was marginal (0.85(0.72 to 1.01), and the surgery trials result is much less clear due to the small number of trials. The drug trials are subdivided further, as the drugs used varied widely in their composition and degree to which they lower cholesterol levels. Figure 4.9 displays the results of subdividing the drug studies into fibrates, hormones, statins, and other. Unfortunately, this does not provide any further insight - as the pooled estimate for each type of drug is similar, ranging between OR=0.74 to 0.86, and only the statin trials are statistically "significant".

One possibility for this surprising lack of difference in effect size between different treatment regimes is that the effect of baseline risk, which is clearly large, is masking differences in effects of the different interventions. Figures 4.10 and 4.11 display a further partitioning of trials by both baseline risk and treatment type, for drug and diet interventions respectively. These plots suggest that drugs are effective for high and medium risk patients, but possibly not for low risk. Diets are effective for high risk, but non-significant effects are observed for medium and low risk patients; not surprisingly reduced power in much subdivided groups is impeding clear conclusions. Investigating the statin and fibrate trials separately in Figures 4.12 and 4.13 respectively is illuminating. There are no statin trials with patients in the high risk group, for the medium and low risk groups the analysis suggests statins are beneficial. The fibrate trials, suggest a beneficial effect for the high-risk patients, a non-significant benefit for medium risk, and a non-significant harmful effect for low risk.

These are very interesting findings but potentially open to familiar problems of interpretation of subgroup analyses stemming from multiple estimates/testing and lack of power. This analysis is extended via regression analysis, where baseline risk is also modelled as a continuous covariate in Sections 4.9.4 and 4.9.5.

Figure 4.8 Subgroup analysis partitioning the trials by type of intervention administered

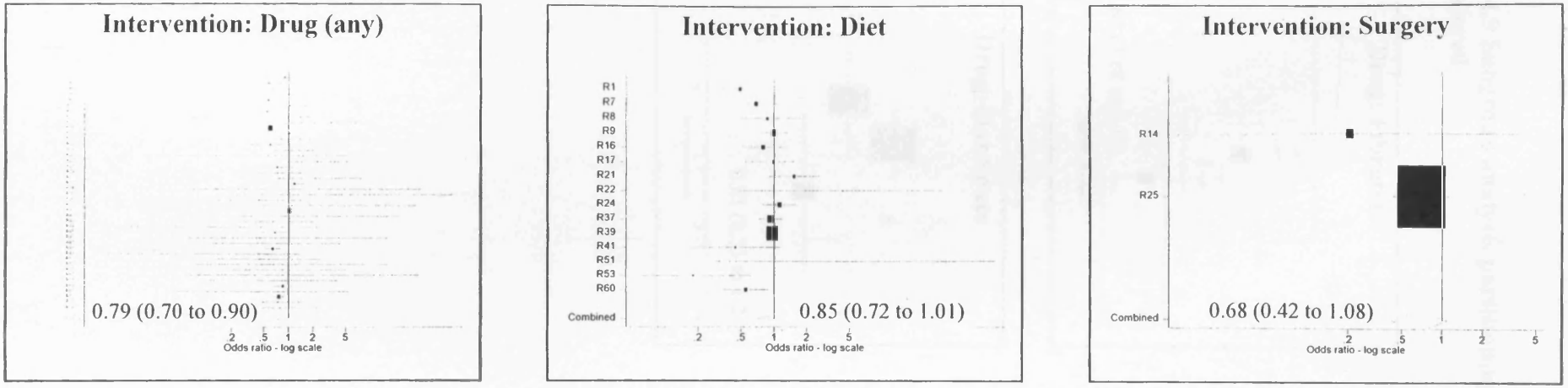


Figure 4.9 Subgroup analysis partitioning the drug trials further, by type of drug administered

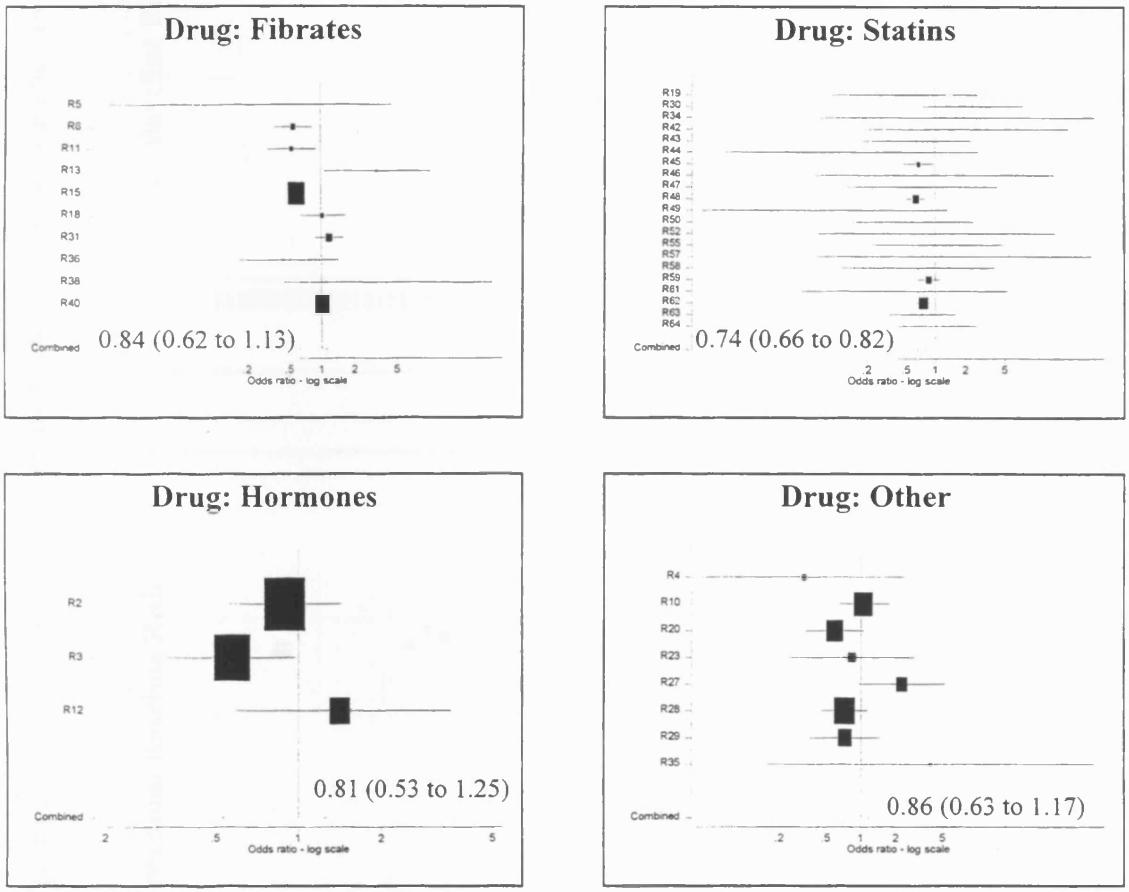


Figure 4.10 Subgroup analysis of CHD mortality endpoint partitioning the *drug* (any) intervention trials by baseline risk of CHD

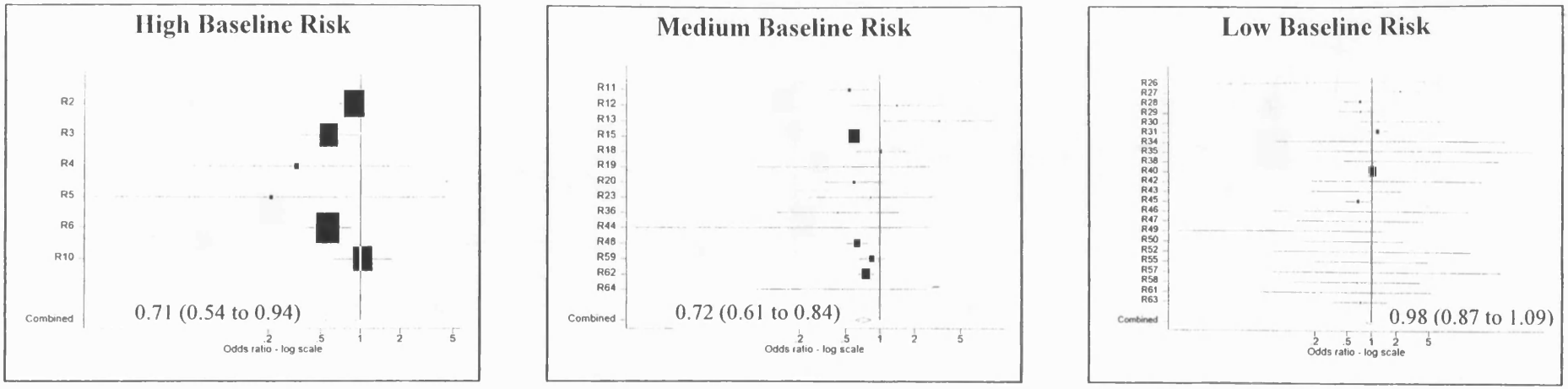


Figure 4.11 Subgroup analysis of CHD mortality endpoint partitioning the *diet* intervention trials by baseline risk of CHD mortality

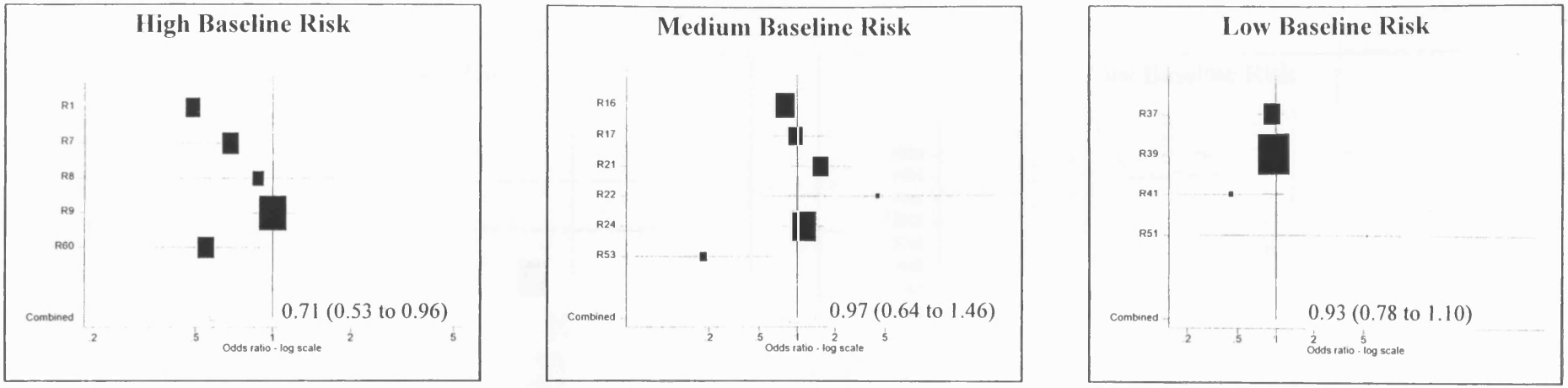
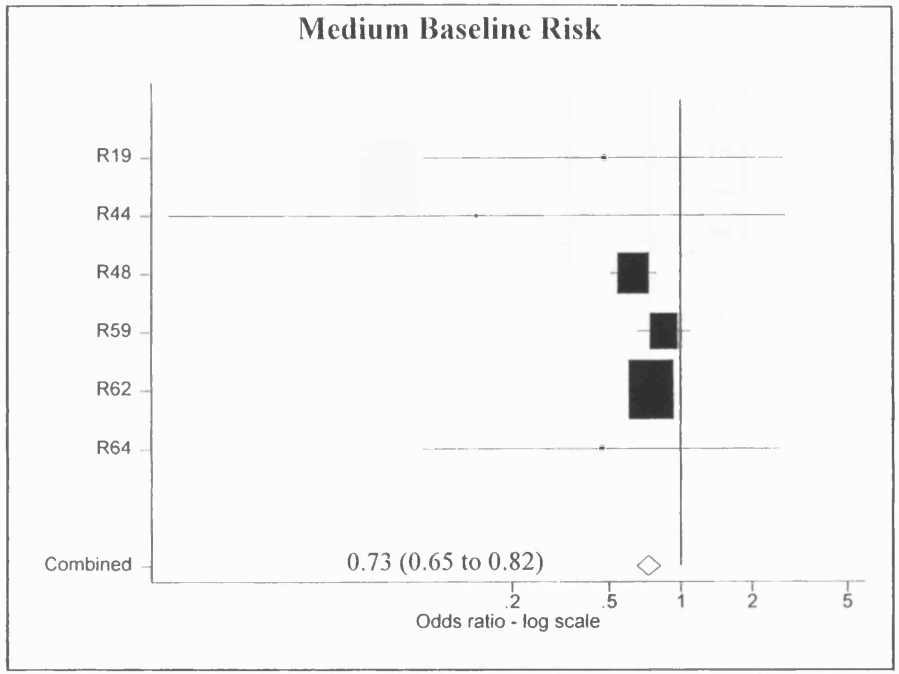


Figure 4.12 Subgroup analysis of CHD mortality endpoint partitioning the *Statin (drug)* intervention trials by baseline risk of CHD mortality



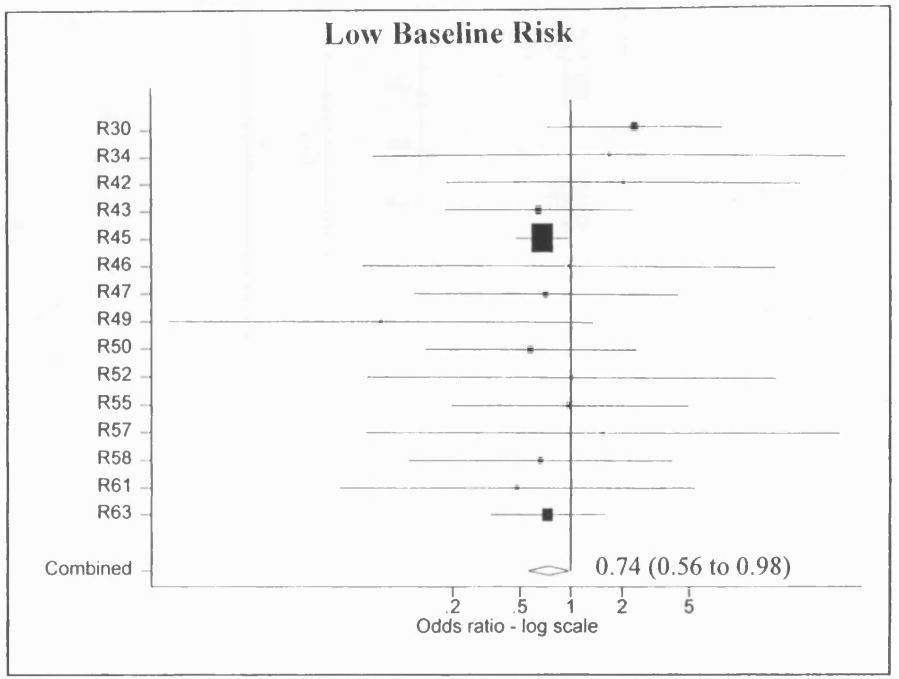
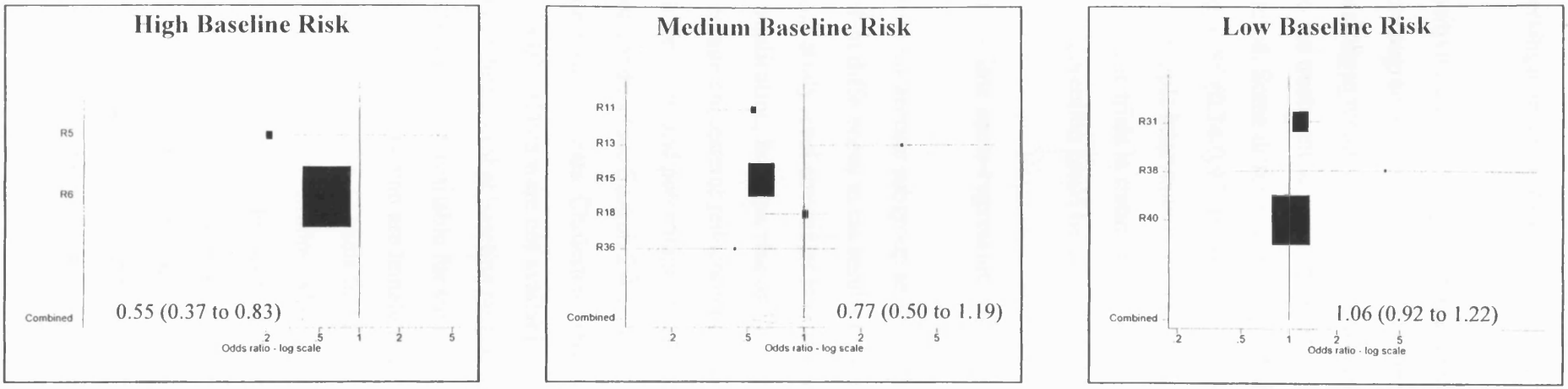


Figure 4.13 Subgroup analysis partitioning the *Fibrate (drug)* intervention trials into primary and secondary prevention studies



Multiple/single intervention

As a sensitivity analysis of the initial categorisation of trials by intervention type, trials were segregated into those which applied a single intervention in the treatment group, and those which administered multiple, for example several trials give dietary advice to the treatment arm, and then, if this was not effective, drugs were administered. Some difference in the pooled estimates and confidence intervals were observed (0.85 (0.74-0.97) for single intervention group compared with 0.75 (0.64 to 0.89) for multiple intervention group). Since the most common combination of treatment in the trials is statin drug and dietary advice, the apparent benefit of multiple intervention could be confounded by statin treatment.

4.9.4 Univariate meta-regression analysis

In the previous section subgroup analysis was used to examine discrete factors which may explain differences in the results between studies. The effect of several continuous study level covariates is examined using meta-regression in this section. Year of publication, baseline risk of CHD in the control group, length of trial follow up, percentage cholesterol reduction between groups achieved on average, baseline serum cholesterol, and percentage female enrolled – are examined using mixed meta-regression models (see Section 2.4.2). Follow-up ranged from ten weeks to ten years, with a mean of 3.7 years. Cholesterol reduction ranged from 0.03% to 32.2%, with a mean of 14.0%. (Data were not available from trials R2, R5 or R39 on cholesterol reduction.) Cholesterol at baseline ranged from 5.4mmol/l to 9.6mmol/l, with a mean of 6.6. (Data was not available for trials R2 and R5 on baseline cholesterol.) The percentage of patients who are female ranges from 0 to 71%, with a mean of 14.1% (26.5% female for those 32 trials that included some female subjects). Twenty-eight of the trials included no females, while data were not available on gender composition from trials R22 and R36. The covariate values for each RCT are provided in Appendix A.V. The Stata macro *metareg* (Sharp, 1998) using restricted maximum likelihood (REML) fitting procedures was used together with the SAS PROC MIXED procedure for the analysis. The problems of modelling baseline risk (see Section 2.4.5) are ignored in this section, but the Bayesian analysis that follows (Section

4.9.6) rectifies this shortcoming. Table 4.9 provides the results of including each of these covariates individually in the meta-regression model.

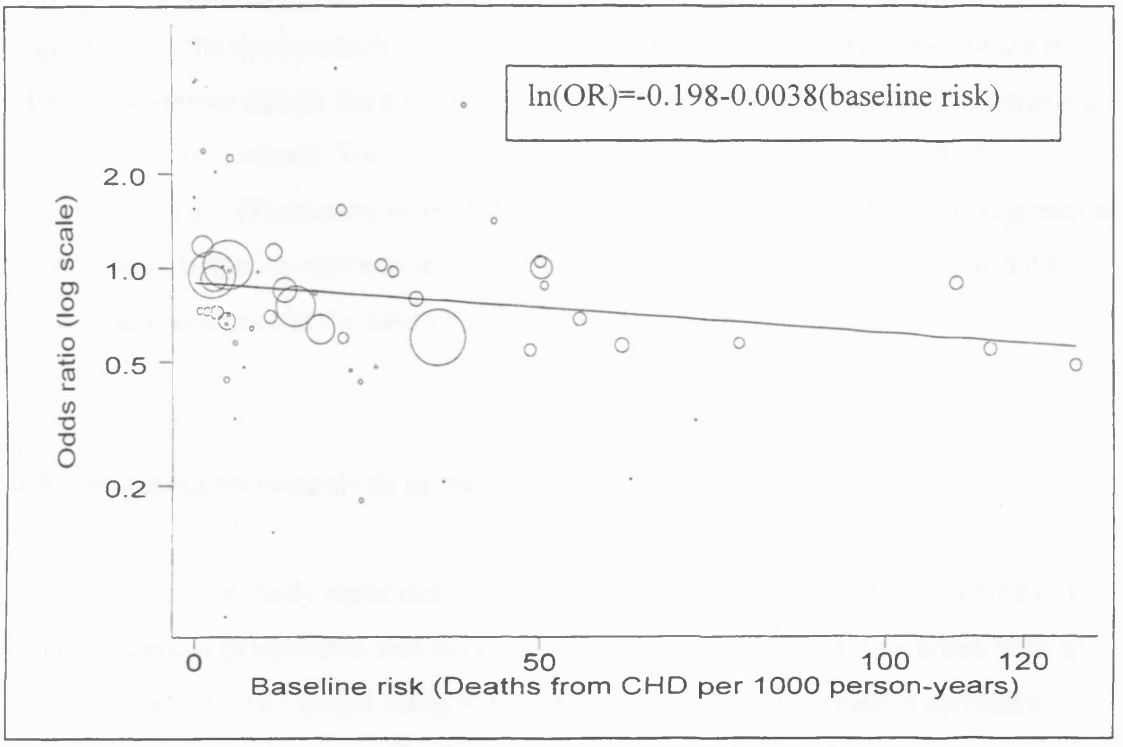
Table 4.9 Results of univariate meta-regression analyses (on log odds ratio scale)

Covariate	*Intercept (95% CI)*	*Slope (95% CI)*	*P-value for slope coefficient*	*Tau-squared*
Year of publication	5.51 (-13.0 to 24.0)	0.003 (-0.01 to 0.01)	0.55	0.041
Length of follow-up	-0.22 (-0.46 to 0.17)	0.002 (-0.04 to 0.04)	0.94	0.041
% reduction in cholesterol	-0.18 (-0.37 to 0.01)	0.004 (-0.10 to 0.02)	0.60	0.042
Baseline cholesterol	-0.24 (-1.28 to 0.80)	0.004 (-0.16 to 0.17)	0.96	0.043
% Female	-0.20 (-0.32 to –0.09)	-0.001 (-0.008 to 0.006)	0.83	0.040
Baseline risk in control group	-0.11 (-0.23 to 0.01)	-0.004 (-0.007 to –0.001)	0.01	0.027

Little evidence of linear variations existed with any of the continuous covariates except baseline risk (p= 0.01). Its effect was consistent with that observed in the subgroup analyses described previously, cholesterol lowering being more effective at preventing CHD mortality in higher risk groups. Figure 4.14 displays a regression plot of the studies by baseline risk, the size of the plotting circle being proportional to the precision of the study effect estimate. The regression line resulting from fitting baseline risk (centred at 24.3 - the mean value of the 61 studies used in the model) is plotted on this graph, and suggests, for people at lowest risk of CHD mortality, little benefit is to be gained from lowering cholesterol levels, however the populations at

greatest risk suggest a reduction in odds to nearly 0.5 by lowering cholesterol levels. The value for the between study variance, having adjusted for baseline risk is 0.027, which can be compared to the value obtained from the random effects model (0.048) at the beginning of the analysis. Including this covariate has reduced the between study variance by around 44 percent.

Figure 4.14 Scatter plot of studies by baseline risk in the control group, and mixed (weighted) regression model fitted to the data



4.9.5 Multivariate meta-regression analysis

In order to explore the combined effect of covariates, examined using subgroup analyses and regression above, a multivariate mixed model was employed. Specifically, it is of interest whether different interventions have different effects at different baseline risk levels, and hence interactions between baseline risk and intervention type were explored. Using both the broad categorisation of interventions

into drug, diet and surgery, and the more detailed categorisation subdividing the drug trials further into statins, fibrates, hormones and others, little evidence of baseline risk × intervention interactions existed. However, since such analysis will have very low power for detecting such interactions, (Lambert et al. 2001a) such findings should not be over interpreted.

4.9.6 Drawback of above regression modelling

It has been mentioned previously that this analysis on baseline susceptible to regression to the mean, which has the potential to make the model over estimate the effect of baseline risk. In the next section, parts of the RCT analysis are implemented using Bayesian methods. This includes the implementation of the method of Thompson et al., (Thompson et al. 1997) which overcomes the problem of regression to the mean. Bayesian methods are also used to combine the observational and the randomised evidence in the next chapter.

4.9.7 Bayesian meta-analysis of the RCT data

This section, essentially replicates as a Bayesian the analysis previously carried out from a classical perspective, and compares the results obtained. It concludes with a full Bayesian MCMC model using WinBUGS in which baseline risk is correctly modelled. Vague priors were fitted to all parameters of interest. Various convergence diagnostic tests and plots available in the CODA software (Cowles et al. 1994) were used to assist with the decision of when convergence has been reached.

Random effects model

Initially, a model of the form outlined by equation 2.18, implemented in BUGS as described by Smith et al. (Smith et al. 1995) is fitted to the data. This models the events "directly" using a binomial distributions. Vague priors were placed on all

required parameters. A Normal $(0,10^6)$ distribution was placed on the pooled log odds ratio; a Normal $(0,10^5)$ distribution on each study's individual log odds ratio; and an Inverse-Gamma(0.001,0.001) distribution on the between study heterogeneity parameter. A run of 40000 iterations was carried out; the first 10000 were designated as burn in and discarded; leaving the remaining 30000 to be used for estimating parameters. Examining the diagnostic output provided by CODA, it would appear that convergence was achieved, and the chain had been run long enough to provide accurate estimates.

The median estimate of the odds ratio was 0.81, with 95% Credible Interval (CrI) (0.72 to 0.89), and between study variance of 0.047 with 95% CrI(0.016 to 0.130). Thus the pooled estimate is very similar to that obtained from the classical random effects model (OR = 0.81 (0.73 to 0.91)). The slight discrepancy observed could be due to the need to use continuity correction factors for the Classical model, in addition to the different model specifications and method used to evaluate the model.

Including baseline risk

The random effects model described above forms the basis to which the extension of including a covariate for baseline risk can be added. As previously mentioned, the method of Thompson et al. (Thompson et al. 1997) is used to circumvent the problem of regression to the mean. The model used is outlined in equation 2.21. In order to use this method, a modification in definition of baseline risk is necessary. This has the drawback over the previous definition is that it does not take into account the length of follow up in each study.

In order to estimate the effect of regression to the mean in this example, a mixed effect model including a baseline risk covariate as described in equation 4.2 was also fitted using the standard Bayesian mixed regression model (equation 2.20).

In both models the covariate for baseline risk (μ) is centred, i.e. its mean value across trials (-2.6) is taken away from each trial's value. This is done to reduce correlation

between parameters in the simulation procedure. The same priors as used above were used for this analysis, with the addition of a Normal $(0,10^6)$ prior distribution placed on the regression slope parameter in both models.

For both models, a burn in of 10000 iterations was used, with results being derived from a run of a further 20000 iterations. Parameter estimates from these two models, accounting and not accounting for regression to the mean, are displayed in Table 4.10.

Table 4.10 Parameter estimates for models adjusting and not adjusting for regression to the mean

Model parameter estimate (95% CrI)	Model not accounting for regression to the mean (equation 2.20)	Model accounting for regression to the mean (equation 2.21)
Baseline risk (centred at -2.612)	-0.102 (-0.170 to –0.026)	-0.103 (-0.170 to -0.030)
Log(OR)	-0.150 (-0.253 to –0.049)	-0.149 (-0.253 to -0.050)
Between study variance	0.024 (0.004 to 0.092)	0.024 (0.004 to 0.091)

From these estimates two regression models (removing centring) can be derived:

Not accounting for regression to the mean

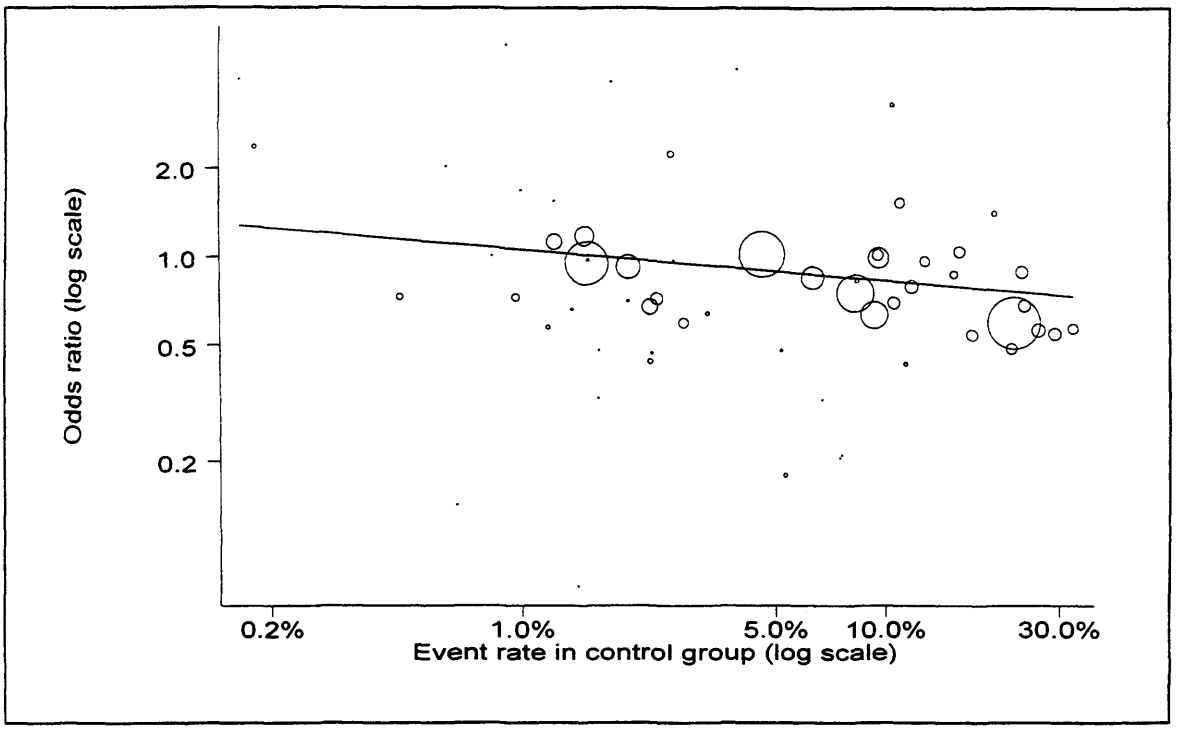
$$\text{Log(OR)} = -0.418 - 0.102 \times \text{Log(Baseline risk)} \qquad (4.4)$$

Accounting for regression to the mean

$$\text{Log(OR)} = -0.418 - 0.103 \times \text{Log(Baseline risk)} \qquad (4.5)$$

The results from the two models are almost identical suggesting that regression to the mean does not pose a serious problem in this analysis. The adjusted regression line together with the 60 RCTs is plotted in Figure 4.15. Note this figure has a different x-axis scale than 4.18 which reported the classical analysis, this is due to the different definitions of baseline risk used. It can be seen that, after adjusting the coefficients accordingly, a strong relationship still exists between outcome and baseline risk. The appearance of Figure 4.15 is quite different from the classical equivalent (Figure 4.14), which can be explained by the fact that 4.15 is on a log scale and the definition does not take length of follow-up into account. However, it would appear that the results are qualitatively robust to baseline risk definition.

Figure 4.15 Baseline risk regression line for 60 RCTs adjusting for regression to the mean



4.10 Summary/discussion

This chapter has considered the topic of cholesterol reduction and its impact on CHD and overall mortality. Over forty meta-analyses have been carried out on the RCTs in this area previously, and have produced contradictory results. Only one of these previous meta-analyses considered relevant observational evidence but did not combine it with the RCTs. The issue of expressing the results of observational studies in a compatible form with the trials is considered and a solution described in detail. A meta-analysis of the observational evidence is then reported which indicates a strong association between cholesterol levels and incidence of CHD mortality. Then an updated meta-analysis of the RCTs using more sophisticated regression methods than has been done previously is described which accounts for regression to the mean when including patient baseline risk as a covariate in a meta-regression model. There is some difficulty comparing the results of the classical and Bayesian regression analysis since, by necessity, a simpler definition of baseline risk, not taking into account length of

study follow-up is used in the Bayesian analysis, but a comparison between Bayesian models suggests regression to the mean was not great in this example. Development of a Bayesian method that does allow length of follow-up to be included in the baseline risk definition while avoiding regression to the mean would be desirable.

Although different combinations of the RCTs had been meta-analysed before, this was perhaps the most all-encompassing analysis as it attempted including trials using all evaluated modes of reducing cholesterol. This analysis suggested that treatment effectiveness is related to patients' risk of disease. The findings of the Bayesian analysis are similar to those from the classical model, since the adjustment due to regression to the mean was small.

Although these are important findings, it has been noted previously that it is difficult to define an individuals' baseline risk, and hence the model is limited in its application to individuals. (Sharp et al. 1996) Thompson et al. (Thompson et al. 1997) have suggested that one could develop a prognostic score based on patient covariates from related cohort studies and relate treatment effects to this score for individual patients. (Senn et al. 1996) Such an analysis would remove the need for considering 'underlying risk' directly. They suggest the prognostic score would best be based on data other than that from the trials which form the meta-analysis for treatment effects. Another possibility, if individual patient data are available, is to include them in a hierarchical model with an extra level added, that of individual patents. This is theoretically a simple extension of the models presented here. (Higgins et al. 2001) Chapter 8 considers a model which uses prognostic information provided by related cohort studies to estimate individual patient benefit from treatment. Chapter 5 considers how the data from the RCTs and the cohort studies can be combined while preserving the necessary complexities of this model for the RCTs.

A further drawback of the analysis is that the comparison group in the RCTs changes in the trials. For example may of the early trials had a comparison group given only placebo, while many of the later trials comparison group was given dietary advice, which itself was administered as the experimental treatment in a proportion of the early trials. This issue is not properly addressed here, and would appear to have been glossed over in previous analyses also. A method, which models the intervention given to the

control group, is described in Section 5.2. Additionally, the indirect comparisons model described in Section 5.4 could potentially address this issue.

Chapter 5 Generalised synthesis methods including further analysis of the cholesterol lowering evidence

5.1 Combining the observational and randomised evidence using Bayesian hierarchical modelling

In Chapter 4, meta-analyses of cohort studies reporting levels of CHD for different serum cholesterol values, and RCTs of cholesterol lowering interventions were performed. In this section, a three level hierarchical model, implemented using Bayesian MCMC methods, is used to combine the results from both study types. The first model used is of the form described in Section 3.8.1, where between study within study type and between study type heterogeneity is accounted for. An extension of this model is also fitted which includes the study level covariate baseline risk discussed at length in the previous chapter.

5.1.1 Pooling the RCTs and Cohort studies ignoring covariates

This model could be viewed as a natural extension of the standard meta-analysis random effect model used for combining studies of a single type. (Dersimonian and Laird, 1986) Of course, without this extension, all studies could be pooled as a simple set using the standard random effect model, and ignoring the type of study involved. That would, however, assume all studies to be exchangeable, an assumption which is unrealistic if different study designs have been used. Another possibility would be to include study type as a fixed covariate. This allows the different types of studies to have a different mean effect sizes, but the corresponding variance term would be assumed the same for all study types. The proposed model allows study types to have distinct effect sizes and corresponding variance terms.

The model was fitted to the 60 RCTs and the 5 cohort studies for which the relevant data could be extracted. The code used to fit the model, in WinBUGS is given in Appendix A.VII and the model is expressed algebraically below including the priors specified.

RCT "exact" binomial model

$$r_{cj} \sim Bin[p_{cj}, n_{cj}] \qquad r_{tj} \sim Bin[p_{tj}, n_{tj}] \qquad j = 1.....60$$

$$\text{logit}(p_{cj}) = \mu_j \qquad \text{logit}(p_{tj}) = \mu_j + d_j$$

$$d_j \sim \text{N}[\theta_1, \tau_1^2]$$

$$\mu_j \sim N(0,10^5) \qquad \tau_1^2 \sim IG(0.001,0.001)$$

Cohort study model

$$T_k \sim N[\Psi_k, \sigma_k^2] \qquad k = 1....5$$

$$\Psi_k \sim N[\theta_2, \tau_2^2]$$

$$\tau_2^2 \sim IG(0.001,0.001) \tag{5.1}$$

Pooling both study types

$$\theta_m \sim N(\phi, \nu^2) \qquad m = 1,2$$

$$\phi \sim \text{N}(0,10^6) \qquad \nu^2 \sim IG(0.001,0.001),$$

where nc_j and nt_j are the total number of persons, rc_j and rt_j are the number of CHD deaths, and pc_j and pt_j are the estimated probabilities of events in the treatment and control arms of the jth RCT. μ_j is the estimated ln(odds) of an event in the control group, and d_j is the estimated ln(odds ratio) in the jth group. θ_1 is the estimated pooled odds ratio and τ_1^2 is the between study variance term for the RCTs. T_k is the estimated

ln(odds ratio) and σ_k^2 the estimated variance for the *k*th cohort study. Ψ_k is the model shrunken ln(odds ratio) from the *k*th cohort study. θ_2 is the estimated pooled odds ratio and τ_2^2 is the between study variance term for the cohort studies. ϕ is the overall mean effect of the two populations of studies and ν^2 represents the between study type variance.

This model deviates from the model specified in equation 3.8.1 in that different models are used to combine the RCTs and the cohort studies. The RCTs are pooled using the "exact" model given by equation 2.18, while it is necessary to pool the cohort studies using the simpler model expressed in equation 2.17since odds ratios were derived from regression coefficients for these studies as opposed to estimated directly from comparative group data. Such flexibility is one of the appeals of the MCMC approach.

When this model was originally implemented in 1997, many problems were experienced getting it to run in BUGS. Fortunately, due to improvements in the WinBUGS program these problems were all alleviated when updating the analysis in 2001. Increased computer power allowed long runs of 10,000 burn in followed by 20,000 iterations to be completed relatively quickly. Two such runs using the two sets of staring values for the parameters of interest given below in Figure 5.1 were carried out.

Figure 5.1 Initial values assigned to the MCMC chains used

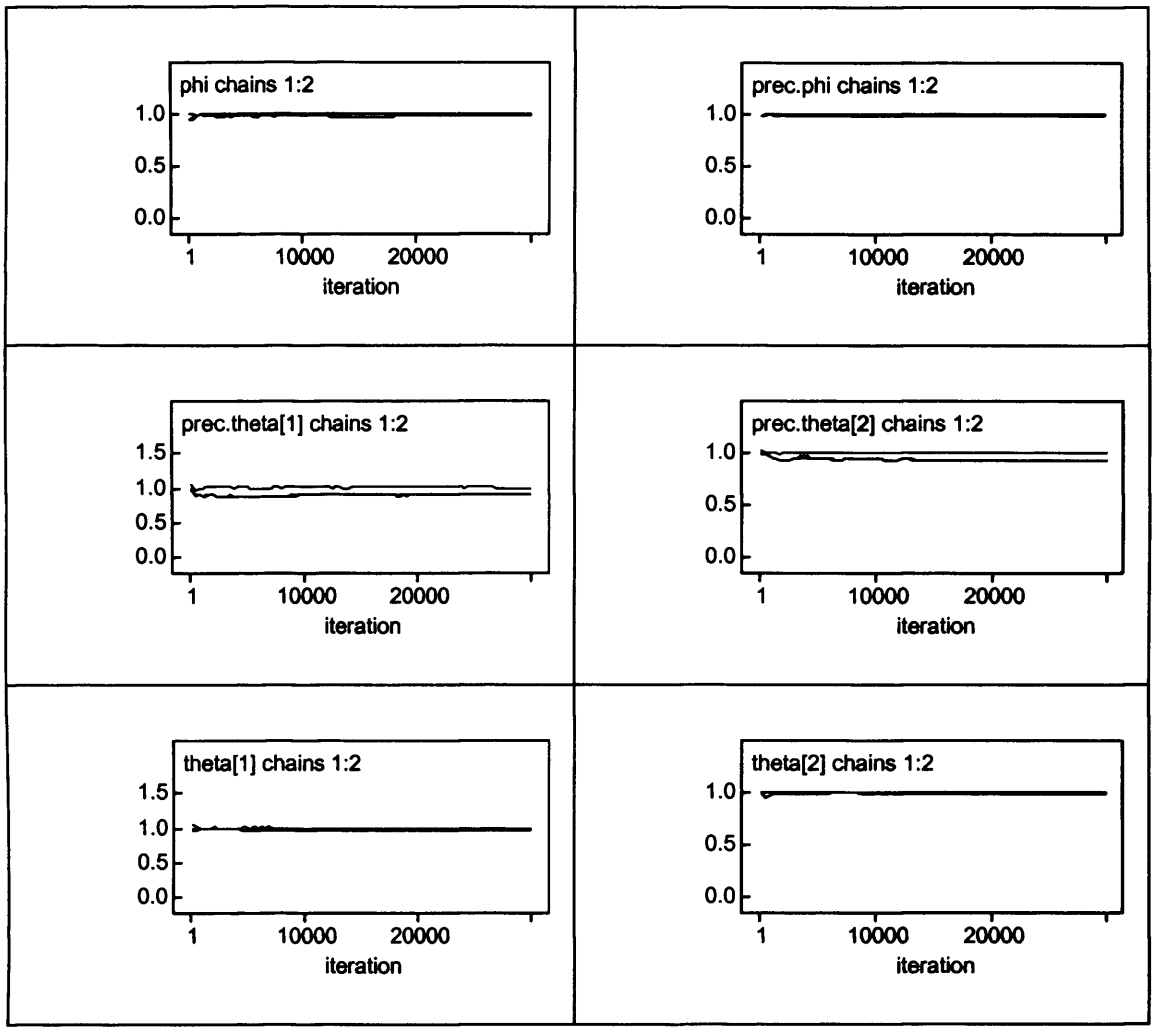
Chain 1

list(prec.theta=c(0.5,0.5), theta=c(0,0), phi=0, prec.phi=1)

Chain 2

list(prec.theta=c(0.1,0.1), theta=c(5,5), mean=5, prec.mean=0.1)

Figure 5.2 Gelman-Rubin convergence plots



The modified Gelman-Rubin convergence statistic (Brooks and Gelman, 1998) is used to assess convergence of the chains. Figure 5.2 displays the associated plots for the Gelman-Rubin statistic for the two chains. The width of the central 80% interval of the pooled runs is green, the average width of the 80% intervals within the individual runs is blue, and their ratio R is red. In the figure both the ratio and the within individual run interval are normalised to have an overall maximum of one.

The convergence of the ratio to 1 and the pooled within interval widths to stability is of interest. In all six parameters displayed in Figure 5.2 it would appear that a convergence of the red line to 1 and stability of the blue line are achieved quickly (confirmed by examining the underlying statistics the plots are based on), suggesting that a burn-in of 10,000 iterations is adequate to be confidence of convergence.

Figure 5.3 Displays plots of the chain histories for the parameters of interest based on the 20,000 iterations used to calculate parameter estimates and make inferences.

Figure 5.3 Plots of the chain histories for parameters of interest

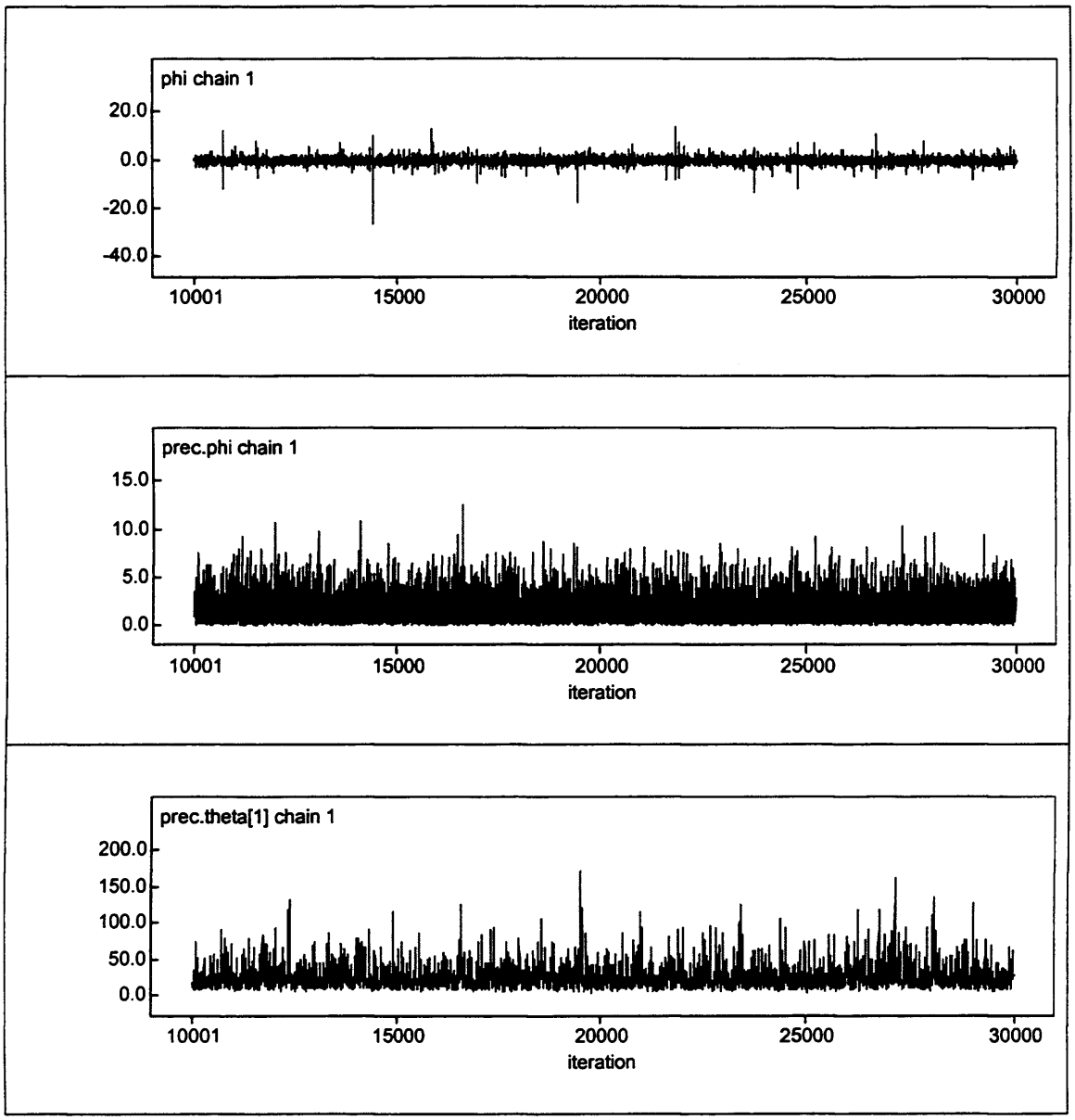
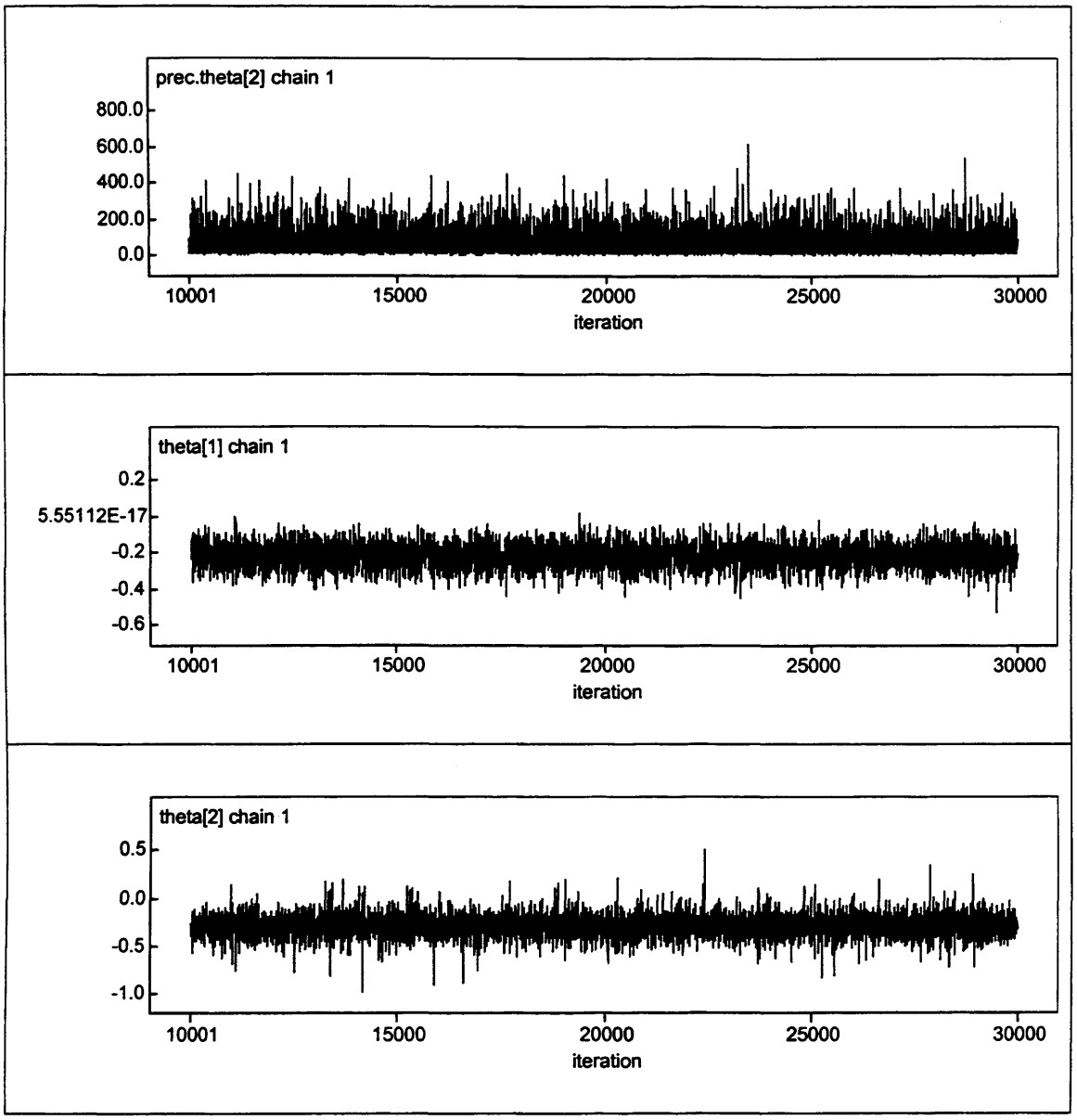
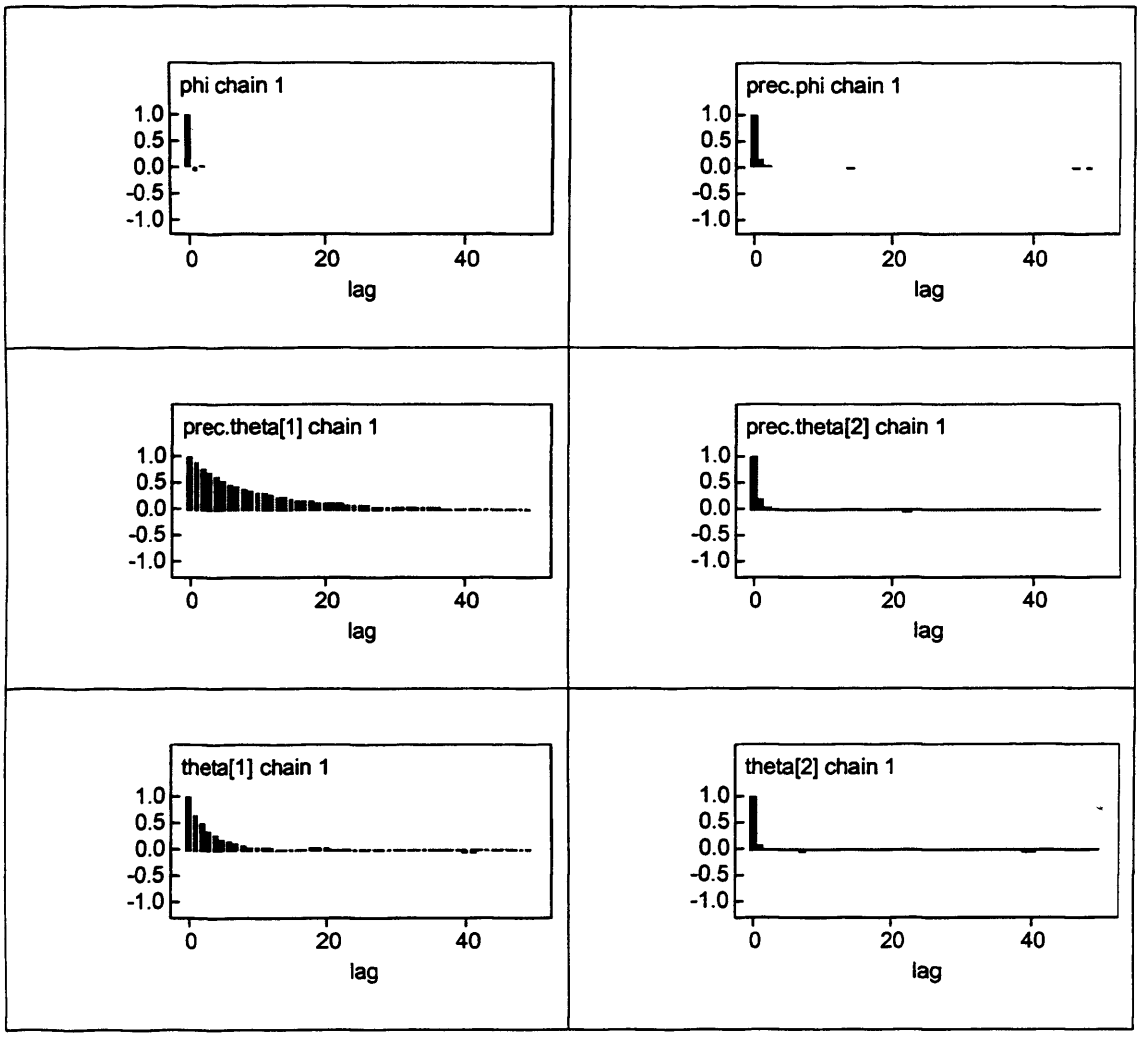


Figure 5.3 (Continued)



The plots in Figure 5.3 are highly "spiky" suggesting that the sampler is moving around the sample space quickly. This is confirmed by examining the autocorrelation plots in Figure 5.4 which informs that the correlations between successive iterations of the sampler are small. The lack of any apparent drift in the history plots also confirms it would appear convergence has been achieved.

Figure 5.4 Autocorrelation plots for parameters of interest



Smoothed posterior kernel densities for the parameters of interest are displayed in Figure 5.5 and summary statistics for these parameter estimates are given in Table 5.1.

Figure 5.5 smoother posterior kernel densities for the parameters of interest

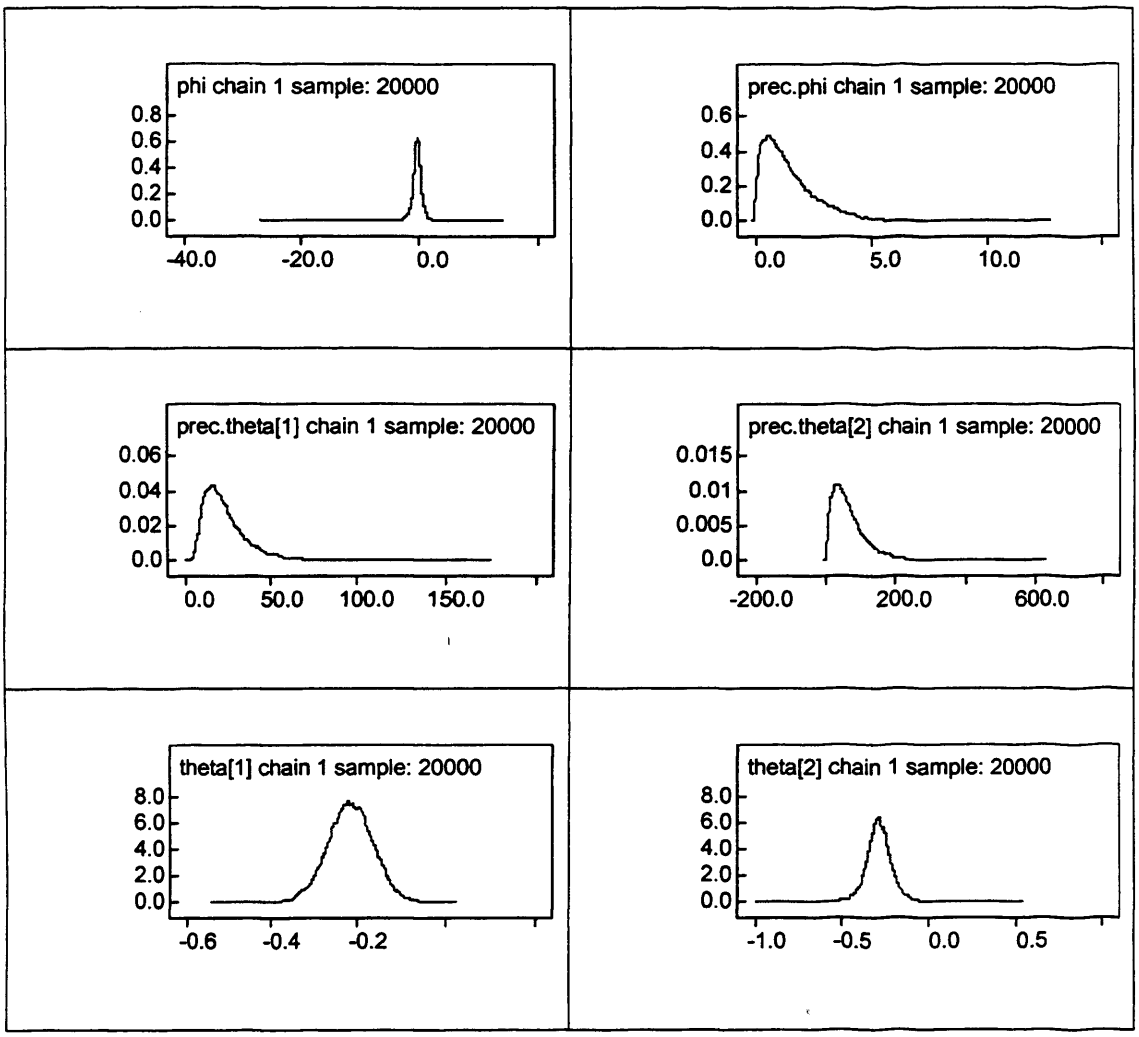


Table 5.1 Transformed estimates and interpretation

Parameter	Interpretation	Median estimate (2.5 & 97.5 percentiles)
e^{ϕ}	Overall pooled OR (pooling the RCT and Cohort study population results)	0.78 (0.40 to 1.53)
e^{θ_1}	OR pooled estimate for RCTs (comparison of treatment and control groups)	0.80 (0.72 to 0.88)
e^{θ_2}	OR pooled estimate for Cohort studies (corresponding to a 0.87mm/l difference in cholesterol)	0.76 (0.67 to 0.87)
v^2	Between study population type variance	0.010 (0.0006 to 5.57)
τ_1^2	Between study variance within RCTs	0.048 (0.017 to 0.128)
τ_2^2	Between study variance within Cohort studies	0.018 (0.005 to 0.111)

These results are presented graphically in the form of a Forrest plot in Figure 5.6. In addition to the pooled estimates for each study type and the overall result, shrunken estimates for each of the individual studies, generated by the model, are plotted. To make a comparison with earlier results, a further plot (Figure 5.7) displaying the results of combining RCTs and Cohort studies individually, combining all the studies using a standard random effect model ignoring study type, together with the results from the hierarchical model on a magnified scale, for closer inspection.

The pooled point estimates for the RCTs, cohort studies and combined analyses from the three-level model differ only very slightly from the results obtained using a standard random effects model. The credibility intervals for the RCTs and cohort studies are fractionally narrower using the hierarchical model, but a much wider interval is obtained for the overall pooled result. This transpires because heterogeneity at the study type level is being accounted for in the general synthesis model which is ignored when

using a simple random effects model on all studies. Not only is this credibility interval wider (and non-significant), it is wider than either of those produced when considering the RCT and cohort studies individually. This may seem counter intuitive because the hierarchical model is using much more information when combining than the individual study types separately, which would normally lead to a narrowing of confidence intervals - rather than a widening. However, the overall pooled estimate represents the mean of the population of study *type* effects, rather than the mean of the population of study effects in the standard random effects model. As only two study types are included in this example there is a lot of uncertainty in the estimation of the between study type random effect (0.01 (0.001 to 5.57)), this in turn leads to the uncertainty observed in the overall mean effect (0.78 (0.40 to 1.53))

The effect of 'borrowing strength' can be observed at two levels. Individual studies results are shrunk towards the pooled estimate for that study type (this is shown by the dashed lines in Figure 5.6), and pooled results for study types are shrunk towards the overall pooled estimate (this can be observed by comparing the study level results using model 1 and using a simple random effects model - plotted in Figure 5.7).

5.1.2 Sensitivity analysis to estimation of the between study type variance

In the previous section the initially counter-intuitive result that the credible interval for the overall mean effect size combining all the evidence using the generalised synthesis model was wider than that estimated for either study type individually was considered. It was noted that this was largely due to the poor estimation of the between study type variance parameter since only two study types are included in the model. This is despite the fact that the estimates for the RCTs and the Cohort studies were in quite close agreement. In such circumstances the prior distribution placed on this parameter may be influential since there is little data available. It has been noted previously that specifying truly 'vague' parameters for variance components in hierarchical models is non-trivial. In the model presented above a InverseGamma(0.001,0.001) was used. In this section other prior distributions are used to assess the influence of prior choice on the overall results.

Several different distributional forms for priors for variance components are reasonable. The influence of several of these are explored in depth elsewhere, (Lambert et al. 2001b) and in Chapter 7, but for this analysis focus is restricted to the InverseGamma distribution. Less "vague" distributions of InverseGamma(0.01,0.01), InverseGamma(0.1,0.1) and InverseGamma(1,1) were used. The parameter estimates produced are given in Table 5.2

Table 5.2 Results of sensitivity analysis to the prior distribution placed on the between study type variance parameter

Prior distribution for the between study type variance (ν^2)	ν^2 **Between study type variance (95% CrI)**	e^{ϕ} **Overall pooled OR estimate (95% CrI)**
InverseGamma(0.001,0.001)	0.010 (0.001 to 5.57)	0.78 (0.40 to 1.53)
InverseGamma(0.01,0.01)	0.051 (0.005 to 19.57)	0.79 (0.22 to 2.54)
InverseGamma(0.1,0.1)	0.336 (0.038 to 68.92)	0.78 (0.06 to 12.35)
InverseGamma(1,1)	0.849 (0.213 to 9.44)	0.77 (0.12 to 5.09)

The table shows that while the point estimate for the overall pooled estimate appears robust over the range of priors specified, the corresponding confidence interval varies greatly, as does the estimate and credible interval for the between study type variance parameter, indicating the lack of the robustness of the latter to the choice of prior distribution. This is an important issue when fitting such models as this will always be a problem when the number of study types is small, which will usually be the case. Three potential approaches are suggested which may address this problem. Firstly, data-inflation methods are discussed later in this thesis (section 7.4). These are used to remove the undesired influence of prior distributions and may be able to be utilised

here. Secondly, it may be possible to increase the number of study types, and hence the number of units in the top level of the hierarchy. For instance, in the cholesterol example, trials using different interventions could be included as different study types. Increasing the number of units in this way will improve estimation of the between study type variance, but it may also have other effects on parameter estimation. Finally, instead of trying to specify a vague prior distribution for the between study type variance, it may be desirable to derive an informative empirical prior. Higgins and Whitehead (Higgins and Whitehead, 1996) derive empirical priors for the between study heterogeneity parameter in a random effects meta-analysis model by examining the estimates of this parameter from previous related meta-analyses. In a similar fashion it could be possible in the generalised synthesis context to examine the variability that exists between study types in syntheses where more than one study type has been examined. A recent investigation by et al. (Ioannidis et al. 2001) may provide data for such an exercise.

SPECIAL NOTE

THIS ITEM IS BOUND IN SUCH A MANNER AND WHILE EVERY EFFORT HAS BEEN MADE TO REPRODUCE THE CENTRES, FORCE WOULD RESULT IN DAMAGE

Figure 5.6 Forrest plot displaying results of combining RCTs and Cohort studies

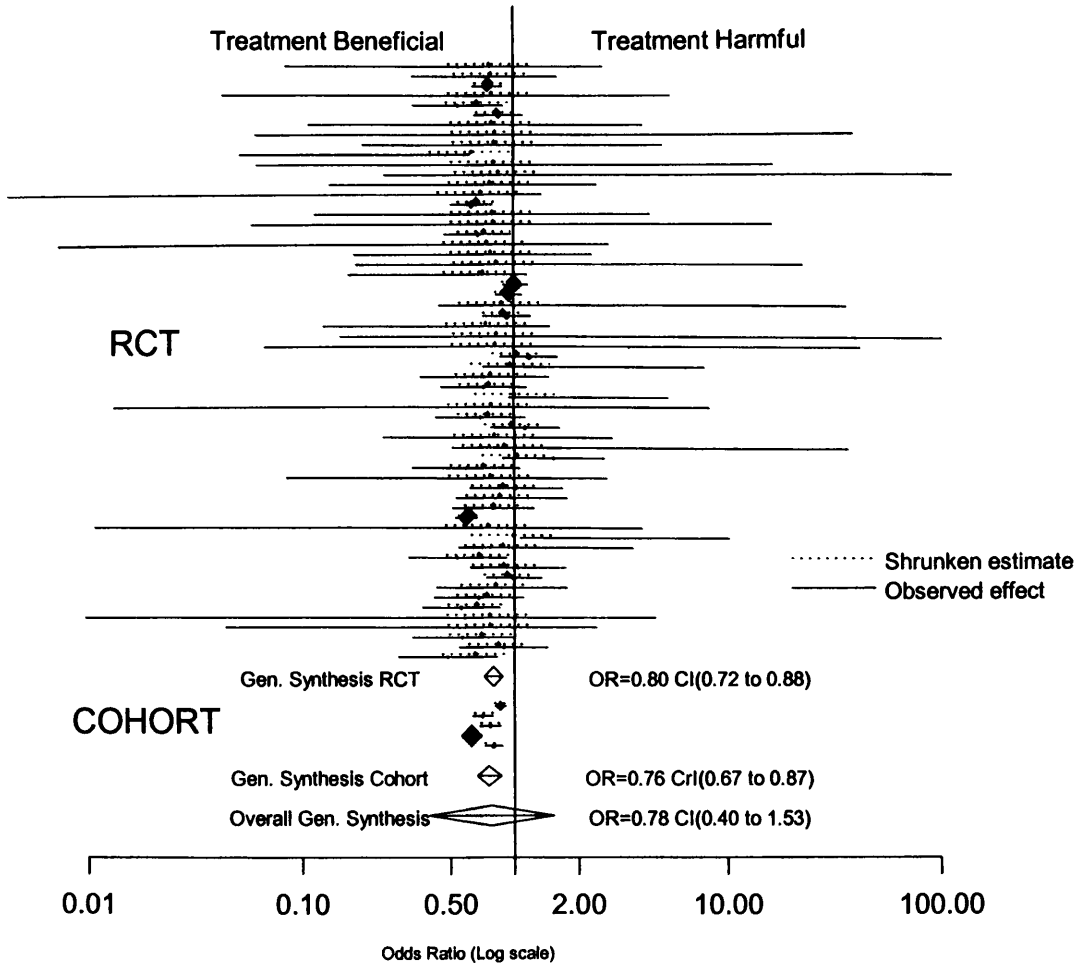
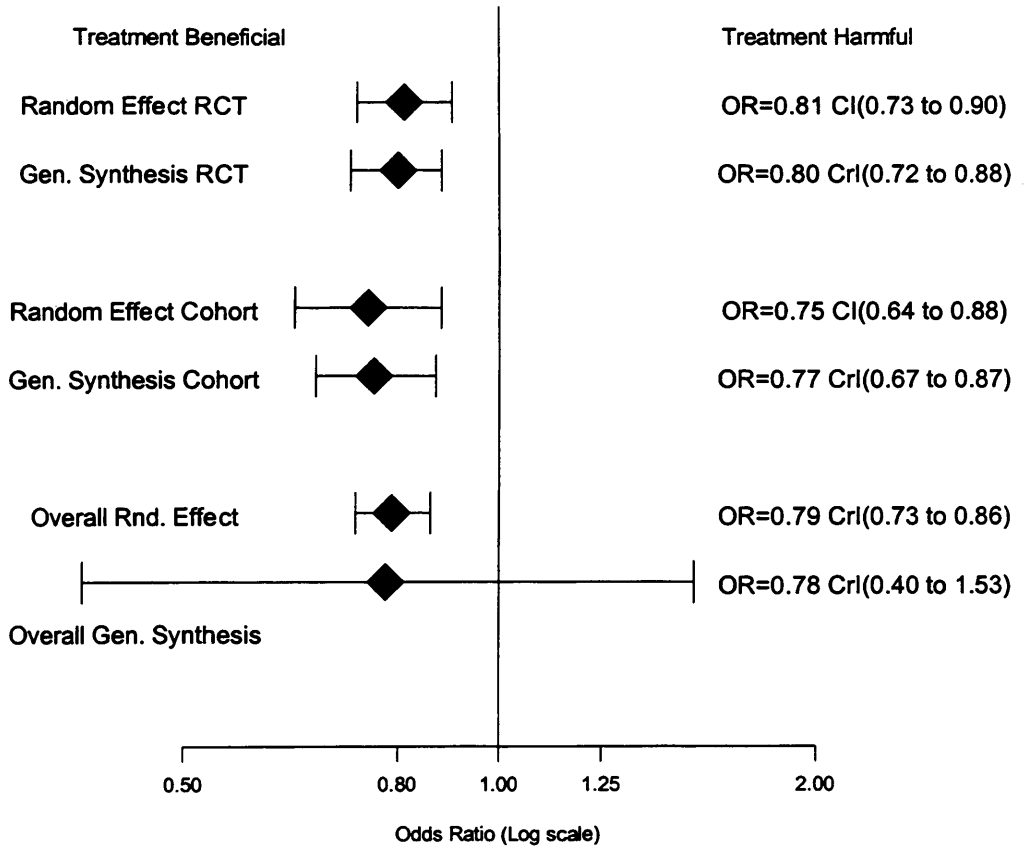


Figure 5.7 Pooled estimates derived from combining study types individually and together



5.1.3 Inclusion of covariates

Using the above model as a basis, it is possible to adjust the RCT results for baseline risk. An amalgamation of the hierarchical model of equation (5.1) and the meta-regression model (2.21) adjusting for regression to the mean is constructed.

RCT "exact" binomial model

$$r_{cj} \sim Bin[p_{cj}, n_{cj}] \qquad r_{tj} \sim Bin[p_{tj}, n_{tj}] \qquad j = 1.\ldots.60$$

$$\log it(p_{cj}) = \mu_j \qquad \log it(p_{tj}) = \mu_j + delta_j$$

$$delta_i = delta_j' + \beta(\mu_i - \bar{\mu}) \qquad delta_j' \sim \mathrm{N}[\theta_1, \tau_1^2]$$

$$\mu_j \sim N(0,10^5) \qquad \tau_1^2 \sim IG(0.001,0.001)$$

$$\beta \sim \mathrm{N}[0,10^6]$$

Cohort study model

$$T_k \sim N[\Psi_k, \sigma_k^2] \qquad k = 1\ldots.5$$

$$\Psi_k \sim N[\theta_2, \tau_2^2]$$

$$\tau_2^2 \sim IG(0.001,0.001) \qquad (5.2)$$

Pooling both study types

$$\theta_m \sim N(\phi, \nu^2) \qquad m = 1,2$$

$$\phi \sim \mathrm{N}(0,10^6) \qquad \nu^2 \sim IG(0.001,0.001)$$

The only changes in equation 5.2 from 5.1 are in the RCT part of the model where the regression coefficient for baseline risk (centred), β, is included with a vague prior placed on it. Hence, now θ_1 has the interpretation of the pooled odds ratio estimate for the RCTs *adjusted* for the average baseline risk observed in the trials. In 2001 this model ran with no problems and in a reasonable amount of time, unlike in 1998 when it was infeasible to fit, due to slower computing/memory constraints, and an earlier, inferior, version of BUGS.

The model was run using a burn in of 10000 iterations followed by a sample of 40000 iterations to base estimation and inferences on. Informal assessment of convergence indicated there was no reason to believe there were any problems with convergence. The parameter estimates are provided in Table 5.3.

Adjusting for baseline risk has increased the pooled odds ratio for the RCTs from 0.80 to 0.85. This result is consistent with the analysis of the RCTs described in section 4.9.7. This increases the discrepancy between RCT and cohort results, which in turn inflates the between study type variance estimate, which in turn inflates the credible

Table 5.3 Transformed estimates and interpretation

Parameter	Interpretation	Median estimate (2.5 & 97.5 percentiles)
e^{ϕ}	Overall pooled OR (pooling the RCT and Cohort study population results)	0.81 (0.35 to 1.87)
e^{θ_1}	OR pooled estimate for RCTs (comparison of treatment and control groups)	0.85 (0.76 to 0.94)
e^{θ_2}	OR pooled estimate for Cohort studies (corresponding to a 0.87mm/l difference in cholesterol)	0.76 (0.67 to 0.89)
β	RCT baseline risk regression coefficient	-0.098 (-0.164 to –0.022)
ν^2	Between study population type variance	0.018 (0.001 to 8.84)
τ_1^2	Between study variance within RCTs	0.024 (0.004 to 0.091)
τ_2^2	Between study variance within Cohort studies	0.018 (0.005 to 0.119)

interval for the overall pooled odds ratio, while increasing its point estimate slightly also.

5.1.4 Discussion of the use of hierarchical modelling for combining the cholesterol data

Section 5.1 has illustrated how hierarchical modelling can be used to synthesise evidence from different sources while explicitly allowing for heterogeneity in effects from the different sources. With current software such modelling is very feasible, but issues of problems with the estimation of the between study type population effects due to small numbers of study types has been noted. The feasibility of using different

statistical models to combine different study types, and the adjustment of results for study level covariates has been demonstrated.

The hierarchical approach taken here is appealing as it provides a transparent model for evidence synthesis, and hence a framework in which to explore the impact of different sources of evidence. For example, an extension could be to downweight the observational evidence if it was thought to be potentially more biased. This idea is explored further in section 7.8.

A general criticism that could be laid at this hierarchical modelling approach to evidence synthesis, is that the emphasis is on producing an overall pooled effect size and does not address the wish, as described fully in section 2.9 to estimate treatment effects for individuals or groups of patients. For example, it may be more desirable to use the observational evidence to inform treatment and policy decisions in patients under-represented, or not represented at all in the RCTs, and in a sense extrapolate the results to contexts not addressed by trials. In a cholesterol lowering context, this could mean using observational evidence to inform about the very elderly, or patients with very low risks of CHD (NB. this was one of the initial aims of cross design synthesis).

A different approach to modelling the cholesterol data is described in the next section. Further consideration of the hierarchical model is provided in section 5.3 later in this chapter.

5.2 Combining the observational and randomised evidence using a variance components model

The standard meta-analysis models for combining comparative studies utilise a comparative measure of effect, such as the odds ratio or relative risk for binary outcomes, or the standardised mean difference for continuous outcomes. Typically the comparison is made between two groups of patients in each study. In situations where all studies to be combined are simple 2-arm single period RCTs comparing comparative treatments, where exchangeability can be assumed between studies, and only one outcome is of interest, this model is entirely appropriate. However, where the types of

studies to be combined are more heterogeneous in design, and outcome measures reported, the use of a more sophisticated model may be advantageous. Possible benefits of adopting such an approach include the incorporation of more evidence, and more exact specification of the data included in the model.

DuMuchel (DuMouchel, 1998) describes a model which allows many extensions to the standard meta-analysis model (see also section 3.5). Firstly, it has the capacity to model multiple outcomes from studies, provided they are all reported on the same scale, using variance component modelling. In addition, it is possible to specify more than two treatment or intervention states being compared by the studies in the analysis. Different study designs are accounted for, as different treatment groups can consist of results from separate groups of subjects, or from groups that cross-over and are subject to multiple treatments.

A key feature which distinguishes this model from previous ones is that each group of patients are modelled separately. So, for example, if one is considering a binary outcome, the log odds in each group is considered, rather than a comparative measure such as the log odds ratio. The most general form of the model is given below:

$$y_{ijktm} = \mu_m + \beta_k + \gamma_{mk} + a_i + b_{im} + c_{ik} + d_{j(i)} + e_{t(i)} + \varepsilon_{ijktm}, \qquad (5.3)$$

where y is the observed outcome, i indexes the N studies in the analysis, m indexes the M different outcomes considered, j indexes the J_i cohorts of patients in the ith study, k indexes the K treatments being compared in the analysis, and t indexes the T_i time periods or cross-over states in the ith study. A variance component model is used to model the expected correlations among reported results from the same study and cohorts of subjects. Outcome and treatment are considered fixed, while study, cohort and period are considered random effects. Hence, μ, β, and γ denote fixed effects due to outcome, treatment and their interaction, and a, b, c, d, and e are random effects due to study, study*outcome, study*treatment, cohort within study, and period within study, respectively, and ε denotes pure error. Simple extensions of this model allow study or cohort level covariates to be included. (DuMouchel, 1998)

DuMouchel notes that it is assumed that each study reports results for two or more of the treatments under investigation. This, however, is not a restriction of the model because it is possible to include single arm studies without modification, but rather reflects a concern about including uncontrolled evidence[1]. A further issue not highlighted in the original paper is that because individual arms of studies are modelled separately the benefits of randomisation are largely lost because *direct comparison between arms of the same studies are not made.* This essentially has the effect of turning the meta-analysis into a synthesis of uncontrolled data, which clearly has serious implications.

Details have been given (DuMouchel, 1998) on how to implement this model using PROC MIXED in SAS. (SAS Institute Inc., 1992) Here maximum likelihood or restricted maximum likelihood is used to estimate the random effects and then weighted least squares is used to estimate the fixed effects. At the time the method was described (1997) it was necessary to perform a grid search to identify the variance components, and it usually required several runs to narrow the search down. Fortunately, improvements to PROC MIXED have been made since then and in SAS V6.12 this grid search is no longer required. Instead, a more efficient Newton-Raphson search can be applied to estimate the unknown random effects, while using the known values of the standard deviations of the response data. Additionally, a simplification of the code is also now possible.[2] The original code, published by DuMuchel and the improved version which will run on SAS v6.12 are both given in Figure 5.8 (Note: the original code does not run on SAS v6.12).

[1] Personal communication with William DuMouchel

[2] These modifications were identified through e-mail discussions with both William DuMuchel and Russel Wolfinger at the SAS institute

Figure 5.8 Original and improved SAS PROC MIXED code to implement the variance component model of DuMuchel

Original code

```
proc mixed data = dataset sigiter;
class outcome treatment study cohort period;
weight w;                                /* define w = 1/s**2  */
model y = outcome treatment outcome*treatment  / noint solution;
random intercept outcome treatment cohort period / subject = study
solution;
repeated / local localw;
parms (tausq-alist) (tausq-blist) (tausq-clist) (tausq-dlist) (tausq-
elist) (1)/noiter;
run;
```

(Where (tausq-alist) is a set of values for which the REML likelihood is calculated etc.)

Improved code

```
proc mixed data = dataset noprofile;
class outcome treatment study cohort period;
weight w;                                      /* define w = 1/s**2  */
model y = outcome treatment outcome*treatment  / noint solution;
random intercept outcome treatment cohort period / subject = study
solution;
parms (0.5) (0.5) (0.5) (0.5) (0.5) (1)/eqcons=6;
run;
```

5.2.1 Applying DuMouchel's model to the cholesterol RCTs

This section considers how the method described by DuMouchel and outlined above can be applied to the cholesterol RCT dataset. There are several theoretical benefits of using this model in this instance, and it circumvents several shortcomings of previous analyses. These benefits are discussed below.

Note: This analysis was not updated in 2001, and hence uses slightly fewer trials than all previous examples, explicitly trials R61 through to R64 are not included in the analysis. Since no direct quantitative comparisons with previous model estimates are made it is perceived that this loss of continuity will have minimum impact.

Examining several outcomes simultaneously

In the review of previous meta-analyses on the cholesterol lowering RCTs (Section 4.4) it was noted that several different outcomes had been examined, most commonly total mortality, CHD mortality, CHD events, and also non-CHD mortality. However, no one meta-analysis considered more than one outcome, and this was one reason different meta-analyses came to different conclusions. Although it is always possible to carry out individual meta-analyses for different outcomes, this model combines all available data in one analysis, making comparison between outcomes clear, and makes the exploration of covariates over different outcomes easy (note: in doing so it assumes that the same model structure is appropriate for all outcomes and "borrows strength" across them). In the analysis which follows, total mortality, CHD mortality and total CHD events outcomes are all examined.

Including trials with more than two arms

In section 4.5 it was noted that several of the trials had more than two randomised arms. Previous meta-analysis had dealt with these differently, usually either excluding certain arms or merging certain ones in order to reduce the data to a two-arm comparison. Neither of these approaches is ideal, and a third alternative of including the comparative effect of each experimental arm with the standard/placebo as a separate estimate results in the standard/placebo arm data being included in the analysis multiple times. The

model described above allows each individual arm to be modelled separately and hence no excluding or merging of arms is required.

Modelling the intervention given in the control arm(s) of the trials

The intervention administered to the control arm(s) of the trials varied considerably. Early trials typically gave no treatment, or a placebo, while in the later drug trials strict diets were often given. These diets often reduced cholesterol levels more than diets given in the experimental arms of the earlier trials! Such heterogeneity of study design brings into question the assumption that comparative measures of treatment effect are exchangeable across studies. Using the model above, the intervention given in the control arm of the trial can be explicitly modelled.

Including arm level covariates

In section 4.9.4 it was found that the baseline risk in the control group was highly correlated with outcome i.e. the higher the risk of the patients the greater the treatment effect. The model of DuMouchel also allows the inclusion of study level covariates, however it also allows covariates to be included at the cohort, or study arm level. Hence, one could explore the effect on absolute level of cholesterol reduction in each arm opposed to the difference between (two) arms. When a study has more than two arms, a different intervention is administered to each, and cholesterol is reduced by different amounts in each arm, then this extra information can be included in the analysis (an example of the data that are included in the model is given in section 5.2.2). Note that using the current formulation of the model no adjustment is made for regression to the mean when including baseline risk as a covariate. In the cholesterol dataset this was not found to be a large problem, however development of a method to compensate for the problem in a variance components model would be desirable, but is not pursued here.

Including observational evidence

Modelling the event rate in each arm of each study, allows for a natural extension to included data from single arm or non-comparative studies. Hence this model provides the opportunity to include observational data in a different way from that of section 5.1. This possibility is explored in section 5.2.11.

5.2.2 Illustration of the data included in the model

In order to clarify the information being included, and to further illustrate the benefits of this model, an individual trial (R10) is examined and the data included from it in the synthesis described. (The complete datasest used for all RCTs is given in Appendix A.VIII.). Data available for trial R10 is given in Table 5.4.

Table 5.4 Illustration of the potential extra data included in model from a 6 arm RCT (R10)

Arm	Number patients	Total mortality	CHD mortality	Total CHD events	Baseline risk (event rate in the control group)	% decrease in cholesterol level from baseline
Placebo	143	27	23	42	50.3	-0.2
Estrogen	141	27	25	51	50.3	-0.8
Dextrothyroxine	74	10	8	20	50.3	-4.4
Nicotinic acid	77	15	13	25	50.3	-7.9
Est. + Dex.	67	13	10	21	50.3	-12.3
Est. + Nicot.	68	16	15	24	50.3	-11.5

This trial has six arms, each with data on the three outcomes under investigation. Both a study level covariate (baseline risk) and an arm level covariate (% decrease in serum cholesterol) are shown. For analysis in SAS a separate line of data is required for every

arm/outcome combination. Figure 5.9 presents the raw data corresponding to this trial. (Note that it was necessary to merge the last 4 arms of Table 5.4 because they all have the same treatment code (6 = other drug) and there were problems fitting the nested cohort random effect (see below). The average cholesterol reduction over these four arms was calculated for the *cholpc* variable.)

Figure 5.9 Illustration of data used model when fitting the model of DuMuchel

study	*outcome*	*treat*	*y*	*w*	*br*	*cholpc*
10	1	7	-1.44	22.23	50.3	-0.2
10	1	2	-1.42	22.16	50.3	-0.8
10	1	6	-1.45	44.13	50.3	-9.0
10	2	7	-1.63	19.65	50.3	-0.2
10	2	2	-1.51	20.90	50.3	-0.8
10	2	6	-1.64	38.95	50.3	-9.0
10	3	7	-0.87	29.91	50.3	-0.2
10	3	2	-0.56	32.76	50.3	-0.8
10	3	6	-0.77	61.91	50.3	-9.0

study - trial number
outcome - 1 - total mortality, 2 - CHD mortality, 3 - coronary events
treat 1 - fibrate class drug, 2 - hormonal class drug, 3 - statin class drug, 4 - diet/diet advice, 5 - surgery, 6 - other drugs, 7 - placebo/nothing/usual care
y - log odds (ln(number of patients having event/number of patients not having event))
w - weighting = 1/var(ln(odds))
br- baseline risk - (as described in section 4.9.4)
cholpc - percentage decrease in total serum cholesterol during trial for that group

5.2.3 Model & code used

The cholesterol studies did not require the full general model (equation 5.3) because no patients were given a series of treatments (i.e. no cross-overs) which meant that the period term was obsolete and hence not fitted. In addition, after some preliminary modelling, it appeared that including a study*cohort interaction was problematic. The reason for this is that there are few studies which administer the same treatment to different cohorts of people, so there is very little data available to distinguish between a study*treatment and a study*cohort interaction. Hence, a compromise was necessary; the study*cohort interaction was dropped and arms of trials which administered the same treatment (or more exactly, had the same treatment code) were merged. Trials which included different arms with treatments categorised by the same code were R1, R10, R15, R22, R38, R40, R56 and R57. Exploratory analysis suggested that the changes in the final estimates produced by the model after doing this were minimal. Additionally, due to some combinations of outcome and treatment having few data points, there were problems estimating all the *outcome*treatment* interaction coefficients (the effects of statins were particularly problematic) in the model. DuMuchel noted this problem also in the application the model was developed for (see below) (DuMouchel, 1998) and suggested fitting the interaction as a random effect, which assumes the interactions are distributed normally, with mean 0 and variance to be estimated. This allows the interaction estimates to "borrow strength" from each other.

The respective modifications required to the model code are relatively straightforward. The code used which incorporates the necessary modifications, and also includes the single covariate baseline risk is presented in Figure 5.10.

Figure 5.10 SAS code used to fit model to cholesterol data

```
proc mixed data = chol1j noprofile;
      class outcome treat study;/* removed period & cohort*/
      weight w;
      model y = outcome treat br br*treat/ noint solution;
      random outcome*treat study study*outcome study*treat/solution;
      parms (0.05)(0.5)(0.5)(0.5)(1)/eqcons=5;
run;
```

5.2.4 Results

It appears that percentage decrease in cholesterol did not explain heterogeneity between estimates, however baseline risk was highly statistically significant ($p = 0.004$), as was its interaction with treatment ($p = 0.005$)). The fixed effect term for outcome was also highly significant ($p = <0.001$), suggesting differences in effect for different outcomes. The overall treatment (type) variable was not formally significant ($p = 0.08$) however individual indicator variables for some treatment indicators were small. The parameter estimates from this model are given in Table 5.5.

Table 5.5 Parameter estimates fitting model with baseline risk covariate

Parameter	Estimate/ Standard error	P-value
Random effects		
Outcome*Treatment	0.016	
Study	0.596	
Outcome*Study	0.310	
Treatment*Study	0.074	
Fixed effects		
Outcome		
Total mortality	-2.57 (0.16)	<0.0001
CHD mortality	-2.99 (0.16)	<0.0001
Coronary events	-2.03 (0.17)	<0.0001
Treatment		
Fibrates	-0.054 (0.18)	0.77
Hormones	-0.849 (0.32)	0.02
Statins	-0.670 (0.29)	0.04
Diets	-0.159 (0.14)	0.27
Surgery	-0.840 (0.43)	0.07
Other drug	-0.115 (0.174)	0.52
Covariate		
Baseline risk	0.029 (0.005)	<0.0001
Br*fibrate	-0.008 (0.007)	0.23
Br*hormone	0.023 (0.006)	<0.001
Br*statins	-0.014 (0.016)	0.39
Br*diets	-0.003 (0.003)	0.33
Br*surgery	-0.044 (0.034)	0.20
Br*other drug	-0.013 (0.006)	0.05

Of the four estimable random effect parameters *study* and *outcome*study* are the largest. This suggests that there is considerable variability between studies, both generally and across outcome measures. It is interesting to note that these components were also largest in the meta-analysis of clinical reminder systems, for which this methodology was originally developed (DuMouchel, 1998). Smaller in magnitude are the variability parameters corresponding to *study*treatment*, indicating the variation in intervention effects from study to study, and the interaction between treatment and outcome measure (*outcome*treatment*).

Focusing on the fixed effects, interpretation is non-trivial, but clearly different effects from different forms of treatment are estimated for different baseline risks. Although these coefficients are in themselves informative, more directly relevant are comparative effects and confidence intervals calculated using linear combinations of the fixed and random effects. In this way log odds ratios comparing different treatments can be produced. For example, if the comparative effects of statins v diet interventions, for each of the three outcomes, for various specified levels of baseline risk is of interest, comparative estimates can easily be derived from the above model. Importantly, in these estimates all studies which included a diet *or* a statin arm *or* both are included "directly" in the analysis, while strength is being borrowed from all studies via the random effect terms. Hence, more information is being included than if just studies comparing statins with diets directly had been combined.

A similar approach to estimation is possible when estimating direct comparisons using models such as that described by Higgins and Whitehead (Higgins and Whitehead, 1996) (See equation 3.2). The differences between models are that, while the model of Higgins is more restrictive regarding the types of studies it can synthesise, it does have the advantage that randomisation is not 'broken' during the analysis. Further consideration is given to such models that 'maintain' randomisation in the analysis in Section 5.4.

Sample code to work out the estimates and 95% confidence intervals for the example comparisons suggested above is given in Figure 5.6. Three baseline risk levels were examined, a typically low risk (3 CHD deaths per 1000 person years, the average risk

across trials (un-weighted average = 25 CHD deaths per 1000 person years), and a typically high risk (80 CHD deaths per 1000 person years). Results from these contrasts are provided in Table 5.11.

Figure 5.11 Code to calculate estimates & 95% confidence intervals for treatment comparisons taking into account baseline risk

This comparison is for statins v diet intervention for the outcome total mortality for patients at a typically high risk. Here, treatment 3 (statins) is contrasted with treatment 4 (diet) and the effect of a baseline risk of 80 CHD deaths is contrasted for these interventions. The random effects for the 1st outcome (total mortality) for the 3rd and 4th interventions are also included in the contrast.

```
estimate 'st v di CHDm hBR'
treat 0 0 1 -1 0 0 0
br*treat 0 0 80 -80 0 0 0 |
treat*outcome 0 0 0  0 0 0    1 0 0   -1 0 0
              0 0 0  0 0 0    0 0 0  / cl;
```

Table 5.6 Comparison of statin v diet treatments for total mortality, CHD mortality and CHD events over three levels of baseline risk

Outcome	Baseline risk (CHD deaths per 1000 person years)	OR (95%CI)
Total mortality	80	0.33 (0.03-3.57)
CHD mortality	80	0.29 (0.03-3.03)
CHD events	80	0.32 (0.03-3.44)
Total mortality	25	0.60 (0.29-1.22)
CHD mortality	25	0.52 (0.27-0.99)
CHD events	25	0.58 (0.29-1.17)
Total mortality	3	0.76 (0.46-1.24)
CHD mortality	3	0.66 (0.45-0.96)
CHD events	3	0.73 (0.46-1.18)

These results suggest that statin intervention is preferable to diet for all three of the baseline risk levels explored. Although no result is statistically significant, a clear trend of increasing benefit of the use of statins over dietary interventions as baseline risk increases is evident. At the high baseline risk level the benefit of statins is four fold, while it reduces to around two fold at average risk in the trials and reduces to about 20% benefit at low levels of baseline risk. It is also worth noting the change in width of confidence intervals over different baseline risks this is primarily due to the data available (i.e. there were no statin trials done on patients at high baseline risks). Interestingly, the comparative effects for each outcome appear to be almost identical at a given baseline risk.

5.2.5 Summary

A novel meta-analysis model has been applied to the cholesterol randomised trials with considerable success and benefits. Several of the shortcomings of previous modelling approaches have been so alleviated. These include modelling of treatment in the control arm of the trial, inclusion of data from studies with more than two arms and the exploration of study arm level covariates.

However, the problem of regression to the mean when modelling baseline risk, for which there are methods of adjusting for using the standard meta-analysis model, are not available using this approach. This is a potentially serious issue. In addition, all the usual advantages of using a Bayesian approach to meta-analysis, including the proper accounting of uncertainty due to estimating the random effect parameters are not present in the empirical Bayes approach employed above. A variance component model is implemented from a Bayesian perspective in the next section.

5.2.6 Implementing DuMouchel's variance components model using a Bayesian framework

This section describes a Bayesian implementation of the most general model described by DuMouchel (DuMouchel, 1998) using WinBUGS. In order to assess the plausibility of implementing the full model, the dataset of RCTs to evaluate computer-based clinical reminder systems for preventive care in the ambulatory setting, (Shea et al. 1996) for which this model was originally developed, is examined in section 5.2.7. The model is then fitted to the cholesterol lowering dataset in section 5.2.8.

5.2.7 Application to the computer-based reminder systems meta-analysis

This dataset consisted of data from 16 primary studies investigating the effects of computer-based clinical reminder systems on rates of recommended preventive care practices for outpatient visits. (Shea et al. 1996) Six preventative practice groups are examined; vaccinations, breast cancer screening, cervical cancer screening, colorectal cancer screening, cardiovascular risk reduction, and other preventative services. Four

treatments were reported: computer reminders (to the physician), manual reminders, both computer and manual reminders, and control. The studies included cross-sectional and multi-period designs, and all were randomised at the patient or physician level. The response of interest are the logits of the proportion of eligible patients for which the various preventative practices were carried out.

The structure of the model is the same as that defined by equation 5.3 (with the exception that the *outcome*treatment* variable was specified as random rather than fixed). Prior distributions are required for the model parameters to be estimated. For all parameters, vague priors were used; Inverse-Gamma(0.001,0.001) for the variance components; and Normal(0,1000) for the fixed effects. Annotated WinBUGS code to implement the model, including the dataset, is given in Appendix A.VIII.

The modelling aspect of the code is relatively straightforward, but the way data is indexed needs some explanation. The main loop of the program iterates around all lines of the data. Each response is then uniquely identified in a six dimensional matrix. The dimensions of this matrix represent study, treatment, outcome, group, time, and the replication number. This last dimension is required because some lines of data in the dataset are not uniquely identified by the first five dimensions (usually because different doses of the same drug have been given and the model is not detailed enough to examine different dose levels per se). The specification of the regression equation, although lengthy, is straightforward and equivalent to the SAS code. Next, the nature of the random effects are specified. The variables tau[1] through to tau[6] specify the random effects for the *treatment*outcome*, *study*, *treatment within study*, *outcomes within studies*, *cohorts of patients within studies*, and *time periods within studies*, in that order. Note how the highest level of the treatment variable is set to zero. This is required for a unique solution, allowing full-rank constraints to hold (i.e. one level of the categorical variable is not estimated). Nodes representing the comparative estimates are defined. Here, each treatment is compared to placebo for each outcome. Conveniently, data can be entered in exactly the same format as that required for the SAS implementation. In this example, the 16 studies provide 330 separate outcome responses and hence 330 lines of data are included.

Following preliminary examination, a 5000 iteration burn-in followed by a run of 25000 further iterations was considered adequate. This took under two and a half minutes on a pentium II 450Mhz (which is computationally quicker than using PROC MIXED!). Examination of the MCMC chains indicated that autocorrelation between successive iterations was considerable, but not impractical to work with (hence the relatively long chain required). The estimates from this model are compared to the classical approach in SAS in the next section.

5.2.8 Comparison of classically and Bayesian derived estimates for the computer-based reminder systems meta-analysis

Estimates of the odds ratios comparing the three interventions to placebo for each of the six outcomes generated using PROC MIXED, as reported by DuMouchel, (DuMouchel, 1998) and obtained usin WinBUGS are reproduced in Table 5.7 The random effect estimates for the same model are provided in Table 5.8.

Table 5.7 Odds ratios and 95% confidence intervals for the effect of three interventions versus control on six classes of outcomes using PROC MIXED (Reproduced from DuMuchel (DuMouchel, 1998)) and from the WinBUGS implementation

Intervention & Analysis		Computer Remind. OR (95% CI/CrI)	Manual Reminder OR (95% CI/CrI)	Both OR (95% CI/CrI)
Vaccinations	Classical	3.09 (2.39 to 4.00)	2.46 (1.86 to 3.25)	3.06 (2.25 to 4.16)
	Bayesian	*3.07 (2.43 to 3.87)*	*2.45 (1.90 to 3.17)*	*2.58 (1.97 to 3.37)*
Screen	Classical	1.88 (1.44 to 2.45)	1.63 (1.21 to 2.18)	1.88 (1.44 to 2.45)
Breast	*Bayesian*	*1.83 (1.44 to 2.31)*	*1.60 (1.22 to 2.09)*	*1.80 (1.37 to 2.37)*
Screen	Classical	1.15 (0.89 to 1.49)	1.10 (0.82 to 1.46)	1.12 (0.82 to 1.51)
Cervical	*Bayesian*	*1.25 (0.99 to 1.57)*	*1.17 (0.90 to 1.52)*	*1.47 (1.12 to 1.92)*
Screen	Classical	2.25 (1.74 to 2.91)	1.85 (1.39 to 2.47)	2.71 (2.01 to 3.66)
Colon	*Bayesian*	*2.20 (1.75 to 2.78)*	*1.79 (1.38 to 2.33)*	*2.15 (1.64 to 2.82)*
Screen	Classical	2.01 (1.55 to 2.61)	1.86 (1.41 to 2.47)	2.57 (1.89 to 3.51)
CV risk	*Bayesian*	*2.05 (1.62 to 2.60)*	*1.90 (1.47 to 2.46)*	*2.48 (1.89 to 3.25)*
Other	Classical	1.02 (0.79 to 1.32)	0.99 (0.71 to 1.37)	2.59 (1.73 to 3.86)
Preventatives	*Bayesian*	*1.05 (0.83 to 1.33)*	*1.04 (0.76 to 1.40)*	*2.28 (1.55 to 3.37)*

Table 5.8 Random effect variance estimates for model using PROC MIXED as reported by DuMuchel (DuMouchel, 1998) and by MCMC implementation in WinBUGS

	Classical	Bayesian
Random effect	Estimate	Estimate (95% CrI)
Treatment*outcome interaction	0.24	*0.21 (0.15 to 0.32)*
Study	0.38	*0.12 (0.03 to 0.46)*
Study*treatment interaction	0.25	*0.24 (0.17 to 0.35)*
Study*outcome interaction	0.78	*0.84 (0.69 to 1.05)*
cohort within study	0.14	*0.16 (0.11 to 0.26)*
period within study	0.16	*0.17 (0.07 to 0.45)*

It can be seen that the odds ratios agree very closely between models, with the exception of slight disagreement for the both intervention effect for vaccination reminder, where the benefit is a little lower for the Bayesian model. Surprising, however, is the fact that generally the Bayesian credibility intervals are fractionally narrower than the classical confidence intervals. This is unexpected because the uncertainty in estimating the random effect terms is accounted for only in the Bayesian model and hence wider intervals would be expected there. The random effect estimates are all comparable except the between study term which is estimated to be a lot smaller in the Bayesian model. This could well account for the narrower confidence intervals, but, the reason for this occurrence is not clear.

In summary, it would appear that fitting a variance component model of this type in WinBUGS is very feasible. Although the results of the two models agree closely, some concerns remain why the Bayesian credibility intervals are estimated as being narrower than the classical confidence intervals, and why there is discrepancy in the estimate of between study variability. It has not been possible to establish a reason for this, largely because of inaccessibility of PROC MIXED preventing a deeper understanding of the mechanisms of the program. A further problem is that the original published code will not run on the latest version of SAS, (v6.12) and although steps were made to establish equivalent code that would run slight discrepancies remain between the two versions (and WinBUGS) for which no explanation is available. It may be argued that the WinBUGS implementation should be favoured because the model specification method is much more transparent, but, this issue does require further investigation. In the next section a Bayesian variance components model is applied to the cholesterol RCT meta-analysis dataset.

5.2.9 Applying the fully Bayesian random components model to the cholesterol RCT data

The code required to fit the Bayesian version of the model fitted using the SAS code of Figure 4.25 only requires minor modifications from the more general model described

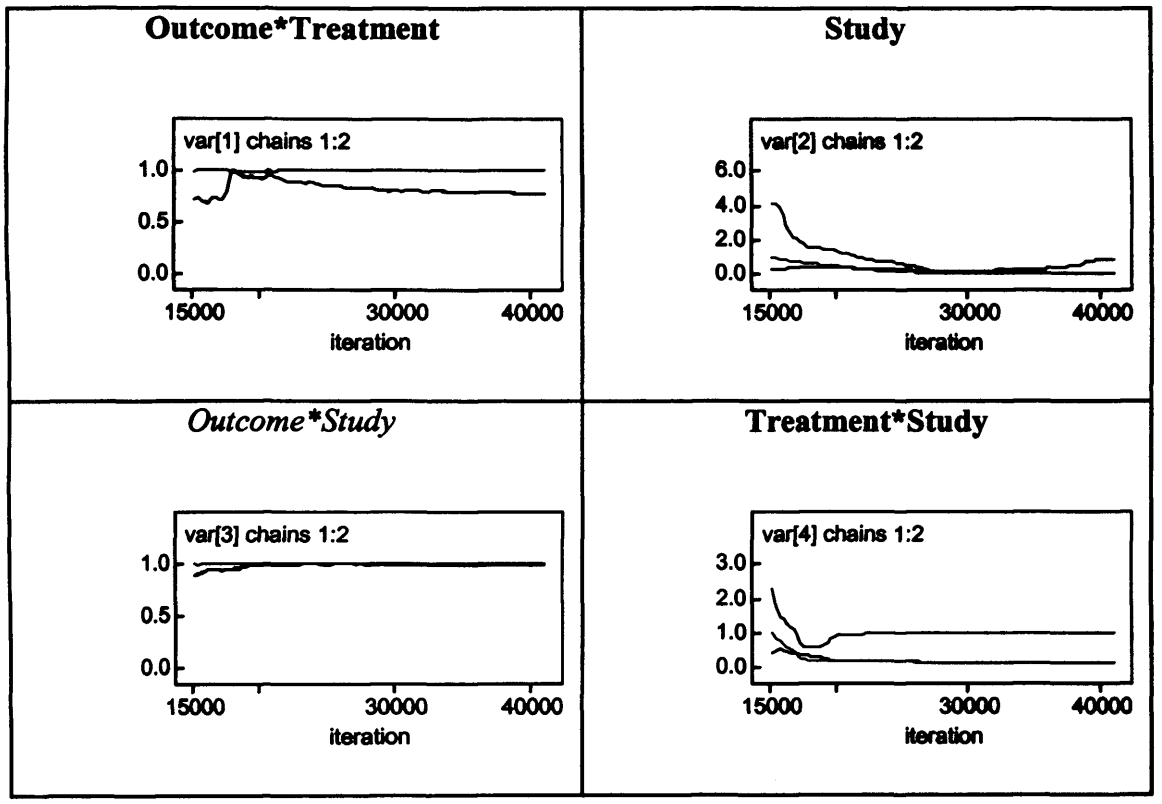
Table 5.9 Parameter estimates fitting model with baseline risk covariate

Parameter	Classical estimate/ Standard error (reproduced from Table 5.3)	Bayesian estimate/ Standard error
Random effects		
Outcome*Treatment	0.016	*0.133 (0.035)*
Study	0.596	*0.760 (0.102)*
Outcome*Study	0.310	*0.558 (0.047)*
Treatment*Study	0.074	*0.280 (0.045)*
Fixed effects		
Outcome		
Total mortality	-2.57 (0.16)	*-2.57 (0.16)*
CHD mortality	-2.99 (0.16)	*-3.00 (0.17)*
Coronary events	-2.03 (0.17)	*-2.05 (0.17)*
Treatment		
Fibrates	-0.054 (0.18)	*-0.043 (0.20)*
Hormones	-0.849 (0.32)	*-0.822 (0.33)*
Statins	-0.670 (0.29)	*-0.649 (0.30)*
Diets	-0.159 (0.14)	*-0.163 (0.14)*
Surgery	-0.840 (0.43)	*-0.851 (0.44)*
Other drug	-0.115 (0.174)	*-0.117 (0.18)*
Covariate		
Baseline risk	0.029 (0.005)	*0.030 (0.005)*
Br*fibrate	-0.008 (0.007)	*-0.008 (0.008)*
Br*hormone	0.023 (0.006)	*0.023 (0.007)*
Br*statins	-0.014 (0.016)	*-0.012 (0.017)*
Br*diets	-0.003 (0.003)	*-0.003 (0.004)*
Br*surgery	-0.044 (0.034)	*-0.044 (0.034)*
Br*other drug	-0.013 (0.006)	*-0.013 (0.007)*

above. This includes the covariate baseline risk, but does not adjust for regression to the mean; this problem is considered in Section 5.2.10. The estimates of the model parameters from WinBUGS are presented in Table 5.9.

If the Bayesian results in Table 5.9 are compared with the classical results (also reproduced from Table 5.5 in Table 5.9) it can be seen that the fixed effects are in very close agreement. However, there is considerable discrepancy between the random effects estimates. The classical model estimates are lower than the Bayesian ones, in some instances, to such an extent that the Bayesian 95% credibility interval does not include the classical estimate. These findings are qualitatively similar, but more extreme, than those observed for the computer based reminder system meta-analysis (i.e. the random effect variance estimates were all lower in the classical model). These differences suggest that the PROC MIXED code is not fitting exactly the model intended. As noted previously, it is difficult to examine what PROC MIXED is doing due to the nature of the program. However, it should not be ignored that there may be convergence issues regarding the random effects in the Bayesian model. Figure 5.12 displays the plots associated with the Gelman-Rubin convergence criteria (briefly described in Section 5.1.1) for the random effects. From these it can be seen there is some suggestion that stability of the lines, and hence convergence is not achieved.

Figure 5.12 Gelman-Rubin convergence plots for the variance-component parameters in the model



Importantly, these apparent differences in the random effect component of the model specification have substantial effects on estimated treatment contrasts that can be seen by comparing the examples in Table 5.10 (Classical results reported previously in Table 5.6 are reproduced). As before, the unexpected result that the Bayesian credible intervals are often narrower than the corresponding Classical confidence intervals is observed.

Table 5.10 Comparison of statin v diet treatments for total mortality, CHD mortality and CHD events over three levels of baseline risk - Classical and Bayesian estimates

Outcome	Baseline risk (CHD deaths per 1000 person years)	Classical estimates OR (95% CI) (Reproduced from Table 5.3)	Bayesian estimates OR (95% CrI)
Total mortally	80	0.33 (0.03-3.57)	*0.40 (0.04 to 4.48)*
CHD mortality	80	0.29 (0.03-3.03)	*0.36 (0.03 to 4.01)*
CHD events	80	0.32 (0.03-3.44)	*0.35 (0.03 to 3.97)*
Total mortality	25	0.60 (0.29-1.22)	*0.65 (0.35 to 1.21)*
CHD mortality	25	0.52 (0.27-0.99)	*0.59 (0.32 to 1.09)*
CHD events	25	0.58 (0.29-1.17)	*0.58 (0.31 to 1.06)*
Total mortality	3	0.76 (0.46-1.24)	*0.80 (0.55 to 1.16)*
CHD mortality	3	0.66 (0.45-0.96)	*0.73 (0.50 to 1.06)*
CHD events	3	0.73 (0.46-1.18)	*0.71 (0.50 to 1.01)*

5.2.10 Extensions to the Bayesian variance component model

Missing/uncertain covariate data

For four groups of patients the baseline risk was unknown because the number of CHD deaths was not available. In a classical model these patients are excluded, however WinBUGS can deal with the missing data, for example, by placing vague priors, centred at mean baseline risk across studies, on the relevant nodes. A further refinement would be to build a model to predict the number of CHD deaths based on the total mortality figures from the trials where both were available, however this was not pursued here. A third approach is possible in this instance. The study for which there are no CHD mortality figures is number R54. However, since, there was only one death in total in the control group of this trial, only two values are possible for the CHD mortality

deaths: 0 or 1. The two baseline risk values corresponding to 0 and 1 deaths (after centring) are -13 or -23. It would be sensible to randomly impute one of these two values into the model on every iteration. This can be achieved in WinBUGS, by generating a random number from a Uniform(-1,1) distribution, then using the *step* function to generate a variable (*int*) which takes the value 0 if the random number is <0 and 1 if it is >0. The baseline risk for study R54 can then be defined as: *int**(-10)-13. Although in this instance, the amount of missing data is so small that such methods have only a minute effect on the parameter estimation, there is potential for their use in instances where larger amounts of data are missing. See Section 7.7 for a more detailed example illustrating how the impact on the uncertainty in meta-analysis data can be explored.

Allowing original binary data to be used

An extension to the model, easily implemented in WinBUGS, allows modelling of binary outcomes using the original aggregated outcome data opposed to a summary effect size and corresponding standard error. Such a specification has several advantages which include: 1) the assumption that that each outcome is normally distributed is no longer required; and 2) no continuity correction is required for groups of patients which have zero events. Hence, such an extension would be particularly valuable for meta-analyses of rare events; this issue is considered in Chapter 7. This extension borrows directly from the standard Bayesian meta-analysis model for binary outcomes described by Smith et al. (Smith et al. 1995b) which models the events in each group using Binomial distributions (see Section 2.3.3) and is similar in spirit to the extension of the generalised synthesis model in Section 5.1 which also modelled the binary data directly.

Specifying the model in this way for the cholesterol example produced similar results to previously. The largest differences were for parameters where little data were available. Autocorrelation was similar to previously. The theoretical advantages make this model more desirable if data are available since assumption of normality of effects from individual patient groups is no longer required (see Chapter 7).

5.2.11 Incorporating the non-randomised studies into the variance components model

One of the motivations for exploring this model was the potential to include non-randomised evidence. Since this approach models each group of patients of a study separately, it provides a natural framework for including non-comparative single arm studies. If this is done the model can be viewed as a generalisation of the model of Begg and Pilote (Begg and Pilote, 1991), which allowed the inclusion of single arm studies which used historical controls (see Section 3.4).

Uncontrolled studies investigating a single treatment of interest could be included in a straightforward manner. For example, in the computer-based reminder system meta-analysis, a hospital audit of the use of a single reminder method could be included directly. Unfortunately, the inclusion of the non-randomised evidence for the cholesterol lowering meta-analysis is less straightforward because no intervention was administered in these studies; rather, mortality rates in groups with different cholesterol levels were reported. Perhaps the simplest way of including them is to use the number of events for each outcome over the whole cohort population in each study and to give the observational studies the same treatment code as that for the placebo arms in the trials. The covariate baseline risk can be calculated for these studies and included in the model. This is important because the large cohort studies included generally healthy individuals and hence levels of baseline risk will generally be lower than those in many of the trials.

This analysis was carried out, incorporating evidence from the ten largest non-randomised studies described in Chapter 4. Table 5.11 reports the odds ratios for receiving diet versus statins (reported previously) and statins versus placebo for high, medium and low levels of baseline risk for models including the RCTs only and RCTs and non-randomised studies. For contrasts both directly including the non-randomised evidence (statins v placebo) and those not including it (statins v diet) there was little change in the estimated odds ratios and credible intervals in this example.

SPECIAL NOTE

THIS ITEM IS BOUND IN SUCH A MANNER AND WHILE EVERY EFFORT HAS BEEN MADE TO REPRODUCE THE CENTRES, FORCE WOULD RESULT IN DAMAGE

Table 5.11 Comparison of model estimates including and excluding the non-randomised evidence

Outcome	Baseline risk (CHD deaths per 1000 person years)	Statin versus diet - OR (95% CI)		Statin versus placebo - OR (95% CrI)	
		Randomised evidence only	Randomised and Non-randomised evidence	Randomised evidence only	Randomised and Non-randomised evidence
Total mortally	80	0.40 (0.04 to 4.48)	0.44 (0.05 to 5.25)	0.28 (0.03 to 3.06)	0.31 (0.03 to 3.75)
CHD mortality	80	0.36 (0.03 to 4.01)	0.39 (0.04 to 4.75)	0.25 (0.03 to 2.72)	0.27 (0.03 to 3.32)
CHD events	80	0.35 (0.03 to 3.97)	0.39 (0.04 to 4.66)	0.26 (0.03 to 2.79)	0.28 (0.03 to 3.4-0)
Total mortality	25	0.65 (0.35 to 1.21)	0.66 (0.37 to 1.26)	0.55 (0.30 to 1.01)	0.54 (0.30 to 1.03)
CHD mortality	25	0.59 (0.32 to 1.09)	0.60 (0.34 to 1.14)	0.49 (0.26 to 0.90)	0.48 (0.27 to 0.91)
CHD events	25	0.58 (0.31 to 1.06)	0.59 (0.33 to 1.11)	0.50 (0.27 to 0.92)	0.49 (0.28 to 0.93)
Total mortality	3	0.80 (0.55 to 1.16)	0.80 (0.55 to 1.13)	0.73 (0.50 to 1.05)	0.70 (0.47 to 0.98)
CHD mortality	3	0.73 (0.50 to 1.06)	0.72 (0.49 to 1.03)	0.65 (0.44 to 0.95)	0.62 (0.42 to 0.89)
CHD events	3	0.71 (0.50 to 1.01)	0.70 (0.50 to 0.98)	0.66 (0.46 to 0.94)	0.63 (0.44 to 0.88)

Including the observational evidence in this way is conceptually very different from the three level hierarchical modelling approach of Section 4.10. There, a measure of relative benefit was derived from the observational studies using regression methods. The influence the observational evidence has on the outcomes is also considerable. In the hierarchical model, the observational studies were highly influential due to their relative magnitude, though less influential than if they had been incorporated directly into a standard random effect model. Here, because they only provide information on one group, the placebo or control group, they have much less influence, and only marginally change the parameter estimates in the model despite their relative size.

Further approaches to including the non-randomised evidence using a single arm modelling approach are conceptually possible. One possibility would be to model total serum cholesterol level as an intermediate outcome. If this were done, then the observational studies could be used to provide information on the mortality and CHD event rates for different levels of blood serum cholesterol. The RCTs, and other studies could be used to model the typical degree of cholesterol lowering achieved by the different interventions. Linking these two pieces of information could allow inferences to be made about the effect of lowering cholesterol levels on the mortality and CHD event rate outcomes. Models using intermediate outcomes are described in the context of the confidence profile method, (Eddy et al. 1992) are not pursued but fit naturally into the way models are specified in WinBUGS; a synthesis which does included intermediate outcomes is described in Chapter 8.

5.2.12 Further consideration of different multiple outcome meta-analysis models

The model of DuMouchel allows the modelling of multiple outcomes simultaneously, and allows the modelling of binary, or continuous outcomes, provided they are all the same in any one particular model. For continuous outcomes, other models allowing the simultaneous modelling of outcomes have been suggested. Perhaps the most general and flexible of these is the multiple outcome random effect meta-regression model of Berkey et al. (Berkey et al. 1998) (see Section 5.3). A major way this model differs from that of DuMouchel is that the correlations between outcomes are explicitly

modelled, hence the model can be considered as truly multivariate. The model of DuMouchel assumes that variances for each outcome are equal, and that all covariances are equal, hence only two parameters have to be estimated. DuMouchel notes (personal communication) that for the computer based reminder system meta-analysis where there are 6 outcomes, a fully multivariate model would require the calculation of a 21-parameter covariance matrix. Unless there was information from external sources as to the nature of this matrix it would be hard to estimate with the data available. However it would be possible to try and fit intermediate models with between 2 and 21 parameters in the covariance matrix.

A multivariate model for binary data has never been developed for meta-analysis. In many instances it makes little sense because binary events are commonly competing alternatives, e.g. death from cardiovascular disease, non-cardiovascular death or survival. Occasionally, however, where responses are not competing, it would be possible. For example, a pain relief treatment for rheumatoid arthritis could be assessed to examine effectiveness (pain relief/no pain relief) at different sites of the body, each of which would be correlated. Additionally, no meta-analysis model, truly multivariate or not, has been developed which allows the combination of binary and continuous outcomes simultaneously. Such a model would clearly have uses since outcomes are rarely all measured on the same scale. Recent developments in the cost-effectiveness literature have begun to address a similar problem, where clinical outcomes are binary and costs are, obviously, continuous and the correlation between both needs to be taken into account. (O'Hagan and Stevens, 2001)

A Bayesian implementation of a multivariate model for continuous outcomes is described in section 5.3.

5.2.13 Summary/discussion of the cholesterol synthesis and the approaches explored

Two different statistical models were used to combine the randomised and observational evidence. The first was a three level hierarchical model, which accounts for heterogeneity between study types, as well as within studies of the same type. The

second model used a variance components model that reduced all studies to individual study arms for the purposes of modelling. This framework is very flexible in the study designs it can synthesise, but it should not be forgotten that the benefits of randomisation are largely lost when the model is used. An advantage of the variance component approach is that it allows comparison of different treatment options at levels of baseline risk, even if such comparisons were not made directly in the trials.

In summary, it was possible to synthesise the data using both modelling approaches, and both produced sensible answers, nonetheless, it is difficult to see how such models could be validated empirically. It is important to note that the two types of studies being combined in this synthesis are perhaps more dissimilar than in many other areas. Here, the observational evidence was not based on patients given interventions (whereas elsewhere such before and after studies or audits as will often exist – see Chapter 6), but based on aetiological data. In areas where single arm studies are being compared with comparative studies the variance components model offers an appealing amount of flexibility.

An important assumption that has been made in this analysis should not be overlooked that the cohort studies and RCTs are estimating comparable phenomena. The RCTs investigate the effect of 'artificially' lowering cholesterol levels, while the cohort studies observe the incidence of CHD for people with different cholesterol levels. Clearly these are not the same, although, if the reduction in risk from lowering cholesterol levels is present immediately after the reduction then they could be considered as broadly equivalent. Care is needed however, in interpreting the evidence of efficacy if life-threatening side effects are associated with the interventions.

Any systematic review/meta-analysis is a lengthy process. If a broader range of evidence is considered in that review than normally, then naturally it will take longer still. The literature on cholesterol and its effect on CHD is very large indeed. In this example only ten extra non-randomised studies were considered in addition to the randomised evidence. This was done largely due to practical time constraints, although it should be appreciated that there are no theoretical limitations on the amount of evidence combined in the modelling approaches used.

Another interesting issue is the effect of including other sources of evidence. For example, the estimates from the international survey studies identified by Law et al. [R28] could have been included (assuming CHD mortality estimates are available). For the three-level hierarchical model, the uncertainty at the between study type level is heavily influenced by the number of sources of evidence considered.

Another issue, related to the inclusion of more evidence, is the specification of prior distributions for model parameters. All the priors used in the two synthesis models were intended to be non-informative, and hence, influenced the analysis very little. However, this need not be the case. The previous implementation of the three level model (Prevost et al. 2000) showed how constraints could be placed on the model through the specification of the priors. In this previous implementation the constraint that the randomised studies are less biased than the non-randomised was enforced. These had the effect of the pooled result being closer to that for the RCTs with a tighter confidence interval. Possibly this method could be used in this example as a way of avoiding the somewhat arbitrary relationship between the number of types of evidence and the width of the confidence interval for the pooled estimate, and as a way of restricting the influence on evidence from, possibly, weaker designs.

A further consideration is the aim of the cross design synthesis analysis. It has been stated that there are two kinds of meta-analyses, those which are confirmatory and those which are exploratory. The intention of the first is to produce an effect estimate that is more accurate than those from any individual study. The latter's aim is to explore why results from different studies differ, possibly with the intention of generating hypotheses to test with new studies. Below consideration is given these two distinct types of analysis with respect to cross design synthesis. In the introduction it was mentioned that, in certain situations, the randomised evidence may be insubstantial, and in such circumstances it may be beneficial to consider other sources of evidence also, so that the role of such an analysis would be confirmatory. On the other hand, a further use for cross design synthesis could be in an exploratory role. Here, rather than trying to obtain one overall pooled result, more emphasis is placed on the exploration of variation between studies. The aim of a particular analysis may influence which studies are to be included. Perhaps in an exploratory example, including a broader selection of evidence is more appropriate. For example, considering the cholesterol example, the effect of

cholesterol reduction could be extrapolated beyond patient types included in trials by including data from observational studies of patient groups atypical in the RCTs. Thus, although it has been established that cholesterol lowering is an effective treatment for high risk patients with high levels of cholesterol, it is currently undecided whether there are benefits from reducing cholesterol levels in patients with already low cholesterol levels. (Rubins, 1995) Perhaps, by including data from cohorts with low cholesterol levels, an estimate of the likely effect of reduction in such patients could be possible. Similarly, no RCTs included solely women, so, it was hard it ascertain if there was a gender effect. Several observational studies included data individually on women, and hence could assist in ascertaining if there is a gender effect, and if there is, producing an effect estimate for women.

The issue of publication bias was discussed in the introduction and was touched upon briefly in analysis of the RCTs separately. Chapter 6 considers publication bias in a cross-design synthesis framework, where the possibility that different study types are subject to different publication bias mechanisms is explored.

It is worth considering in what ways an analysis such as this could be expanded further. An appealing use of cross-design synthesis is the idea of explaining the science underlying the trials and observational studies. (Rubin, 1992) It seems logical that information pertaining to the biological mechanisms at work are relevant to an investigation such as this. How such evidence should be included is difficult to imagine, although, evidence from related observational experiments offer an increase in knowledge and hence must also provide some knowledge gain if it can be utilised. For example, many other studies have been carried out examining the degree to which different diets lower cholesterol levels (e.g. see Tang et al.(Tang et al. 1998)). These studies often do not consider event outcomes and hence cannot be included directly in a model investigating mortality, but, they do provide possibly more accurate information on the effect particular interventions have on the reduction of cholesterol levels than the RCTs which were included. Using information such as this, it can be perceived that a more complex multi-stage model could be devised, where different studies provided evidence for different parts of the process in question. In this way different aspects of the cholesterol reduction/CHD process could be modelled, which may provide deeper insight into the underlying science.

Finally, it is worth considering an incident which occurred during the final stages of this theses completion. On the 8th August 2001 Cerivastatin was withdrawn by the manufacturer. The reason for this was that there had been 31 deaths in the US from severe rhabdomyolysis, an adverse event, in patients taking the drug. The issue of adverse events was not dealt with explicitly in the analyses described in this and the previous chapter, but consideration to this topic is given in other settings later in Chapters seven and eight where it is acknowledged that observational evidence may be particularly valuable in establishing risks of adverse events since trial data will usually be inadequate.

5.3 Further modelling of multiple continuous outcome measures

This section briefly considers further modelling issues regarding the synthesis of evidence. Although the majority of these have been implemented in a meta-analysis context, they are just as relevant in a generalised synthesis framework. During the past decade many extensions to the basic fixed or random effect models for meta-analysis have been developed. (Sutton et al. 2000a) Many such developments allow more data to be included in a meta-analysis and allow more appropriate analysis of such data. This section discusses an extension to the basic random effects meta-analysis model, and is implemented using Bayesian methods.

A classical literature exists on methods to combine multiple outcome measures simultaneously. When multiple outcomes have been reported for all or a proportion of the relevant studies, it is common practice to conduct separate meta-analyses for each outcome measure, or ignore all but one outcome. (Hedges and Olkin1985)

However, such analyses make questions such as "Does a treatment have larger effects on some outcomes than on others?" and "Does the duration of treatment affect different outcomes differently?" hard to answer. (Raudenbush et al. 1988) Further, these approaches are not optimally efficient as they do not use statistical information about the errors of estimation contained in the other estimated effect sizes. (Gleser and Olkin, 1994) The approach allows multiple continuous outcome measures to be combined using a single model in an alternative way to that described by DuMouchel (DuMouchel, 1998) and considered in detail in section 5.2. Related work has also been carried out combining outcomes measured on different scales, (Whitehead et al. 1999) but this is not considered further in this thesis.

A series of models of increasing sophistication have been described which combine multiple outcomes in a single model. The initial work on this was done by Hedges and Olkin. (Hedges and Olkin1985) Their model assumes the same outcomes are measured on all studies to be combined, and it was superseded by a more general model which allows different outcomes (and different numbers of outcomes) to be measured across studies. (Raudenbush et al. 1988) It also allows different covariates to be included to explain variation in effect sizes for each outcome. An alternative formulation of this model has been given by Glesser and Olkin (Gleser and Olkin, 1994) which corrects a mistake in the previous two approaches. (Hedges and Olkin 1985; Raudenbush et al. 1988) These approaches have several drawbacks, namely, not incorporating any random effects, and only allowing outcomes to be reported as standardized or scale-free differences. Berkey et al. (Berkey et al. 1995; Berkey et al. 1996) provide a generalisation which keeps the measured outcomes in their original units. Recently, Berkey et al. (Berkey et al. 1998) provide a further enhancement by developing models which incorporate random effects. All these models require the correlations between outcomes to be known or estimable, a potential limitation which is discussed below.

We consider a Bayesian formulation of the most general of these models, that of Berkey et al., (Berkey et al. 1998) of which all the previous models can be viewed as simpler sub-models. We apply this to the data used by Berkey, (Berkey et al. 1998) originally reported by Antczak-Bouckoms et al. (Antczak-Bouckoms et al. 1993) from five published trials comparing outcomes of surgical and non-surgical treatments for medium-severity periodontal disease, one year after treatment. Two outcomes are

considered, probing depth (PD) and attachment level (AL), both measured in mm. The mean difference between the two groups, surgical (S) and non-surgical (NS) is the outcome of interest. The covariate year of publication (centred at 1983) is included. For each trial, the within-trial covariance matrix (S_i) of the two outcomes (means) is also required. The data used are reproduced in Table 5.12.

Table 5.12 Results from 5 trials comparing surgical and non surgical treatments for medium-severity periodontal disease, one year after treatment (reproduced from (Berkey et al. 1998))

Trial	Publication Year	Number of patients	Improvement in		S_i	
			Probing depth (S – NS)	Attachment level (S - NS)	PD	AL
1	1983	14	+0.47	-0.32	0.0075	0.0030 0.0030 0.0077
2	1982	15	+0.20	-0.60	0.0057	0.0009 0.0009 0.0008
3	1979	78	+0.40	-0.12	0.0021	0.0007 0.0007 0.0014
4	1987	89	+0.26	-0.31	0.0029	0.0009 0.0009 0.0015
5	1988	16	+0.56	-0.39	0.0148	0.0072 0.0072 0.0304

The general form of the model used is (Berkey et al. 1998):

$$\mathbf{y}_i = X_i\beta + \delta_i + \mathbf{e}_i, \qquad (5.4)$$

where: y_i is a vector of p outcomes ($p = 2$ in the example) reported by trial i; X_i is a matrix containing the observed trial-level covariates for trial i (year of publication in the example); β is the vector of regression coefficients to estimate; δ_i is a vector of p random effects associated with trial i. The cov(δ_i) = D needs to be estimated, and it is

assumed that the δ_i arise from a multivariate Normal distribution (MVN (**0**,D)); and e_i is the vector of random sampling errors within trial i, having $p \times p$ covariance matrix S_i, which is assumed known (but is usually estimated/reported by the individual trials and is included in the example dataset in Table 5.12). If each n_i is sufficiently large, then the vector e_i is approximately MVN (**0**,S_i). This leads to

$$\text{cov}(y_i) = D + S_i \qquad \text{and} \qquad y_i \sim MVN(X\beta, D + S_i).$$

In the example both outcomes are available from all five studies, but with a slight modification, trials can be included if they report only a subset of outcomes. Additionally, the same covariate is used for both outcome measures, but again this need not be the case.

5.3.1 Bayesian specification in WinBUGS

Model 8.1 can be specified using a Bayesian formulation in WinBUGS with the addition of prior distributions. Since WinBUGS works with precisions, the covariate matrices need to be inverted. Vague priors are specified throughout. Normal(0,1000) distributions are specified for the pooled estimates of both outcomes (notated out_1 and out_2 below); a $Wishart\left(\begin{bmatrix} 0.02 & 0 \\ 0 & 0.2 \end{bmatrix}, 2\right)$ prior distribution is placed on the random effect precision matrix. The degrees of freedom of this distribution (in this case 2) need to be at least as large as the rank of the covariance matrix on which it is being placed, but the smaller the number given, the less informative the distribution. (Speigelhalter et al. 2000a) The associated prior matrix suggests the magnitude of the variance terms, but makes no assumptions about the covariances. Finally, Normal(0,1000) prior distributions are placed on the regression coefficients for publication year for both outcomes (notated β_1 and β_2 below). A burn in of 1000 iterations followed by a run of a further 100,000 iterations was carried out. Table 5.11 displays the results of this model, together with the results from the equivalent classical model and from combining each outcome separately.

Point estimates for the four model parameters, out_1, out_2, β_1 and β_2 are approximately the same for all five models. However, the standard errors of all parameters in the Bayesian model were consistently larger than those of their classical counterparts. This can be explained by the incorporation of the uncertainty in estimating the random effect covariance matrix in the Bayesian model. Berkey et al. compared the random effect models in Table 5.13 with fixed effect estimates (not shown) and found the standard errors of the fixed effects ones to be considerably smaller than those for the random effects. This lead the authors to state that the choice between fixed and random effect models was more crucial than the consideration of both outcomes simultaneously. In a similar vein, since the standard errors of the coefficients from both Bayesian models are larger than any of the classical ones, the choice between using a Bayesian or a Classical approach has more influence on the results, in this example, than the choice between modelling outcomes individually or simultaneously. However, there clearly are gains in efficiency using a (Bayesian) multiple outcome model.

SPECIAL NOTE

THIS ITEM IS BOUND IN SUCH A MANNER AND WHILE EVERY EFFORT HAS BEEN MADE TO REPRODUCE THE CENTRES, FORCE WOULD RESULT IN DAMAGE

Table 5.13 Results of classical random effects and Bayesian meta-analysis modelling including regression terms. Estimates reported for outcomes combined individually and in a multivariate model

Bayesian model		Classical random effects model (as reported by Berkey et al. 1998))		
Separate outcomes	Multiple outcomes	Separate outcomes	Multiple outcome GLS*	Multiple outcomes MML**
D matrix (SE for Bayesian models)				
$PD\begin{bmatrix} 0.052(0.442) & 0 \\ AL & 0 & 0.133(0.962) \end{bmatrix}$	$\begin{bmatrix} 0.038(0.042) & 0.021(0.040) \\ 0.021(0.040) & 0.070(0.069) \end{bmatrix}$	$\begin{bmatrix} 0.020 & 0 \\ 0 & 0.036 \end{bmatrix}$	$\begin{bmatrix} 0.022 & 0.013 \\ 0.013 & 0.028 \end{bmatrix}$	$\begin{bmatrix} 0.008 & 0.009 \\ 0.009 & 0.025 \end{bmatrix}$
Outcome PD models				
$Y = out_1 + \beta_1 x$				
(SE(out_1)) (SE(β_1))				
0.360 + 0.003x	0.356 + 0.004x	0.363 + 0.005x	0.359 + 0.005x	0.348 + 0.001x
(0.112) (0.033)	(0.097) (0.028)	(0.073) (0.022)	(0.075) (0.022)	(0.052) (0.015)
Outcome AL models				
$Y = out_2 + \beta_2 x$				
(SE(out_2)) SE(β_2))				
-0.342 - 0.013x	-0.341 - 0.013x	-0.340 - 0.014x	-0.336 - 0.011x	-0.335 - 0.011x
(0.172) (0.050)	(0.129) (0.038)	(0.092) (0.028)	(0.083) (0.026)	(0.079) (0.024)

* Generalised least squares method

** Multivariate maximum likelihood method

5.3.2 Comparison of Berkey and DuMouchel models

The modelling approaches of Berkey and DuMouchel are both possible for datasets such as the dental data originally described by Antczak-Bouckoms et al. (Antczak-Bouckoms et al. 1993) and reanalysed above. Major distinctions between the two approaches are that 1) randomisation is not maintained with the DuMouchel approach; and 2) the correlations between outcomes are not required when using the model of DuMouchel, instead, the correlations are assumed to be the same between outcomes and are estimated using a variance component model.

In order to compare the two approaches the dental data is re-analysed using a variance components model. In order to use DuMouchel's model, the outcomes from both arms of the trial are required, and fortunately were available,(Antczak-Bouckoms et al. 1993) compared to the difference between arms used in the model of Berkey et al. The specific model fitted is given below which is a modification of equation (5.3).

$$y_{ikm} = outcome_m + treatment_k + year_i + (year \times treatment)_{ik} + (r.outcome \times treatment)_{mk} + r.study_i + (r.study \times outcome)_{im} + (r.study \times treatment)_{ik} + \varepsilon_{ikm} \quad (5.5)$$

where y is the observed outcome, i indexes the *1..5* studies in the analysis, m indexes the *2* different outcomes considered, and k indexes the *2* treatments being compared in the analysis. Terms starting 'r.' are fitted as random effects, while the others are treated as fixed. Although this model may appear overtly complex for modelling only five studies, a sensitivity analysis removing interaction terms was carried out and the fixed effect estimation remained robust over several related models.

The parameter estimates that result from fitting this model are presented in Table 5.14.

Table 5.14 Parameter estimates fitting variance component model to the dental meta-analysis data

Parameter	Estimate/ Standard error	P-value
Random effects		
r.outcome × treatment	0.105	
r.study	< 0.001	
r.study × outcome	0.036	
r.study × treatment	0.015	
Fixed effects		
Outcome		
Probing depth	1.515 (0.302)	0.13
Attachment level	0.334 (0.302)	0.47
Treatment		
Surgical	-0.004 (0.337)	0.99
Covariate		
Year (centred)	-0.055 (0.026)	0.10
Year×surgical	-0.004 (0.027)	0.88

To allow comparison with the result of the multivariate model reported in Table 5.13 treatment contrasts are constructed for the variance component model; these are reported in Table 5.15.

Table 5.15 Treatment contrasts for variance component model applied to dental meta-analysis data

Contrast	Estimate (s.e)
Variance component model (Classical)	
Surgery v no surgery for probing depth in 1984	0.318 (0.099)
Surgery v no surgery for attachment level in 1984	-0.327 (0.095)
Independent outcomes (Clasical REML)	
Surgery v no surgery for probing depth in 1984	0.363 (0.073)
Surgery v no surgery for attachment level in 1984	-0.340 (0.092)
Multivariate model (Classical MML)	
Surgery v no surgery for probing depth in 1984	0.348 (0.052)
Surgery v no surgery for attachment level in 1984	-0.335 (0.079)

In Table 5.15 contrasts are calculated for the year 1984. This year was chosen because the covariate for publication date was centred on 1984 allowing treatment effects to be estimated directly without having to take covariate effects into consideration. Hence, the intercept terms for the independent and multivariate models from Table 5.13 (reproduced in Table 5.15) can be compared with the contrasts from the variance components model as reported in Table 5.15. It can be seen that the treatment effects for both outcomes are quite similar from the variance component model when compared to the independent and multivariate model, however the standard errors of the variance components model estimates are larger than even the standard independent outcome analysis confirming that while such a model may allow flexibility in modelling, no efficiency is gained directly by modelling multiple outcomes simultaneously. This is in contrast to the truly multivariate model in which moderate gains in efficiency can be gained when correlations between outcomes are known. However, the variance component model does allow more flexibility in the designs of the study it can synthesise, and it was this increase in flexibility not efficiency which motivated its initial development.

The above has been only a very brief comparison of two competing modelling strategies. Further exploration may be valuable particularly with regard to the degree of bias associated with estimation of treatment effects from each model.

5.3.3 Discussion including extensions/modifications to the model

In the above example all studies included had reported both outcomes of interest, but this may not always be the case. Where studies report one, or a subset of a larger number of outcomes, it will often be desirable to include them in the analysis. Conceptually this is simple to do in WinBUGS, whereas it would be less trivial from a classical standpoint.

Wishart priors are notoriously tricky to comprehend and use (see section 7.5.2). An alternative formulation using the product of normal distributions could be used. This idea is discussed in more detail later in this thesis (section 7.5), however the possibility for its use here is noted. This approach allows univariate normal prior distributions to be placed on the random effect covariance terms. Implementing this modification produced similar, but not identical, estimates when compared with the original formulation. A disadvantage of this method is that it is not possible to produce an estimate of the random effect covariance matrix.

The model above assumed that the correlation, and hence the covariances between outcomes were known. In many situations this will not be the case. Often, only a proportion of studies will report the correlation necessary to calculate the covariance of outcomes. In other instances, estimates of the correlation may not be available for any of the studies, but evidence from external sources may be available from which it can be estimated. (Strube, 1985) Obviously, if little is known about any of the correlations there is little to be gained efficiency-wise from including multiple outcomes in a model, however it may still be useful to do so for other reasons, such as the exploration of covatiates.

In a Bayesian analysis, for studies in which the correlation between outcomes is unknown, it would be possible to i) include the results ignoring the correlations (as

discussed above), ii) specify the correlation parameters, assign no data to them but derive prior distributions for them, possibly based on the studies in the analysis for which there is full covariance information. In the latter option, the modelling will provide an estimate of the unknown correlation.

This section has described how multivariate meta-analysis models for continuous outcomes can be modelled from a Bayesian perspective. Such an approach is relatively straightforward, except perhaps for the specification of a prior distribution for the parameters distributed multivariate normally. In certain situations a Bayesian formulation may be preferable over a classical approach if prior information is available, especially for the correlation between outcomes.

5.4 Estimating indirect comparisons

This section considers a random effect model for synthesising data from RCTs which have made different treatment comparisons. Such models may be of interest if there is little direct randomised evidence for the particular comparison of interest. For example, consider a scenario where there are trials of treatment A v treatment C, and trials of treatment B v treatment C, but few or no trials of treatment A v treatment B, the comparison of interest. In such a situation, estimates of the effect of A v B could be derived by simply constructing an odds ratio using the results from the A arm of the A v C trials and the B arm of the B v C trials, but the benefit of randomisation is lost if this is done, and biased results could result due to differences in the characteristics of patients enrolled into the different trials. (Song et al. 2000a)

More sophisticated is the approach of Bucher et al. (Bucher et al. 1997) where the indirect comparison of A v B is adjusted by results of their direct comparisons with a common intervention C, i.e.

$$\ln OR'_{AB} = \ln OR_{AC} - \ln OR_{BC}, \qquad (5.5)$$

where $\ln OR'_{AB}$ is the indirect comparison of interest, and $\ln OR_{AC}$ and $\ln OR_{BC}$ are the direct comparisons available. The variance of this estimate is

$$Var(\ln OR'_{AB}) = Var(\ln OR_{AC}) + Var(\ln OR_{BC}). \quad (5.6)$$

While this model relaxes the assumption that the patients have the same distributions of characteristics (i.e. different studies sampled patients from the same population), it still assumes that the true underlying odds ratio is the same across trials, so while the absolute efficacy can differ the relative efficacy is constant (i.e. the assumption of the standard fixed effect model). Hence, a limitation of such a model is that it assumes that there is no between study heterogeneity in either of the indirect comparisons. Higgins and Whitehead (Higgins and Whitehead, 1996) describe a Bayesian model which allows the inclusion of a third treatment and heterogeneity parameters in a meta-analysis model. This leads to the estimation of all three treatment effects using both direct and indirect comparisons. The increased flexibility of allowing the inclusion of trials with three arms does come with the drawback that it is necessary to assume the magnitude of the between subject heterogeneity is assumed equal for each of the three comparisons being estimated.

If none of the studies have the three treatment arms of interest, the model can be simplified, and the equality assumption of the between study variances relaxed for the different comparisons. It is this specific model which is considered below.

5.4.1 Example: meta-analysis of RCTs for prevention of Pneumocystis carinii pneumonia in HIV infection

In order to illustrate the Bayesian model described below, the meta-analysis, originally used by Bucher et al. (Bucher et al. 1997) to illustrate their method of combining indirect comparisons, is described. This dataset consists of twenty-two RCTs investigating agents for the prevention of Pneumocystis carinii pneumonia in HIV infection. The comparison of interest is Trimethoprim-sulfamethoxazole (TMP-SMX) vs. dapsone/pyrimethamine (D/P) for which eight of the trials provide direct

comparisons. The remaining fourteen trials provide information for indirect comparisons since they compared one of the two treatments of primary interest with a third – aerosolised pentamidine (AP) (nine compared this to TMP-SMX and five to D/P). The results for these trials are provided in Table 5.14.

Table 5.14 RCTs providing direct or indirect evidence concerning the comparative effect of Trimethoprim-sulfamethoxazole (TMP-SMX) vs dapsone/pyrimethamine (D/P) (compared against each other or against aerosolised pentamidine (AP)) for the prevention of Pneumocystis carionii pneumonia in HIV-infected subjects

(The reference list for the original studies is available elsewhere (Bucher et al. 1997)).

Trial	(Treatment A) TMP-SMX (events / total number of subjects)	(Treatment B) AP (events / total number of subjects)	(Treatment C) D/P (events / total number of subjects)
Rozenbaum 1991	0/29	1/27	
Hardy 1992	14/154	36/156	
Schneider 1992	0/142	6/71	
Smith 1992	3/27	6/26	
Michelet 1993	1/53	4/55	
May 1994	2/108	5/106	
Stellini 1994	0/26	2/23	
Nielsen 1995	1/47	8/48	
Rizzardi 1995	5/95	6/101	
Slavin 1992		8/46	9/50
Girard 1993		10/176	10/173
Torres 1993		15/152	15/126
Opravil 1995		13/242	12/291
Salmon 1995		12/102	5/92
Antinori 1992	1/66		9/63
Mallolas 1992	3/107		8/116
Tocchetti 1994	0/15		1/15
Bozzette 1995	42/276		41/288
Blum 1992	1/39		1/47
Podzamcer 1993	3/81		13/85
Podzamczer 1995	0/104		6/96
Sirera 1995	6/115		9/105

5.4.2 Bayesian random effect model for estimating indirect comparisons

Initially, a model to combine only the indirect evidence is described. This is conceptually simple since a standard random effects estimates of the treatment effect for A v B and B v C are sought using model (2.18). The difference between these pooled log odds ratios is the estimate for the log odds ratio of treatment comparison A v C. The model is given algebraically below for clarity.

$$
\begin{aligned}
&rc.ab_i \sim Bin[pc.ab_i, nc.ab_i] \qquad rt.ab_i \sim Bin[pt.ab_i, nt.ab_i] \qquad i = 1.....9 \\
&\text{logit}(pc.ab_i) = \mu.ab_i - d.ab_i/2 \qquad \text{logit}(pt.ab_i) = \mu.ab_i + d.ab_i/2 \\
&\qquad d.ab_i \sim \text{N}[\phi.ab, \tau^2.ab] \qquad \mu.ab_i \sim Normal[0,10^5] \\
&\qquad \phi.ab \sim \text{Normal}[0,10^6] \qquad \tau^2.ab \sim \text{InverseGamma}[0.001,0.001] \\
&rc.bc_j \sim Bin[pc.bc_j, nc.bc_j] \qquad rt.bc_j \sim Bin[pt.bc_j, nt.bc_j] \qquad j = 1.....5 \\
&\text{logit}(pc.bc_j) = \mu.bc_j - d.bc_j/2 \qquad \text{logit}(pt.bc_j) = \mu.bc_j + d.bc_j/2 \\
&\qquad d.bc_j \sim \text{N}[\phi.bc, \tau^2.bc] \qquad \mu.bc_j \sim Normal[0,10^5] \\
&\qquad \phi.bc \sim Normal[0,10^6] \qquad \tau^2.bc \sim InverseGamma[0.001,0.001] \\
&\qquad\qquad \text{lnor.ac} = \phi.ab - \phi.bc
\end{aligned}
\tag{5.7}
$$

Extensions *ab* and *bc* and *ac* represent treatment comparisons A v B, B v C and A v C respectably. *rc* and *rt* indicate the number of events in the treatment and control groups and *nt* and *nc* the total number of patients in the treatment and control groups. All priors are intended to be vague.

The results of fitting this model are given in Table 5.15 together with the results of combining the direct comparisons using the Bayesian meta-analysis model for binary data outlined in equation 2.11. Additionally, the fixed effect results reported by Bucher et al (Bucher et al. 1997) are included for comparison.

Table 5.15 Results of direct and indirect comparisons for prevention of Pneumocystis carinii pneumonia in persons with HIV

	Odds ratio (95% CI/CrI)	**τ^2 TMP-SMX vs D/P (95% CI/CrI)**	**τ^2 TMP-SMX vs AP (95% CI/CrI)**	**τ^2 D/P vs AP (95% CI/CrI)**
Direct comparison TMP-SMX vs D/P				
Fixed effect model	*0.64 (0.45 to 0.90)*	-	-	-
Bayesian random effect model	0.35 (0.09 to 0.80)	0.75 (0.03 to 6.88)	-	-
Indirect comparison TMP-SMX vs D/P				
Fixed effect model	*0.37 (0.21 to 0.65)*	-	-	-
Bayesian random effect model	0.23 (0.08 to 0.49)	-	0.07 (0.001 to 3.31)	0.02 (0.001 to 0.67)

The Bayesian random effect indirect comparisons model produces a smaller odds ratio than the fixed effect indirect estimate (0.23 compared to 0.37). The differences in these estimates can largely be explained by the different weightings given to the studies using fixed and random effect models. More surprising is the width of the confidence/credible intervals for these two estimates, as the Bayesian one is slightly narrower than the fixed effect one, despite including heterogeneity parameters of both indirect comparisons. If a Bayesian MCMC approach (based on equation (2.12)) is used then the pooled odds ratio is 0.26 (0.14 to 0.48), which is quite different from the classical fixed effect model. These differences can be partially attributed to the need for the use of continuity correction factors (adding 0.5 to data when there are 0 events

in a study) in the classical approach. The use of continuity correction factors are considered further in Chapter 7.

5.4.3 Extension: Including the direct comparison evidence

If direct evidence does exist, as is the case in the example discussed above, several possibilities exist for combining it with the indirect comparison evidence. In a classical framework, an estimate from a meta-analysis of the direct comparisons (using a standard meta-analysis model such as equation (2.18)) can be combined with the meta-analysis results of the indirect comparisons (using equation (8.4) described above) using a standard meta-analysis model (i.e. treating the two sources of evidence as two independent studies). Alternatively, a Bayesian model could be constructed which achieves essentially the same aim only in one step (this is discussed further below). Another option is to use a three-level hierarchical model such as those described in Section 5.1 to combine both sources of evidence. This approach is appealing since, discrepancy (or heterogeneity) between the comparison type estimates is accounted for in the estimation of an overall pooled treatment effect. The specification of a three level hierarchical model is outlined in equation (5.8) below.

Combining the direct evidence: RCT "exact" binomial model

$$rc.ac_k \sim Bin[pc.ac_k, nc.ac_k] \qquad rt.ac_k \sim Bin[pt.ac_{ij}, nt.ac_k] \qquad k = 1.....8$$

$$\log it(pc.ac_k) = \mu.ac_k \qquad \log it(pt.ac_k) = \mu.ac_k + d.ac_k$$

$$d.ac_k \sim \text{Normal}[\theta_1, \tau.ac^2]$$

$$\mu.ac_k \sim Normal(0,10^5) \qquad \tau^2.ac \sim InverseGamma(0.001,0.001)$$

Combining the indirect evidence: Indirect comparisons model

$$rc.ab_i \sim Bin[pc.ab_i, nc.ab_i] \qquad rt.ab_i \sim Bin[pt.ab_i, nt.ab_i] \qquad i = 1.....9$$

$$\log it(pc.ab_i) = \mu.ab_i - d.ab_i / 2 \qquad \log it(pt.ab_i) = \mu.ab_i + d.ab_i / 2$$

$$d.ab_i \sim \mathrm{N}[\phi.ab, \tau^2.ab] \qquad \mu.ab_i \sim Normal[0,10^5]$$

$$\tau^2.ab \sim \mathrm{InverseGamma}[0.001,0.001] \tag{5.8}$$

$$rc.bc_j \sim Bin[pc.bc_j, nc.bc_j] \qquad rt.bc_j \sim Bin[pt.bc_j, nt.bc_j] \qquad j = 1.....5$$

$$\log it(pc.bc_j) = \mu.bc_j - d.bc_j / 2 \qquad \log it(pt.bc_j) = \mu.bc_j + d.bc_j / 2$$

$$d.bc_j \sim \mathrm{N}[\phi.bc, \tau^2.bc] \qquad \mu.bc_j \sim Normal[0,10^5]$$

$$\phi.bc \sim Normal[0,10^6] \qquad \tau^2.bc \sim InverseGamma[0.001,0.001]$$

$$\phi.ab = \theta_2 + \phi.bc$$

Pooling both direct and indirect estimates

$$\theta_m \sim N(\phi, v^2) \qquad m = 1,2$$

$$\phi \sim \mathrm{N}(0,10^6) \qquad v^2 \sim IG(0.001,0.001)$$

The parameters of the indirect estimation part of equation (5.8) are the same as equation (5.7) and the parameters for pooling both direct and indirect estimates are defined in equation (5.1) when the basic form of this model was introduced. Parameters for combining the direct evidence part of the model should be self explanatory since this is the standard "exact" random effect meta-analysis model for binary data outlined in equation (2.12). If the assumption is made that both the direct and indirect comparisons are estimating exactly the same treatment effect then the

between study type variance parameter (ν^2) can be set to 0. Hence, a fixed effect model is used to combine the evidence from each study type having used a random effect model to combine studies of the same study type.

The results of combining the direct and indirect estimates using equation (5.8) are presented in Table 5.16. These results fall into a similar pattern to those observed when fitting the related model to the cholesterol data in Section 5.1. The indirect and direct pooled estimates are in broad agreement and shrunk towards each other. Since the number of studies of each comparison is relatively small there is considerable uncertainty in all between study within type variance components. With only two study type estimates (i.e. direct and indirect) there is a large degree of uncertainty in the between study type variance component. This uncertainty propagates across into the overall pooled estimate which has a very large credible interval. Although not shown, as for the cholesterol example the width of this credible interval is highly sensitive to the choice of prior, and all comments made regarding this in section 5.12 are pertinent here also.

Table 5.16 Transformed estimates and interpretation from hierarchical model combining direct and indirect evidence from HIV infection treatment data

Parameter	Interpretation	Median estimate (2.5 & 97.5 percentiles)
$e^{\phi.ab}$	OR AB direct comparison	0.28 (0.15 to 0.49)
$e^{\phi.bc}$	OR BC direct comparison	1.12 (0.69 to 1.77)
e^{θ_2}	OR AC indirect comparison (shrunken)	0.26 (0.11 to 0.49)
e^{θ_1}	OR AC direct comparison (shrunken)	0.29 (0.12 to 0.60)
e^{ϕ}	OR AC overall (using indirect and direct evidence)	0.27 (0.04 to 2.39)
$\tau^2.ab$	Between study variance for AB comparison	0.043 (0.001 to 1.89)
$\tau^2.bc$	Between study variance for BC comparison	0.020 (0.001 to 0.71)
$\tau^2.ac$	Between study variance for AC comparison (direct)	0.803 (0.103 to 5.67)
v^2	Between study type (direct and indirect) variance	0.048 (0.001 to 53.4)

Figure 5.13 displays graphically several of the different estimates derived from the different models discussed in combining the HIV infection treatment data. The first estimate is derived from a fixed effect model and the second a Bayesian random effect model combining just the direct evidence as reported in Table 5.15. The third and fourth estimates are derived from a fixed effect model and the Bayesian random effect model combining just the indirect comparison data respectively, also reported in Table 5.15. The fifth estimate is from combining all the evidence using a three level hierarchical model, as described above. Finally, the sixth estimate has not been

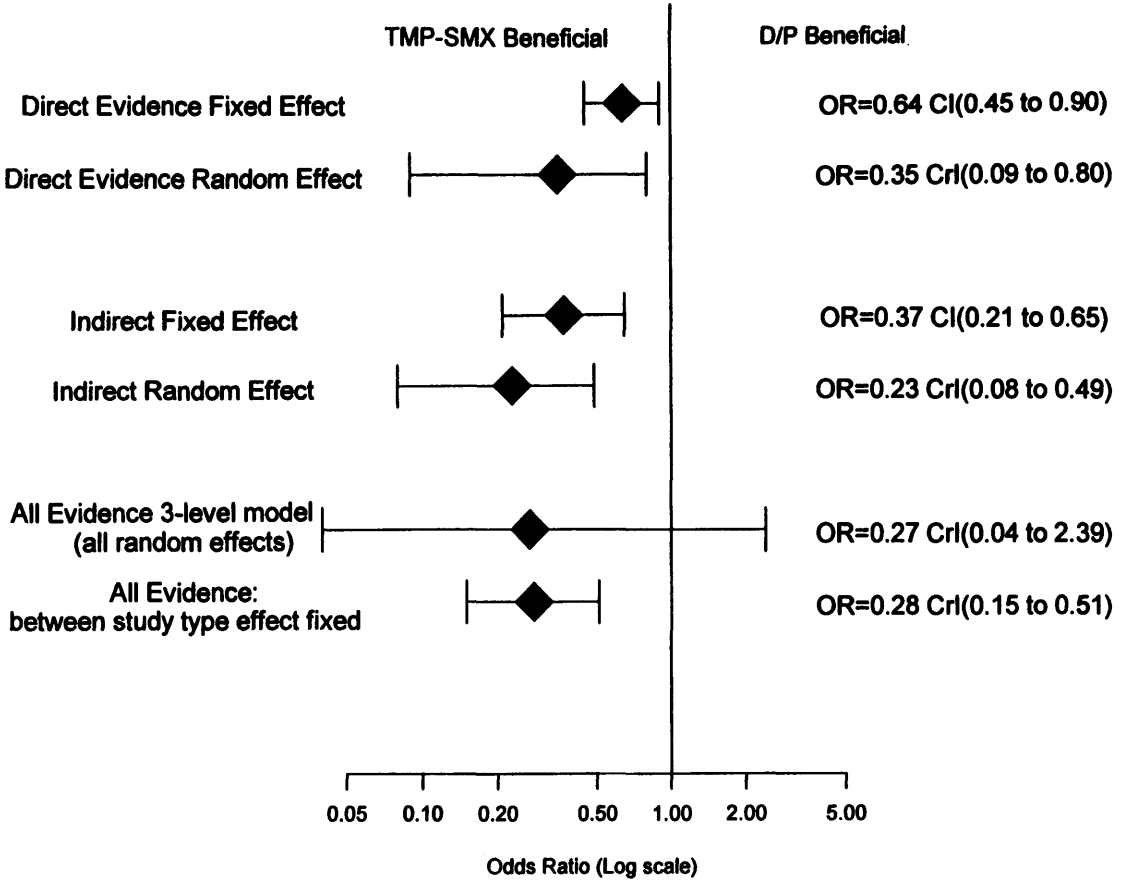
reported previously, and is the result of combining all the evidence assuming no variation between study types, i.e. assuming both direct and indirect studies are estimating exactly the same treatment effect. Perhaps the most striking aspect of these results is the difference in the width of the credibility intervals between estimates five and six. Much has been said regarding the appropriateness and relative benefits of fixed and random effect models for meta-analysis generally, however it has been observed that often they will produce very similar results. (Sutton et al. 1998) As this example highlights, despite the evidence being in qualitatively broad agreement, (i.e. a considerable benefit is observed for TMP-SMX over the alternatives) if more complex synthesis models are constructed results may differ radically depending whether fixed or random effects are assumed between model parameters.

A natural extension of the work presented here would be developing a way of placing a constraint on the modelling restricting the influence the indirect comparisons, under the assumption that the direct evidence is less biased than the indirect estimate. Although such a constraint has been discussed previously for a three level hierarchical model, (Prevost et al. 2000) due to the added complexity of including the indirect evidence model such an extension is non-trivial, although the author suspects it would be eminently possible.

SPECIAL NOTE

THIS ITEM IS BOUND IN SUCH A MANNER AND WHILE EVERY EFFORT HAS BEEN MADE TO REPRODUCE THE CENTRES, FORCE WOULD RESULT IN DAMAGE

Figure 5.13 Comparison of estimates of treatment effect for HIV treatment derived using direct and indirect data and different model specifications



5.5 Summary/discussion of sections 5.3 and 5.4

Sections 5.3.and 5.4 have outlined different models for combining studies where data is available on different outcomes or different comparisons have been made in the studies of interest leading to the possibility of estimating indirect comparison effects. Although neither of these models are directly applicable for combining the cholesterol data, considered in the first half of this chapter, they are very relevant to a generalised synthesis framework and illustrate alternative methods for synthesising data. Deriving Bayesian formulations for these models means that prior information could be included in such syntheses. A further advantage of the MCMC formulations discussed is that these models can be incorporated as components of more complex models as illustrated in section 5.4.3 without the need for specialist software. In section 5.4 the issue of the increased importance of random effect specification was highlighted in generalised synthesis models over standard meta-analysis models. Further work is required both from theoretical and empirical standpoints to establish robust specification and estimation procedures in such contexts. Due to these problems, sensitivity analysis is particularly important in generalised synthesis contexts, and it is sensible to consider estimates from fixed effect models, simpler meta-analysis models and using different prior distributions for the variance components as part of this.

Chapter 6 Generalised synthesis of evidence and the threat of dissemination bias: electronic fetal heart rate monitoring (EFM)

6.1 Introduction

This chapter considers the problem of publication bias, in an evidence base that includes studies of more than one design. There is a general belief that publication bias exists because research with statistically significant, or interesting, results is potentially more likely to be submitted and published than work with null or non-significant, or uninteresting results. Many specific biases sometimes collected under the broad umbrella term publication bias exist, including pipeline bias, language bias, selective outcome reporting, duplicate publication, grey literature bias, citation bias, database bias, retrieval bias and media attention bias. Song et al. describe these biases in detail.(Song et al. 2000b) It has been suggested that the whole collection of publication and related biases be described by the term dissemination bias.(Song et al. 2000b) This follows from the observation that publication is not a dichotomous event: rather it is a continuum.(Smith, 1999) The dissemination profile of research can range from the completely inaccessible to the easily accessible, according to whether, when, where and how research is published.(Song et al. 2000b) Hence, for the meta-analyst the concern that a complete, or at least an unbiased, body of evidence has been identified goes beyond concerns relating to the identification of all known studies, but concerns specific outcomes, groups of patients and analyses.

An array of methods to help deal with the problem of publication bias has been developed; some of these are described in section 2.6 and a fuller treatment is given elsewhere. (Sutton et al. 2000b; Macaskill et al. 2001) These methods are all broadly based on the symmetry of a funnel plot, and although difficult to validate, provide a framework for carrying out a sensitivity analysis regarding the likely impact of publication bias. It should not be forgotten, however, that factors other than publication bias, that are related to both study size and study outcome can cause funnel plot

asymmetry. An example is variable study quality. If generally the larger studies are of better quality, and the better quality studies produce smaller effect sizes than poorer quality studies, due to less bias being present in them, then this will give the appearance of asymmetry in the funnel. (Sterne et al. 2000b) Assessments of the likely impact of publication bias on meta-analyses of RCTs (Sutton et al. 2000a) and studies of diagnostic studies (Song et al. 2002) have been carried out using the Beggs' and Egger's tests for publication bias (Section 2.6.1) and the method of Trim and Fill (2.6.2). However, little empirical work has been carried out in other types of observational study, although publication bias in observational studies on particular topics have been well documented.(Copas and Shi, 2000; Givens et al. 1997)

It may be a reasonable conjecture that the degree to which dissemination bias distorts an evidence base depends to some degree on the type of study designs being considered. For example, concern has been raised that publication bias may be an even greater problem among observational studies results than it is among RCTs.(Givens et al. 1997) Although this has not been investigated empirically, there are sound arguments in its favour. Begg has pointed out (Givens et al. 1997) that when case-control studies are conducted, detailed information will usually be collected on a broad range of potential risk factors. Once completed, investigators will typically publish results in a series of articles, each dealing with a different risk factor or group of factors. There is a danger that only the most interesting associations will be published, some of which may have been data-driven and not the result of examining pre-specified hypotheses. The potential to report only partially on multiple hypotheses examined is also very high in other classic observational designs including cohort and cross-sectional studies. Further, as medical databases and record linkage improve, analysis of routine data becomes more frequent. Such analyses may not even be considered for publication *unless* 'interesting' results are found.

The focus of this chapter is the consideration of the impact of dissemination bias on a generalised synthesis of evidence using the evidence relating the use of electronic fetal heart rate monitoring (EFM) and its effect on preventing perinatal mortality. Data from three study types are considered. Section 6.2 describes the EFM studies, and briefly summarises previous assessments of this body of evidence. It will be seen that events

are rare in all three study designs. Specific issues related to generalised synthesis of rare events are considered in Chapter 7, and many of those issues are pertinent here. In Section 6.3, an assessment of the presence of publication bias is carried out on the EFM data using the tests available. Section 6.4 reports on the use of the method of 'Trim and Fill' to adjust the study results for publication bias as part of a sensitivity analysis. Section 6.5 reports on an updated generalised synthesis of this evidence taking into account the result of the publication bias sensitivity assessment. Section 6.6 reports on further sensitivity analyses of the EFM literature regarding publication bias, and describes limitations of the methods used and some possible modifications to avoid them. Although this may appear to be an ambitiously wide range of issues to consider, there are clearly interactions between them which will become apparent. Section 6.7, the discussion, concludes the chapter.

6.2 The evidence relating to Electronic fetal heart rate monitoring (EFM) for reducing perinatal mortality

A recent meta-analysis (Hornbuckle et al. 2000) identified the data re-analysed here. Over the years, there has been much debate over the effectiveness of EFM in reducing perinatal mortality. While doctors argue such equipment should assist with timely delivery, hence reducing the chance of a stillbirth, EFM has not been shown to reduce perinatal mortality in the nine RCTs that have been identified (Table 6.1). However, since perinatal mortality is rare (only 85 deaths were reported in the 18695 births randomised in the nine trials) a lack of statistical power is a possible explanation for the lack of clear results from a meta-analysis of such studies.

In addition to these RCTs, 17 comparative observational studies have also been reported, comprising 7 comparative cohort studies, and 10 before-and-after studies. In the cohort studies, women who received EFM were compared to others who were not. The before-and-after studies report on practice in two time periods where the use of EFM increased in the second period (Table 6.2). These studies include data on many more subjects than the RCTs (across all 17 observational studies, 377 deaths were

reported in the 284,878 births observed). Generally, as will be seen, these studies are also much more supportive of the effectiveness of EFM on perinatal mortality than the trials.

The results of using standard fixed and random effect inverse variance models (Sutton et al. 2000a) to meta-analyse the data separately for the three study types, and for all the evidence combined, are presented in Table 6.3. While the point estimates from all models suggest EFM is beneficial, the effect size gains statistical significance and is much larger in the observational studies compared to the RCTs. Note that the fixed and random effect pooled point estimates are fairly similar for the RCTs and the comparative cohort studies, while they disagree considerably (-0.72 compared to –1.86) for the before-and-after studies.

SPECIAL NOTE

THIS ITEM IS BOUND IN SUCH A MANNER AND WHILE EVERY EFFORT HAS BEEN MADE TO REPRODUCE THE CENTRES, FORCE WOULD RESULT IN DAMAGE

Table 6.1 Data from 9 RCTs examining the effect of EFM on perinatal mortality

Study ID	Year of publication	Number of subjects given EFM	Number of subjects not given EFM	Number of perinatal deaths in treatment arm	Number of perinatal deaths in control arm	Risk Difference per 1000 births	Standard Error of Risk Difference	95% CI for Risk Difference
R1	1976	175	175	1	1	0.00	8.06	(-15.79 to 15.79)
R2	1976	242	241	2	1	4.12	7.14	(-9.88 to 18.11)
R3	1978	253	251	0	1	-3.98	5.59	(-14.93 to 6.97)
R4	1979	463	232	3	0	6.48	5.03	(-3.38 to 16.34)
R5	1981	445	482	1	0	2.25	3.11	(-3.84 to 8.34)
R6	1985	485	493	0	1	-2.03	2.87	(-7.66 to 3.60)
R7	1985	6530	6554	14	14	0.001	0.81	(-1.58 to 1.59)
R8	1987	122	124	17	18	-5.82	44.54	(-93.11 to 81.48)
R9	1993	746	682	2	9	-10.52	4.76	(-19.85 to -1.18)

Table 6.2 Data from 7 comparative cohort studies and 10 before-and-after studies examining the effect of EFM on perinatal mortality

Study ID	Year of publication	Number of subjects given EFM	Number of subjects not given EFM	Number of perinatal deaths in EFM group	Number of perinatal deaths in non EFM group	Risk Difference per 1000 births	Standard Errors of Risk Difference	95% CI for Risk Difference
Comparative Cohort Studies								
C1	1973	5427	1162	17	2	-1.41	1.43	(-4.22 to 1.40)
C2	1973	6836	150	15	0	-2.19	4.71	(-11.43 to 7.04)
C3	1975	6179	608	37	1	-4.34	1.91	(-8.10 to -0.59)
C4	1977	2923	4210	9	1	-2.84	1.05	(-4.90 to -0.78)
C5	1978	692	554	3	1	-2.53	3.08	(-8.57 to 3.51)
C6	1979	8634	4978	2	0	-0.23	0.23	(-0.69 to 0.22)
C7	1982	66208	45880	45	10	-0.46	0.12	(-0.70 to -0.22)
Before-and-After Studies								
B1	1975	1024	991	0	4	-4.04	2.24	(-8.43 to 0.35)
B2	1975	1080	1161	9	7	2.30	3.58	(-4.71 to 9.32)
B3	1975	1950	11599	1	14	-0.69	0.61	(-1.88 to 0.49)
B4	1976	3529	4323	1	15	-3.19	0.94	(-5.03 to -1.35)
B5	1977	3852	4114	21	53	-7.43	2.12	(-11.59 to -3.27)
B6	1978	7312	15357	6	35	-1.46	0.51	(-2.46 to -0.46)
B7	1980	4503	4240	2	19	-4.04	1.07	(-6.14 to -1.93)
B8	1980	8174	6740	5	15	-1.61	0.64	(-2.86 to -0.37)
B9	1984	7911	7582	2	13	-1.46	0.51	(-2.46 to -0.47)
B10	1986	17586	17409	5	7	-0.12	0.20	(-0.51 to 0.27)

Table 6.3 Results of meta-analysing the EFM evidence individually by study design and combining all the evidence on the risk difference scale

Studies included	Meta-analysis model	Risk difference per 1000 births (95% Confidence interval)	Between study variance estimate	Test for heterogeneity *p*-value
RCTs	Fixed Random	-0.14 (-1.56 to 1.28) -0.22 (-2.03 to 1.58)	NA 0.57	0.40
Comparative cohort studies	Fixed Random	-0.46 (-0.67 to –0.25) -0.68 (-1.25 to –0.11)	NA 0.16	0.08
Before-and-After studies	Fixed Random	-0.72 (-1.03 to –0.42) -1.86 (-2.84 to –0.89)	NA 1.50	<0.001
All three study designs combined	Fixed Random	-0.54 (-0.71 to –0.37) -1.24 (-1.74 to –0.74)	NA 0.48	<0.001

More sophisticated methods have been used to synthesise the evidence from these EFM studies. A recent review (Hornbuckle et al. 2000) considered a Bayesian analysis of this evidence, using the observational evidence to derive a prior for the effectiveness parameter in a meta-analysis of the RCTs (see Section 7.8.2). A direct synthesis of all the data using the Bayesian hierarchical model used to combine the cholesterol data in Chapter 5, has been described by Sutton and Abrams (Sutton and Abrams, 2001) elsewhere. In this previous analysis (using model (3.3)) a pooled risk difference estimate per 1000 births of -1.35 (95% Credible Interval –3.22 to 0.33) was produced. Although the potential for publication bias within this evidence has been acknowledged (Hornbuckle et al. 2000) no formal assessment has been carried out before. A re-evaluation using analytical methods to deal with publication bias is reported in the next section.

6.3. An assessment of dissemination bias in the EFM literature

There is broad consensus that an assessment to potential publication bias should be made as part of a sensitivity analysis when carrying out a meta-analysis. In the assessment that follows, funnel plots are examined, Egger's test for bias is applied, (Egger et al. 1997) and the method of Trim and Fill (Duval and Tweedie, 2000a; Duval and Tweedie, 1998c) used to investigate the potential impact of missing studies on the analysis. These methods were described in Section 2.6.

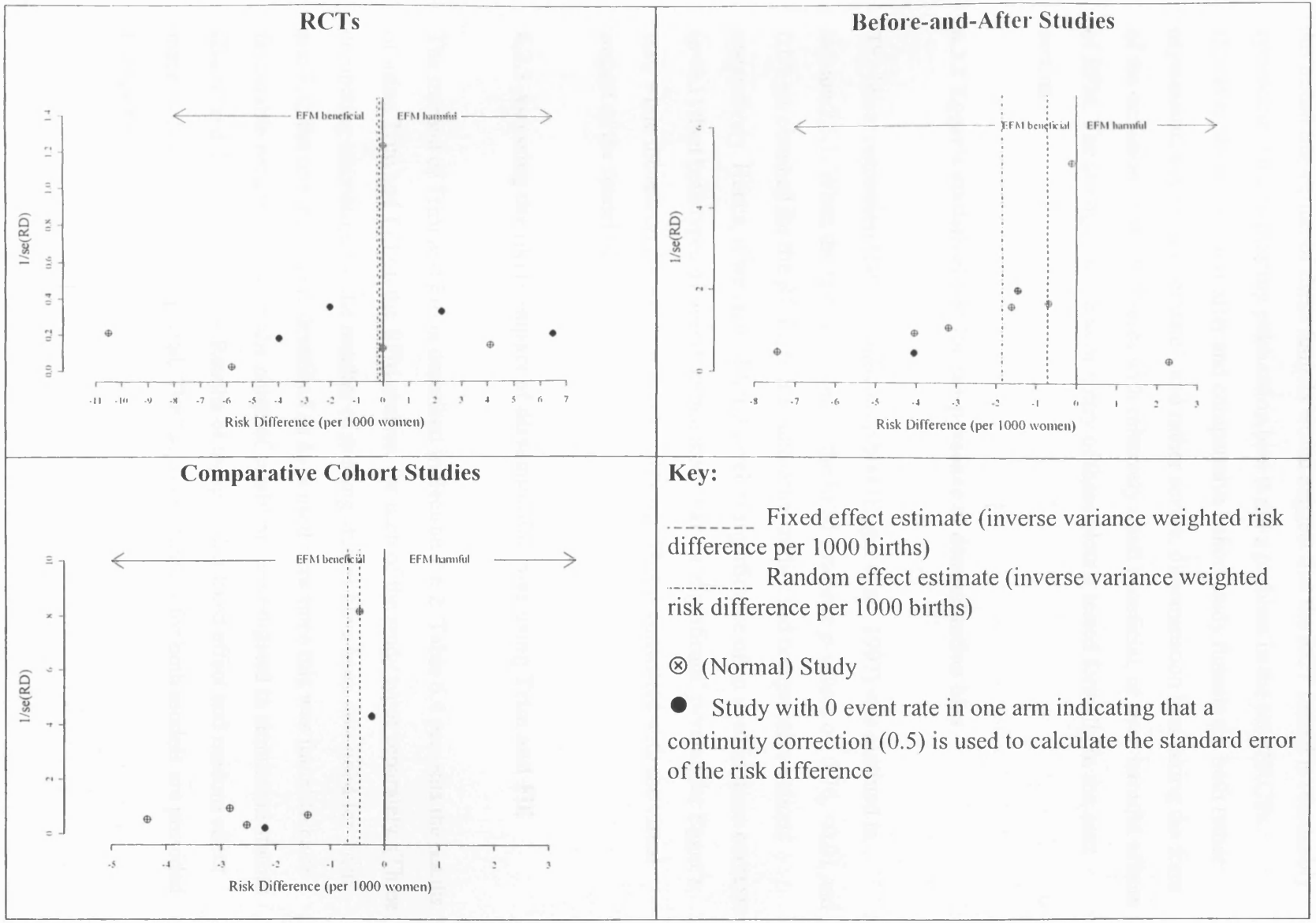
6.3.1 Funnel plots

Funnel plots for each of the three types of studies are presented in Figure 6.1. Included on these plots are the fixed and random effect pooled estimates. The studies requiring a continuity correction factor due to zero event rates (see Section 6.5) are displayed using a different plotting symbol as noted in the key to Figure 6.1.

SPECIAL NOTE

THIS ITEM IS BOUND IN SUCH A MANNER AND WHILE EVERY EFFORT HAS BEEN MADE TO REPRODUCE THE CENTRES, FORCE WOULD RESULT IN DAMAGE

Figure 6.1 Funnel plots for three study types used to evaluate the effect EFM on perinatal mortality on the risk difference scale



An initial inspection of these funnels would suggest that the RCT funnel is reasonably symmetric, thus suggesting publication bias is not a problem in the set of RCTs. However, the before-and-after and comparative cohort study funnels are both rather asymmetric, suggesting 'classic', and rather severe, dissemination bias taking the form of the omission of small studies with relatively small beneficial, or even harmful effects of EFM. The strength of the asymmetry of these plots it tested formally in the next section.

6.3.2 Egger's statistical test for the presence of dissemination bias

The linear regression test for publication bias (Egger et al. 1997) was outlined in section 2.6.1. When the test is applied to the EFM evidence p-values of 0.76, <0.01 and 0.07 are obtained for the RCTs, before-and-after studies and comparative cohorts respectively. Hence, if we take a liberal level of significance often used in these contexts (p<0.1) then both types of observational study have a 'significant' p-value for Egger's test, while the test on the RCTs is non-significant. This is consistent with the visual impact of the funnel plots.

6.3.3 Assessing the likely impact of dissemination bias using Trim and Fill

The method of Trim and Fill is described in Section 2.6.2. Table 6.4 presents the results of using Trim and Fill on the EFM studies, for each of the study types separately. (Three competing estimators for the number of missing studies have been described for Trim and Fill; the one previously described as L_0 is used here since this was found to have favourable performance over the range of conditions investigated in simulation studies (Duval and Tweedie, 2000a)). Results of using both a fixed effect and random effect meta-analysis model are reported. The 'adjusted' funnels for both models are provided in Figure 6.2.

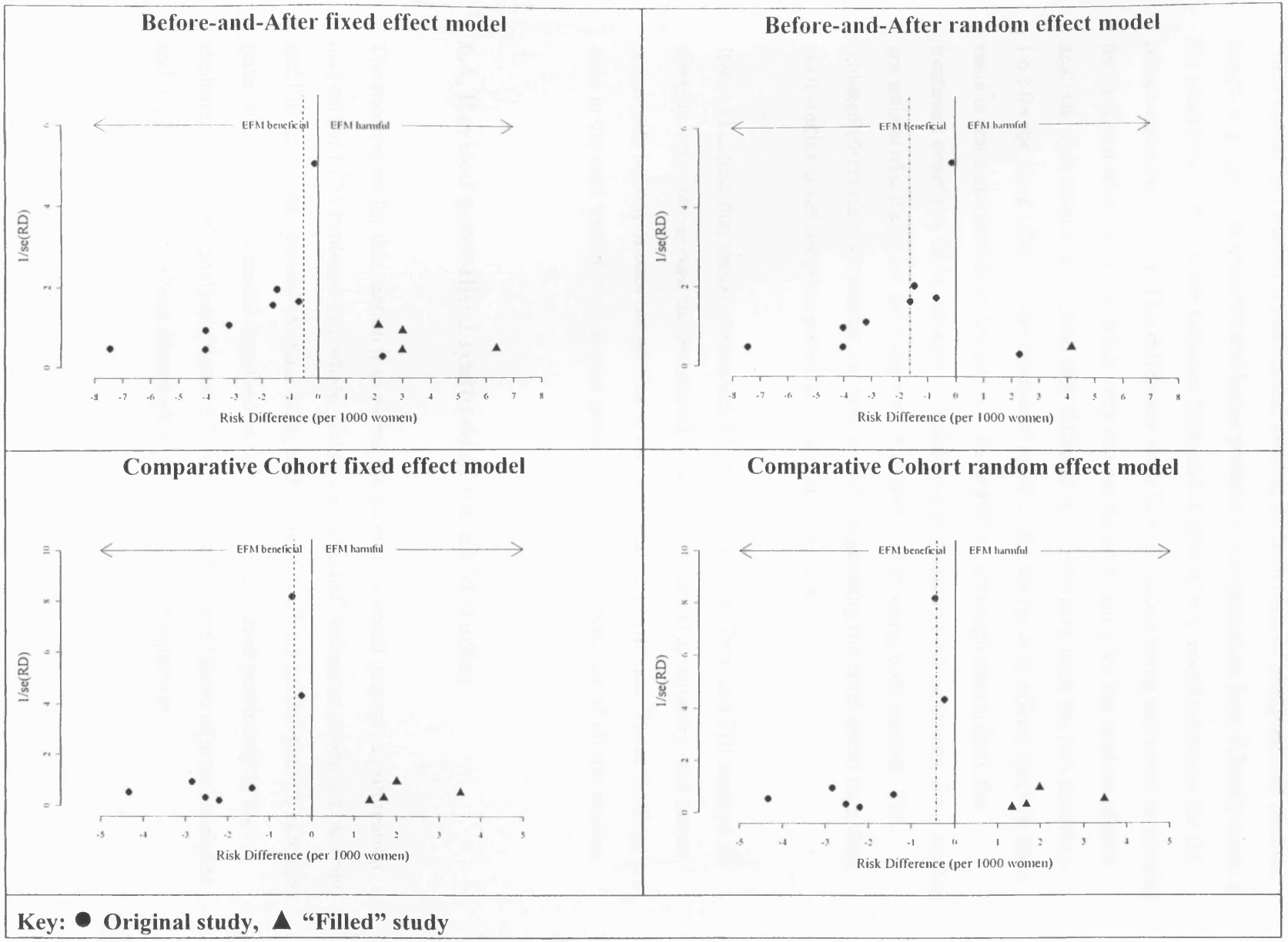
Table 6.4 Results of Trim and Fill analysis for the EFM studies on the risk difference scale

	Number of studies	Meta-analysis model	Original estimate - Risk difference per 1000 births (95% CI)	Number of studies estimated missing	Adjusted estimate using Trim and Fill – Risk difference per 1000 births (95%CI)
RCTs	9	Fixed	-0.14 (-1.56 to 1.28)	0	-
		Random	-0.22 (-2.02 to 1.59)	0	-
Before-and-After	10	Fixed	-0.72 (-1.03 to -0.42)	4	-0.52 (-0.82 to -0.22)
		Random	-1.86 (-2.84 to –0.89)	1	-1.63 (-2.63 to –0.64)
Comparative Cohort	7	Fixed	-0.46 (-0.67 to –0.25)	4	-0.42 (-0.63 to -0.21)
		Random	-0.68 (-1.25 to –0.11)	4	-0.43 (-1.12 to 0.27)

SPECIAL NOTE

THIS ITEM IS BOUND IN SUCH A MANNER AND WHILE EVERY EFFORT HAS BEEN MADE TO REPRODUCE THE CENTRES, FORCE WOULD RESULT IN DAMAGE

Figure 6.2 "Filled" funnels from the Trim and Fill analysis of the observational EFM studies on the risk difference scale



No studies were estimated as missing for the RCTs using either the fixed or the random effect model, but studies are estimated missing for both sorts of observational studies, suggesting funnel asymmetry and hence potentially dissemination bias. Already noted is the considerable difference between fixed and random effect model estimates for the before-and-after studies. This difference leads to four studies being estimated as missing for the fixed effect analysis, while only one estimated missing for the random effects analysis. This results in considerably different point estimates from the two models (-1.63 for the fixed effect model compared to –0.52 for the random effects model); this issue is discussed further in Section 6.4. Nonetheless, although diminished, the treatment benefit of EFM remains statistically significant in both instances. Four studies are estimated missing for the comparative cohort studies using both models. The adjusted effect sizes are similar for both models, suggesting the conclusion regarding such studies is not robust to potential dissemination biases.

It would appear that visual assessment, Egger's test and the Trim and Fill method all come to very similar conclusions, namely that there is funnel asymmetry, and hence potentially serious dissemination bias in the observational evidence. These findings are used in the next section to revise the generalised synthesis analysis of all the studies.

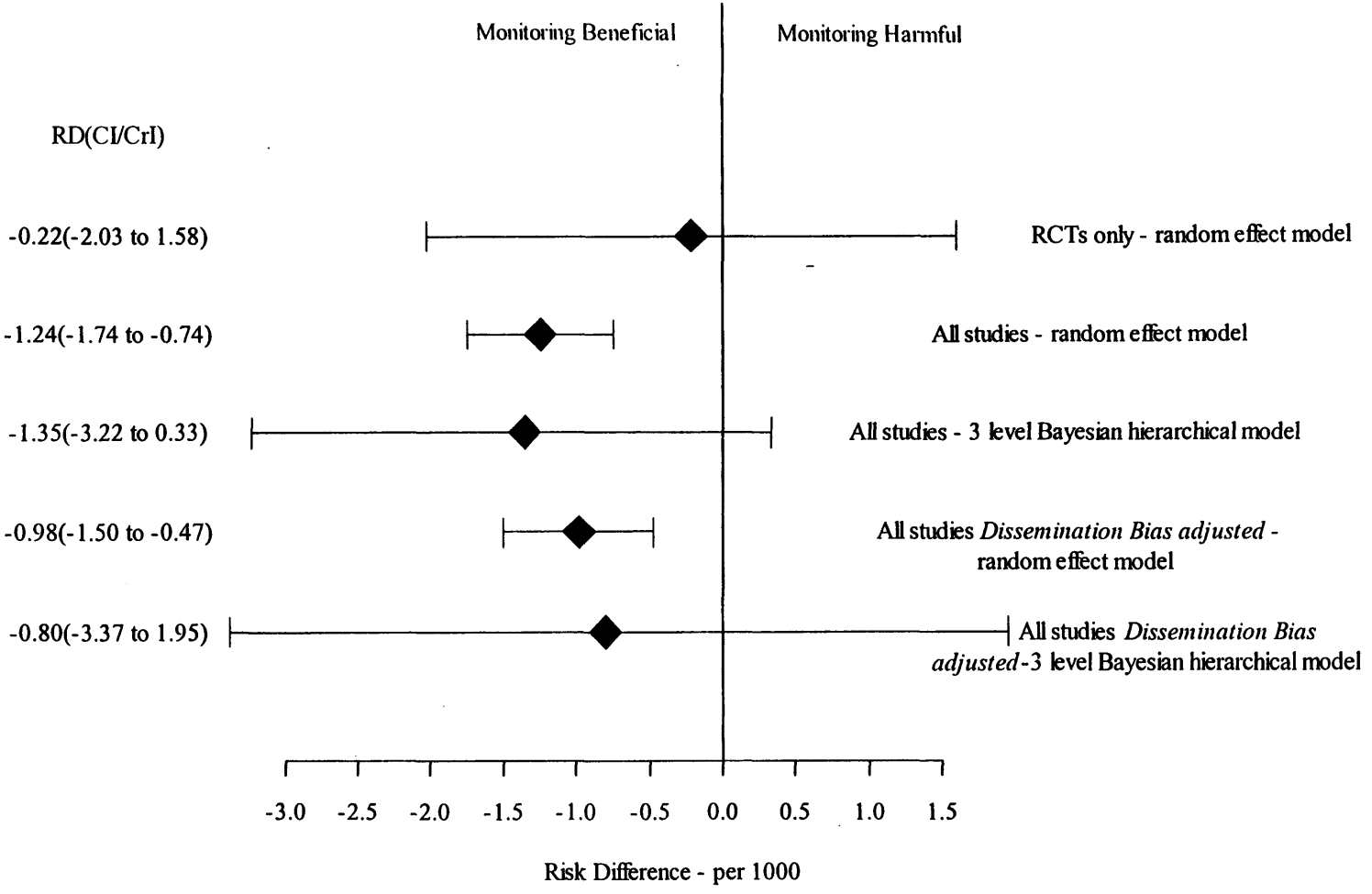
6.4. Revised generalised synthesis of the EFM studies

The motivation for this section is to assess the potential overall impact of publication bias on the EFM evidence as a whole. Using the 'adjusted' estimates produced by Trim and Fill, an overall pooled estimate using both a random effects model and the extended generalised synthesis model described in section 3.8.1 and used previously in the cholesterol lowering analysis. Figure 6.3 displays the results of these adjusted analyses and of several other analyses described in Section 6.2 for comparison.

SPECIAL NOTE

THIS ITEM IS BOUND IN SUCH A MANNER AND WHILE EVERY EFFORT HAS BEEN MADE TO REPRODUCE THE CENTRES, FORCE WOULD RESULT IN DAMAGE

Figure 6.3 Pooled results from using different models to combine the EFM data on the risk difference scale



The first pooled estimate on Figure 6.3 is that obtained combining only the RCTs using a random effect model (see Table 6.3). While the point estimate suggests a slight benefit of EFM, the confidence interval is very wide due to the limited number of randomised patients available.

The second estimate is derived from a random effect model directly combining all the studies (see Table 6.3). In contrast to the RCT result the benefit of using EFM is much larger, and the confidence interval much narrower. This is due to the observational evidence generally being more supportive of EFM than the randomised evidence, and since many of the observational studies are much larger than the RCTs, the RCTs have relatively little weight in this analysis.

The third estimate is derived from the three level hierarchical model. The treatment difference is similar to that obtained using a standard random effect model, but the confidence interval is wider. Here, between-study-type heterogeneity is permitted. This increases the width of the confidence interval, as there clearly is between-study-type heterogeneity with the observational studies generally having more favourable results than the RCTs. Additionally, such an analysis gives each study type a more equal weighting just as individual studies are given a more equal weighting in a random than a fixed effect model when between-study heterogeneity is present.

The fourth estimate is a new result derived from a random effect meta-analysis of all the evidence having 'adjusted' for dissemination bias using Trim and Fill individually in each study type first. As illustrated in Section 6.3, adjusting for funnel asymmetry has reduced the treatment effect in the observational studies, which is reflected in the shift of the estimate towards no treatment effect compared with estimate 2.

Finally, estimate 5 uses the three level model after augmenting the dataset using Trim and Fill in the same manner as estimate four. This model thus accounts for between study type heterogeneity and potential dissemination bias. Using this final model, the most likely treatment benefit is a reduction of around 0.8 deaths per 1000 births, but there is a large degree of uncertainty in the estimation and the data is compatible with

both much larger, and much smaller (even harmful) effects (95% CrI –3.37 to 1.95). Hence, this point estimate suggests a larger effect than the RCTs alone, but smaller than the benefit suggested by the observational studies at face value. The wide credible interval is a consequence of some discrepancy between the estimates from the three study types even after adjustment for funnel asymmetry, which is formally incorporated in the model as between study type heterogeneity. Such residual discrepancies in the effects suggested by the different study types may be the result of other types of biases affecting the studies' results.

6.5 Further dissemination-bias-related sensitivity analyses of the EFM data

In addition to the sensitivity analysis assessing the potential impact of publication bias on the EFM evidence base reported above, further factors related to the assessment of publication bias in the EFM evidence base are considered below.

6.5.1 Change of outcome scale

The choice of outcome scale on which to combine binary outcome data on in a meta-analysis is not straightforward, (Deeks and Altman, 2001) with three mainstream competing choices being the odds ratio, the relative risk and the risk difference. Although similar conclusions will often be drawn whichever of the outcome measures is used, occasionally important discrepancies are found (Deeks and Altman, 2001). Importantly, changing the outcome scale can also change the appearance of the funnel plot and any inferences or adjustments based on it. (Tang and Liu, 2000). Tang and Liu suggest that a ratio measure may be the most desirable measure because the control group event rate is more likely to be associated with treatment effect expressed as risk difference, and size of studies is often associated with event rates, which may cause asymmetry in funnel plots of the risk difference. (Tang and Liu, 2000) Earlier meta-analyses of EFM (Hornbuckle et al. 2000) studies used the risk difference scale, probably because such a measure is more clinically interpretable when considering such

rare events. There has been considerable debate recently in the choice of scale for binary outcomes, generally (Walter, 2000), and specifically in the meta-analysis literature.(Deeks and Altman, 2001; Engles, et al. 2001; Tang, 2000) No one scale can be considered superior, and it has been suggested that the one which fits the data best (in a meta-analysis context this could be considered as the one on which there is least heterogeneity) should be used. However problems with data driven, post hoc assessments are acknowledged and a sensitivity analysis re-analysing on multiple scale may be the most sensible approach.(Deeks and Altman, 2001; Walter, 2000)

Table 6.5 P-values for the heterogeneity test for the EFM evidence using the risk difference and odds ratio outcome scales

Study Type	P-value for the heterogeneity test	
	Risk Difference	Odds Ratio
RCT	0.40	0.61
Comparative Cohort	0.08	0.74
Before-and-After	<0.001	0.07

Table 6.5 displays the p-value results for the standard heterogeneity test in meta-analysis (equation (2.14)). It can be seen the significance of the test is lower on the odds ratio scale compared with the risk difference for all three types of evidence, and only retains statistical significance at the 10% level for the before-and-after studies. In light of this, the assessment in Section 6.4 was replicated using the odds ratio scale. The corresponding funnel plots are presented in Figure 6.4.

Figure 6.5 explores the relationship between odds ratio and risk difference further. In the left hand panels of the figure a scatter plot of risk difference versus odds ratio is shown. This highlights the fact that the relationship between the two measures is not linear, especially for the comparative cohort studies. The right hand panel of Figure 6.5 displays the same scatter plot as the left hand side, but in addition, rectangles indicating

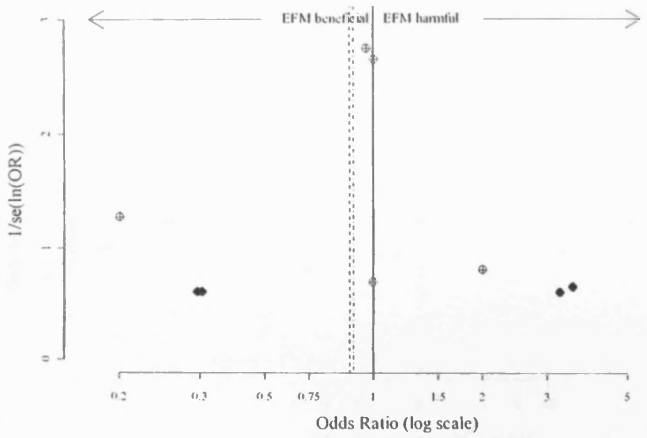
the confidence intervals for studies on both scales are included. Although these plots are rather 'busy', they do give an indication of how the standard errors for both measures also vary, by considering the shape of each rectangle. However, no pattern in the relationship between each scale, and the studies standard errors on each is discernable.

SPECIAL NOTE

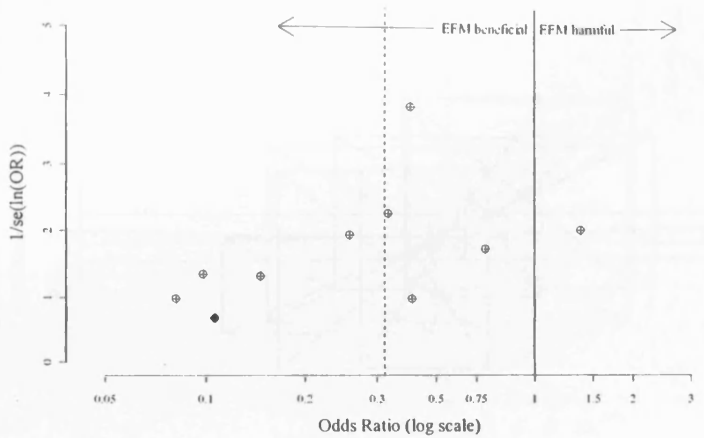
THIS ITEM IS BOUND IN SUCH A MANNER AND WHILE EVERY EFFORT HAS BEEN MADE TO REPRODUCE THE CENTRES, FORCE WOULD RESULT IN DAMAGE

Figure 6.4 Funnel plots for three study types used to evaluate the effect EFM on perinatal mortality on the odds ratio scale

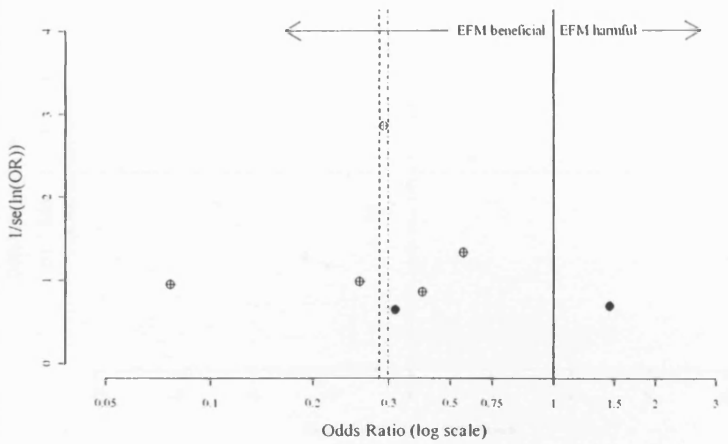
RCTs



Before-and-After Studies



Comparative Cohort Studies



Key:

_____ Fixed effect estimate (inverse variance weighted risk difference per 1000 births)

_____ Random effect estimate (inverse variance weighted risk difference per 1000 births)

⊗ (Normal) Study

● Study with 0 event rate in one arm indicating that a continuity correction (0.5) is used to calculate the standard error of the risk difference

Figure 6.5 Scatter plots of odds ratios versus risk differences for the EFM studies (the right-hand side column including rectangles representing the 95% confidence intervals on both scales

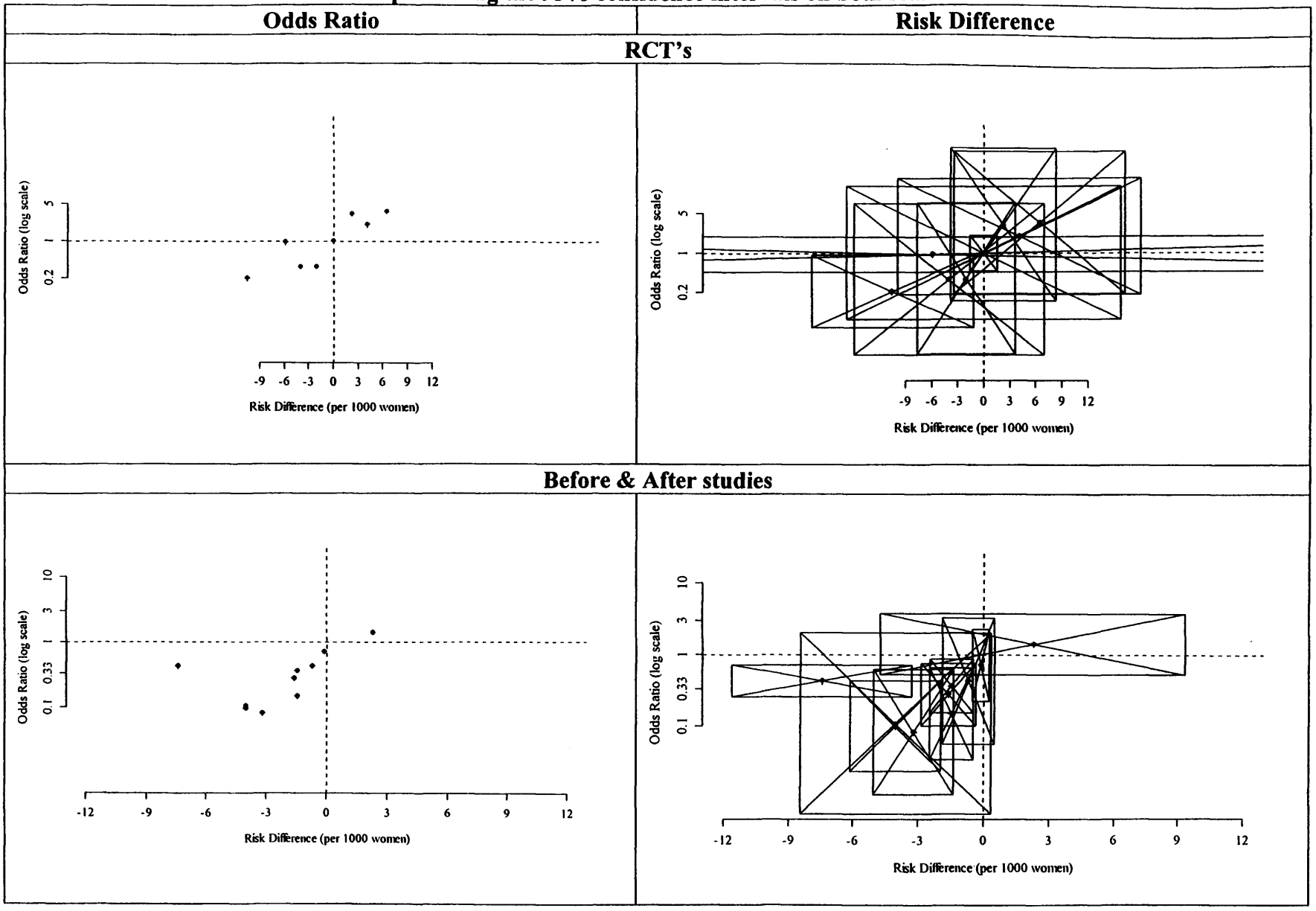
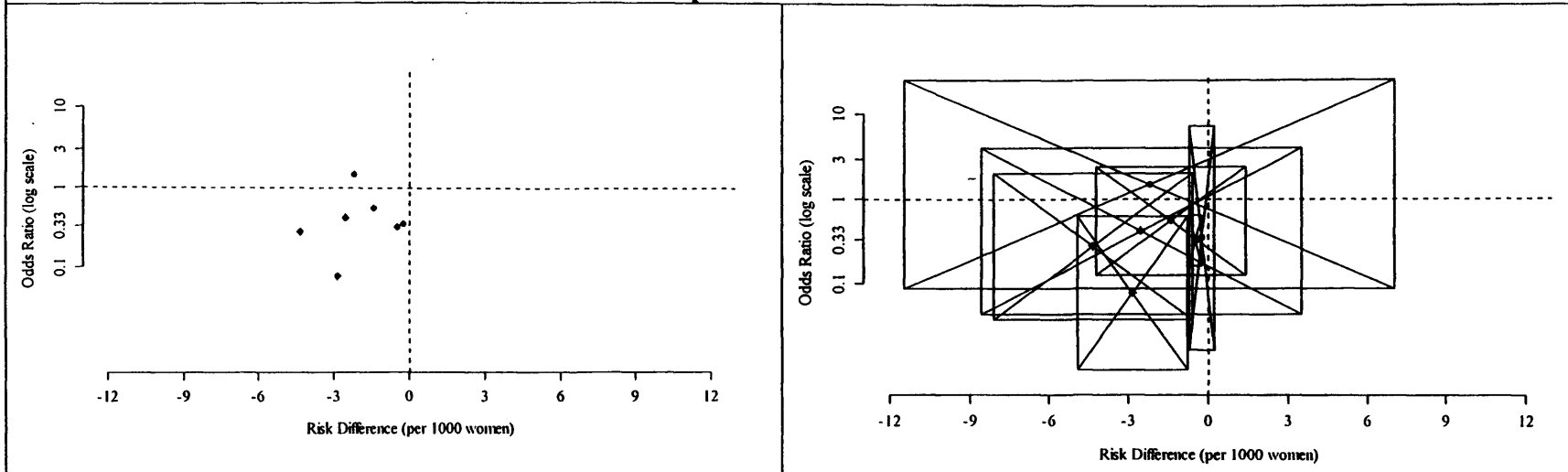


Figure 6.5 Continued

Comparative Cohort studies



Key:

⊗ (Normal) Study

● Study with at least one 0 cell meaning a continuity correction is required (0.5 used) to calculate the odds ratio & its standard error, and the standard error of the risk difference

The odds ratio funnel plot for the RCTs (Figure 6.4) is similar to that for the risk difference in Figure 6.1. However those for both the before-and-after and the comparative cohort studies have changed somewhat. Asymmetry is less extreme in the odds ratio plot, compared to the plot on the risk difference scale, for the before-and-after studies, while no asymmetry is now discernible in the funnel plot for the odds ratio for comparative cohort studies. Applying Egger's test to the three sources of evidence on an odds ratio scale produces a p-value of 0.96 for the RCTs, 0.19 for the before-and-after studies, and 0.21 for the comparative cohort studies. A Trim and Fill analysis on the odds ratio scale (Table 6.6) estimates zero studies are missing for the RCTs and the comparative cohort studies (fixed and random effect models). Three studies are estimated missing for the before-and-after studies leading to a slightly diminished effect of EFM, but one which is still statistically significant using both fixed and random effect models.

Table 6.6 Results of Trim and Fill analysis for the EFM studies on the (log) odds ratio scale

	Number of studies	Meta-analysis model	Original estimate - Odds ratio (95% CI)	Number of studies estimated missing	Adjusted estimate using Trim and Fill – Risk difference per 1000 births (95%CI)
RCTs	9	Fixed	0.88 (0.57 to 1.36)	0	-
		Random	0.88 (0.57 to 1.36)	0	-
Before-and-After	10	Fixed	0.38 (0.28 to 0.52)	3	0.43 (0.33 to 0.60)
		Random	0.33 (0.20 to 0.55)	3	0.44 (0.26 to 0.75)
Comparative Cohort	7	Fixed	0.32 (0.20 to 0.53)	0	-
		Random	0.32 (0.20 to 0.53)	0	-

Hence, all three methods of assessment consistently indicate that funnel plot asymmetry, and hence the potential impact of dissemination bias, is lower when the study results are analysed using the odds ratio measure. Although the results on the risk difference scale

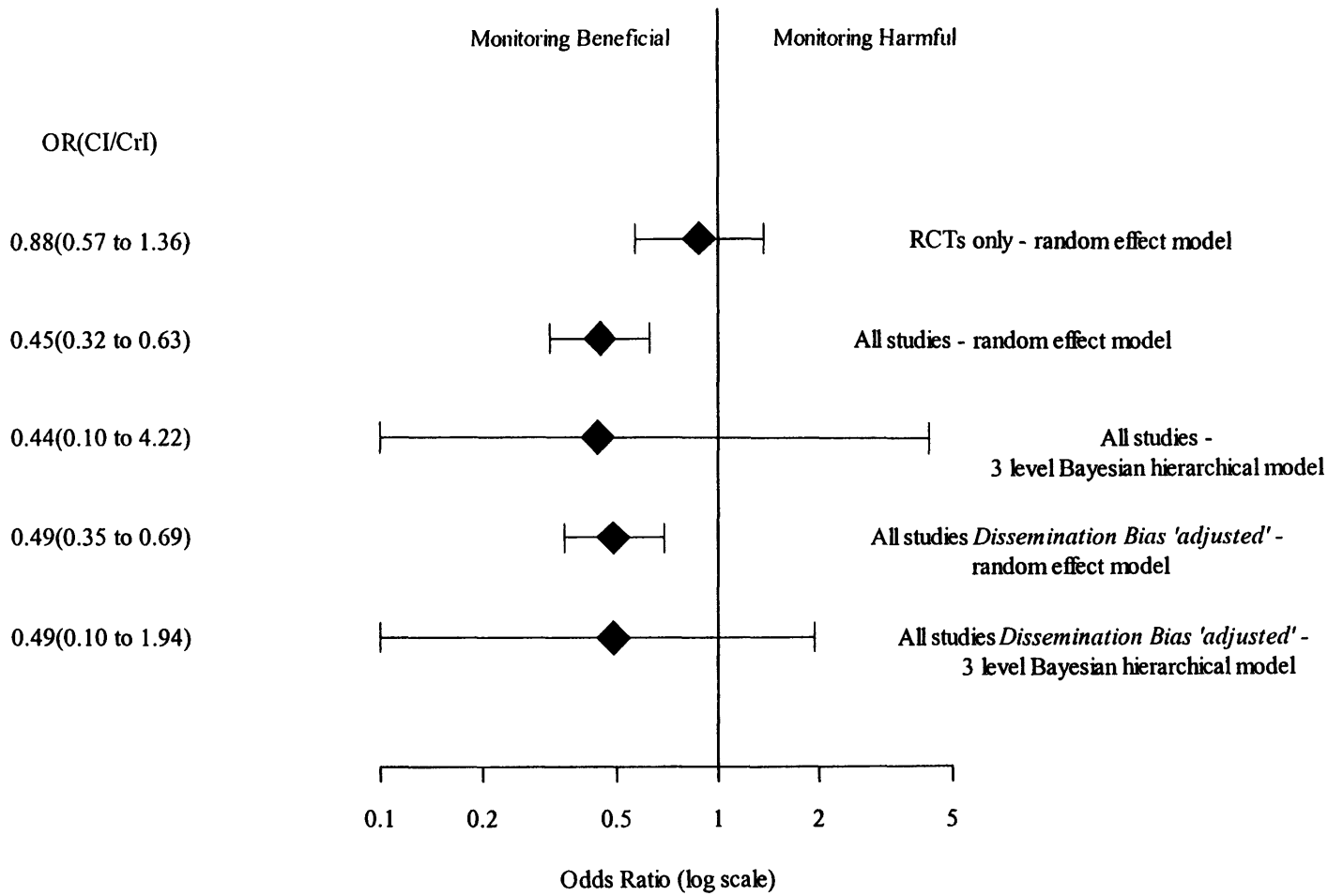
on their own in isolation are persuasive, this lack of robustness in findings to outcome measure choice makes interpretation of the assessment of publication bias problematic, and the possibility that the asymmetry observed on the risk difference scale is due solely, or largely, to the choice of outcome measure cannot be ruled out. Together, with previous findings (Tang and Liu, 2000) this highlights the need for a publication bias assessment measure which is robust to scale choice, if possible.

Figure 6.3 displayed the results for different methods of combining the EFM evidence on the risk difference scale; the equivalent plot for the odds ratio scale is provided in Figure 6.6. Although direct comparison is not possible between results on different metrics, the plots have a qualitatively similar appearance, with wider confidence intervals being produced for the three level model compared to the standard random effects model. One point to note is that for the odds ratio scale, although the pooled estimate changes very little after the adjustment for dissemination bias, the credible interval for the three level model is narrower than that that from the unadjusted data, suggesting that between study type heterogeneity has been reduced by the dissemination bias adjustment.

SPECIAL NOTE

THIS ITEM IS BOUND IN SUCH A MANNER AND WHILE EVERY EFFORT HAS BEEN MADE TO REPRODUCE THE CENTRES, FORCE WOULD RESULT IN DAMAGE

Figure 6.6 Pooled results from using different models to combine the EFM data on the odds ratio scale



Further comparisons between the meta-analysis on the risk difference and odds ratio scale are fruitful. Table 6.7 provides the relative weights given to the individual studies for the two outcome measures using a fixed effect model. As can be seen, the percentage of the total weight given to each study in the analysis can change dramatically depending on the outcome scale chosen. Extreme examples are R8 which is given least weight on the risk difference scale (<0.1%) but most weight on the odds ratio scale (39.8%), and B5 which again gets very little weight on the risk difference scale (0.5%) but by far the greatest weight on the odds ratio scale (40.9%). As can be seen, many other sizable differences in weighting between outcome scales exist. Such differences can be explained by considering the event rates in the studies (also shown in Table 6.7). The risk difference metric is known to give large weight to trials with small event rates and the odds ratio metric gives larger weight to trials as they approach an event rate of 0.5.(Engles, et al. 2001; Tang, 2000) For example, since R8 has event rates in both arms that are an order of magnitude greater than all the other trials, it is given very little weight relative to the other studies on the risk difference metric, but much more weight on the odds ratio scale.

Having examined Table 6.7 it is not surprising that funnel plots on the different scales can look so different. Together, with previous findings (Tang and Liu, 2000) this highlights the need for a dissemination bias assessment measure which is robust to scale choice. While the lower heterogeneity found on the odds ratio scale may add credence to this analysis, problems such as the construction of numbers needed to treat exist with this scale.

Table 6.7 Relative weighting given to the EFM studies using fixed effect analyses on the risk difference and odds ratio scales

Study ID	% of weight given to study in fixed effect analysis		Event rate in non-EFM group	Event rate in EFM group
	Risk Difference	Odds Ratio		
RCTs				
R1	0.8	2.6	0.0057	0.0057
R2	1.0	3.5	0.0083	0.0041
R3	1.7	2.0	0.0000	0.0040
R4	2.1	2.3	0.0065	0.0000
R5	5.4	2.0	0.0022	0.0000
R6	6.4	2.0	0.0000	0.0020
R7	80.3	37.1	0.0021	0.0021
R8	0.0	39.8	0.1393	0.1452
R9	2.3	8.6	0.0027	0.0132
Comparative Cohort Studies				
C1	0.6	13.2	0.0017	0.0031
C2	0.1	3.6	0.0000	0.0022
C3	0.3	7.2	0.0016	0.0060
C4	1.0	6.7	0.0002	0.0031
C5	0.1	5.5	0.0018	0.0043
C6	21.4	3.1	0.0000	0.0002
C7	76.5	60.7	0.0002	0.0007
Before-and-After Studies				
B1	0.5	1.2	0.0000	0.0000
B2	0.2	10.7	0.0083	0.0083
B3	6.7	2.6	0.0005	0.0005
B4	2.8	2.6	0.0003	0.0003
B5	0.5	40.9	0.0055	0.0055
B6	9.4	14.0	0.0008	0.0008
B7	2.1	5.0	0.0004	0.0004
B8	6.0	10.3	0.0006	0.0006
B9	9.5	4.7	0.0003	0.0003
B10	62.3	8.0	0.0003	0.0003

6.5.2 Sparse data

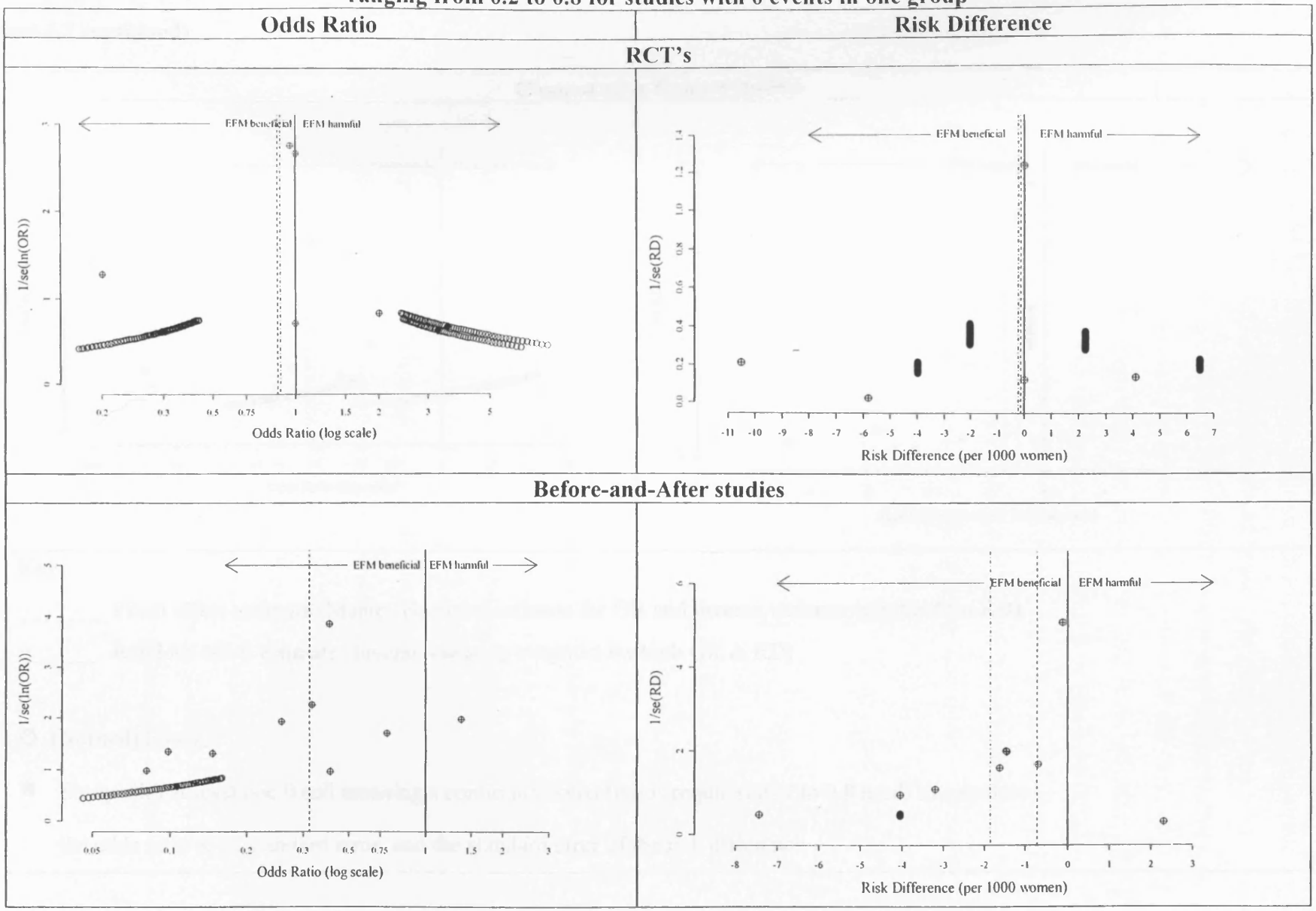
As mentioned previously, the event of interest in the EFM analysis, perinatal mortality, is relatively rare. Indeed there are four RCTs, one before-and-after, and two comparative cohort studies in which there were zero events in one study group, and many more where the number in each group was very small.. When calculating effect sizes from studies with zero events in one arm, it is necessary to add a continuity correction factor for the estimation of an odds ratio *and* its standard error, while it is only necessary to include a continuity correction factor for the standard error of the risk difference. Usually the correction factor of one half is added to each cell of the 2×2 table from which the effect size and standard error are calculated, as done previously in this paper, although it is not clear that this is always optimal, particularly when the number of subjects in each study arm is not balanced, as is the case for many of the EFM observational studies. (Sankey et al. 1996) While certain meta-analysis methods circumvent the need to calculate effect sizes directly for each study, and hence the need to use correction factors, (Sutton et al. 2000a) these factors are necessary for the construction of a funnel plot, Egger's test and Trim and Fill.

The issue of sparse data in meta-analysis datasets is considered in depth in Chapter 7. However the impact sparse data has on the assessment of dissemination bias is considered in this section. In the funnel plots presented in Figures 6.1 and 6.4, studies in which a continuity correction factor are used are plotted using a distinguishing symbol (as indicated in the key). While the use of such corrections may have little impact on the pooled estimate in most meta-analyses, they would appear potentially to have more impact on the appearance of a funnel plot, and hence any statistical methods based on it (such as Egger's test and Trim and Fill). The funnels from Figures 6.1 and 6.4 are redrawn in Figure 6.7, but to illustrate the impact a continuity correction factor could have on the appearance of a funnel, estimates were calculated using continuity correction factors between 0.2 to 0.8, at steps of 0.01 for studies in which 0 events occurred in one group. This produces 61 individual estimates which are plotted in a 'sweep' in the Figure.

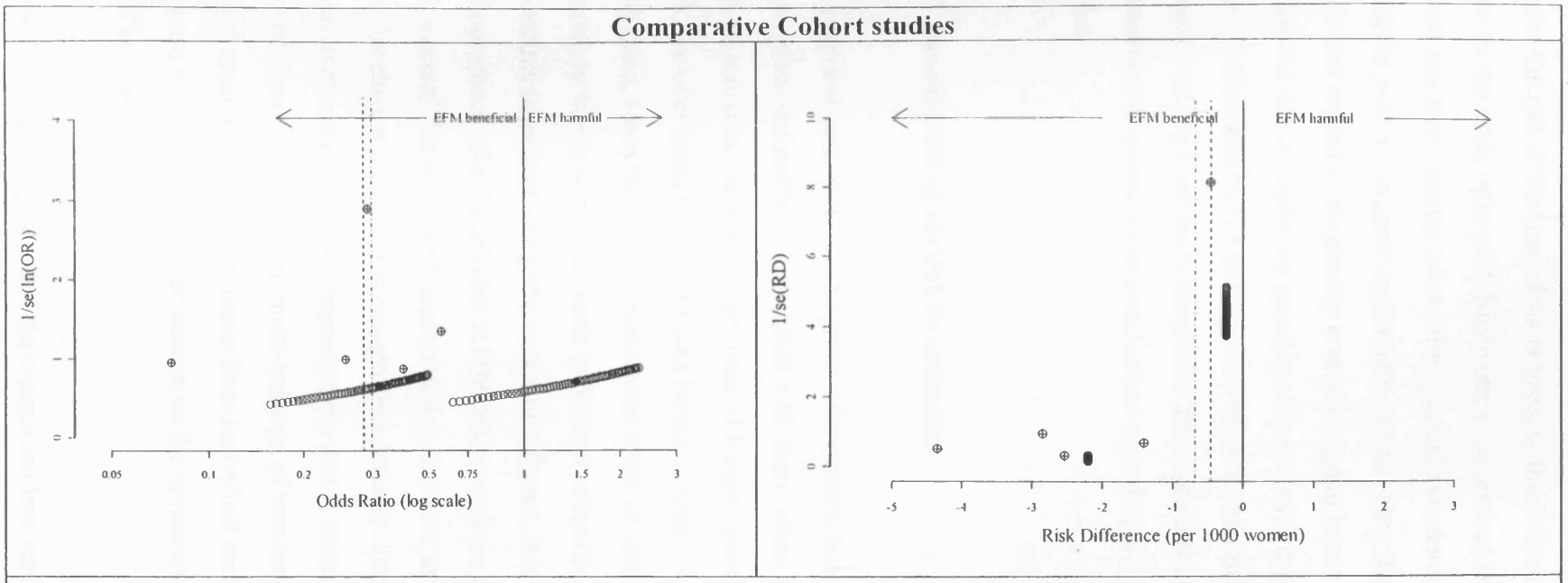
SPECIAL NOTE

THIS ITEM IS BOUND IN SUCH A MANNER AND WHILE EVERY EFFORT HAS BEEN MADE TO REPRODUCE THE CENTRES, FORCE WOULD RESULT IN DAMAGE

Figure 6.7 Funnel plots for EFM studies using (log) odds ratio and risk difference scales, showing effect of using continuity factors ranging from 0.2 to 0.8 for studies with 0 events in one group



(Figure 6.7 continued)



Key:

_____ Fixed effect estimate (Mantel-Haenszel estimate for OR and inverse variance weighted for RD)

_._._._. Random effect estimate (Inverse variance weighted for both OR & RD)

⊗ (Normal) Study

● Study with at least one 0 cell meaning a continuity correction is required (0.2 to 0.8 used) to calculate the odds ratio & its standard error, and the standard error of the risk difference

The impact of the correction factor on the funnels for the odds ratios is clearly larger than for the risk difference. This is because the estimate as well as the standard error varies on the odds ratio plot. Since one is concerned largely with asymmetry, fluctuations horizontally (along the x-axis) have more impact than vertical movement (along the y-axis), suggesting funnels plotted using the risk difference scale are more stable with regard to continuity corrections, and hence sparse data. In this example, the appearance of the funnel on an odds ratio scale for the comparative cohort studies is heavily influenced by the exact nature of the continuity correction factor used. Hence, the nature of any continuity correction factor used could potentially lead to different qualitative and quantitative conclusions regarding the degree of funnel plot asymmetry present.

6.5.3 Funnel plots of shrunken estimates

The previous section has highlighted why caution is needed in the interpretation of funnel plots and methods associated with them when data are sparse. In other areas of medical statistics, such as the analysis of league tables, (Marshall and Spiegelhalter, 1998) the over interpretation of data based on small numbers of events has been highlighted, since misleading results can often be obtained when not considering the uncertainty which surrounds each individual estimate. Although a measure of uncertainty is plotted along the y-axis of a funnel, it is difficult to appreciate how much the appearance of a funnel plot is affected by random error. Random fluctuations will "average out" for plots with many studies, however asymmetry will be more likely to occur by chance in smaller meta-analysis datasets. For this reason, funnel plots may appear asymmetric to some degree purely due to chance alone. (Steichen et al. 1998) This problem is amplified in meta-analyses of rare outcomes due to the large influence random error has on the estimates from individual studies. Hence, it should not be forgotten that the influence of chance on the appearance of the funnel plots in Figure 6.1 could be considerable.

One way of removing some of the chance random variation between studies is to plot the shrunken study estimates from a random-effect meta-analysis model, rather than the

original point estimates. A further advantage of looking at funnel plots of the shrunken estimates is that for the odds ratio scale an "exact" Bayesian model can be used which allows the plot to be constructed without the need for a continuity correction factor (such as that of equation 2.18). These plots of shrunken estimates are compared with the original funnels in Figure 6.8 for the odds ratio scale, and Figure 6.9 for the risk difference scale. Note that both the point estimates and the standard errors are "shrunk" towards the pooled effect. The visual impression of the RCT funnel for the odds ratio scale has not really changed, while the plot for the Before and After studies has, if anything, improved funnel symmetry. The comparative cohort study shrunken plot, however looks more asymmetric, largely due to the large shrinkage observed in the two studies in which there were zero events in one group of subjects. This highlights the issue of continuity correction factors distorting the appearance of a funnel plot.

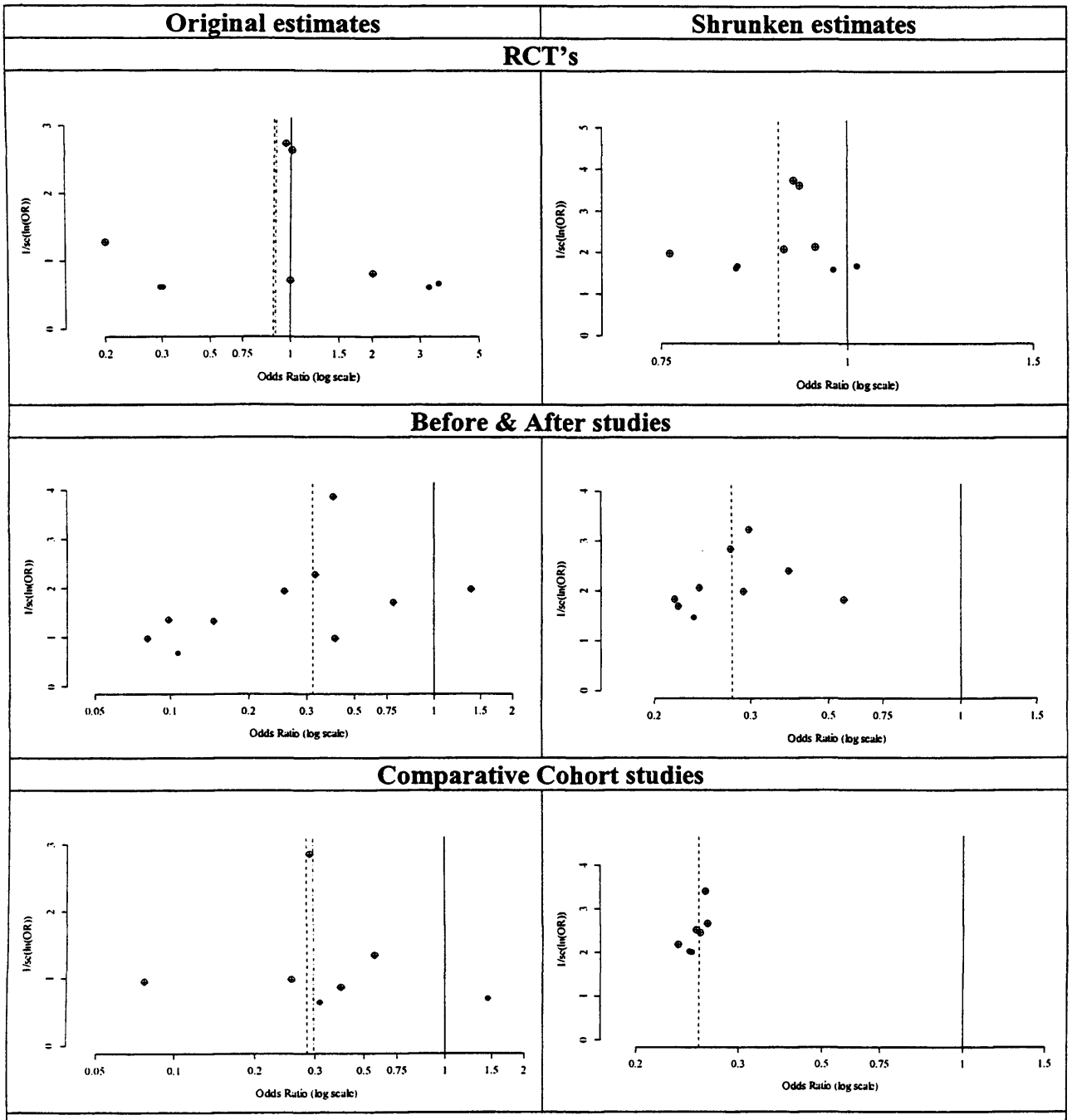
The changes between funnel plots of the original and the shrunken estimates on the risk difference scale are perhaps less striking than for the odds ratio scale. This can largely be attributed to the greater stability of the studies with a zero cell. Visually, publication bias would appear to be a concern for both the before & after and the comparative cohort studies.

Although there would appear to be advantages to considering funnel plots of shrunken rather than the original estimates (especially when data is sparse), there is a potential drawback. If publication bias is present, then the individual study estimates will be shrunken about a biased pooled estimate. In extreme situations this could potentially distort the appearance of the plot. Examination of both funnel plots of direct and shrunken estimates may safeguard against over interpretation. A further idea currently being pursued is to use shrunken funnels in combination with the Trim and Fill algorithm, so studies are shrunken around the pooled effect after Trimming.

SPECIAL NOTE

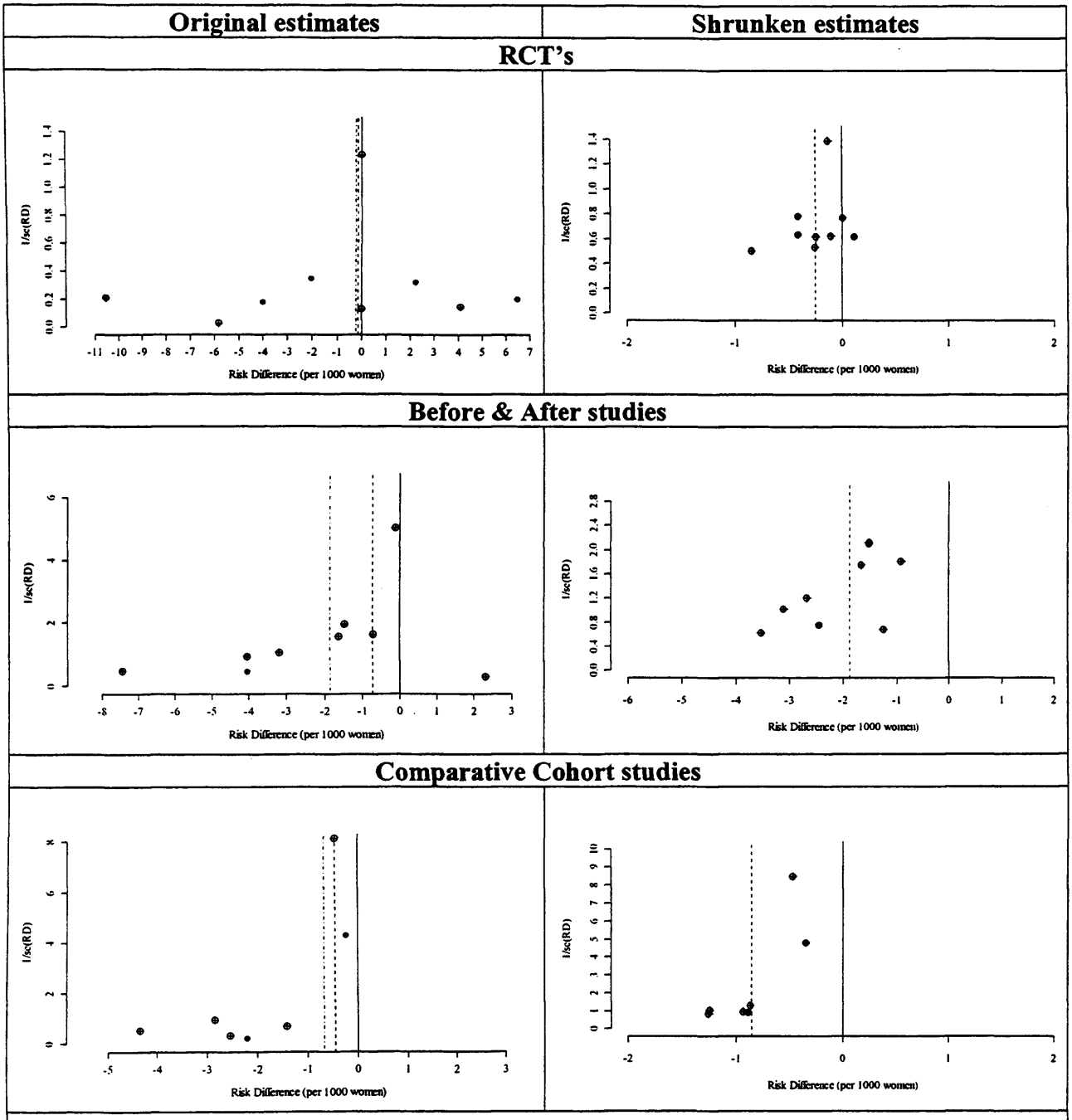
THIS ITEM IS BOUND IN SUCH A MANNER AND WHILE EVERY EFFORT HAS BEEN MADE TO REPRODUCE THE CENTRES, FORCE WOULD RESULT IN DAMAGE

Figure 6.8 Funnel plots of original and shrunken estimates on the odds ratio scale



Key:
LHS panels
_____ Fixed effect estimate (Mantel-Haenszel)
_____ Random effect estimate (Inverse variance weighted)
RHS panels
_____ Bayesian random effect estimate
⊗ (Normal) Study
● Study with at least one 0 cell meaning a continuity correction is required (0.5 used) to calculate
the odds ratio & its standard error, and the standard error of the risk difference

Figure 6.9 Funnel plots of original and shrunken estimates on the risk difference scale



Key:
LHS panels
_____ Fixed effect estimate (Inverse variance weighted)
_____ Random effect estimate (Inverse variance weighted)
RHS panels
_____ Bayesian random effect estimate
⊗ (Normal) Study
● Study with at least one 0 cell meaning a continuity correction is required (0.5 used) to calculate
the odds ratio & its standard error, and the standard error of the risk difference

6.5.4 The impact of model choice on the assessment of dissemination bias

The debate has run for many years regarding the superiority and suitability of fixed and random effect models for meta-analysis, although much of the time both models lead to substantially the same conclusions. Less prominent, although it has been mentioned previously (Givens et al. 1997; Greenland, 1994), is the concern that random effect models may be more seriously influenced by funnel plot asymmetry (and hence publication bias) than fixed effect ones because smaller studies are given relatively greater weighting in the meta-analysis. Despite that, although fixed effect results are included above, prominence are given to the random effect results, due to the fact that there is evidence of between study heterogeneity, and the author's belief that random effect models are generally more appropriate under such conditions.

While Egger's test for heterogeneity essentially uses regression based on a fixed effect model, the Trim and Fill method can be implemented using either model, and as Table 6.4 indicates for the before-and-after studies 'adjusted' results, differing by more than those from a standard analysis are obtained. The impact of model choice on the method of Trim and Fill is examined in the next section.

6.6 Limitations of the method of Trim and Fill

In this section the method of Trim and Fill, as it was implemented on the EFM studies, is brought under inspection. Two separate issues relating to the use of the method are discussed below.

6.6.1 The impact of model choice on the Trim and Fill method - fixed versus random effect models

The previous section noted the problem of different pooled estimates being produced with fixed and random effect estimates when funnel plot asymmetry is present since these differences can be propagated and amplified when using Trim and Fill. For example, the number of studies estimated as being missing for the Before & After

studies is four for the fixed effect model and one for the random effect model. A closer examination of the funnel and the pooled estimates reveals the reason for this discrepancy. Firstly, the fixed and random effect pooled estimates differ considerably; hence the 'centre' of the funnel is quite different for both models (Figure 6.1). This means that different points of symmetry are used by Trim and Fill which leads to different numbers of studies being estimated as missing.

The authors of Trim and Fill have given no guidance as to whether fixed or random effect estimation methods are superior when this method is applied. Indeed, in the original implementation of the method (although not used here) a random effect estimate was used to estimate the number of missing studies, and hence find the centre of the trimmed funnel, irrespective of whether a fixed or a random effect models was used to produce the final 'adjusted' estimate.[1]

An extreme example of dependence on fixed or random effect models on the conclusions is presented in Figure 6.10. This is a funnel plot of a diagnostic test meta-analysis detaset by Huicho et al, 1996 (Huicho et al. 1996) including 19 studies of Fecal leukocytes vs. Stool culture as screening tests for infectious diarrhea, using the diagnostic *d* measure as outcome. (Hasselblad and Hedges, 1995) Here the funnel is highly asymmetric, and clearly there are issues about whether publication bias mechanisms are likely to have caused this appearance. However it would appear that Trim and Fill is not robust to such irregular funnels. In this example, the pooled value for *d* is 0.71 (0.63 to 0.79) and 1.44 (1.05 to 1.83) for fixed and random effect models respectively. When trim and fill is applied to this meta analysis zero studies are estimated missing under a random effects model, while ten are estimated missing under a fixed effect model, and the outcome is adjusted to 0.41 (0.34 to 0.48). These are clearly wildly different answers, and it shows that choice between random and fixed effect often appears more critical *when used in conjunction with trim and fill* than it does in the main decision about assumption for pooling model.

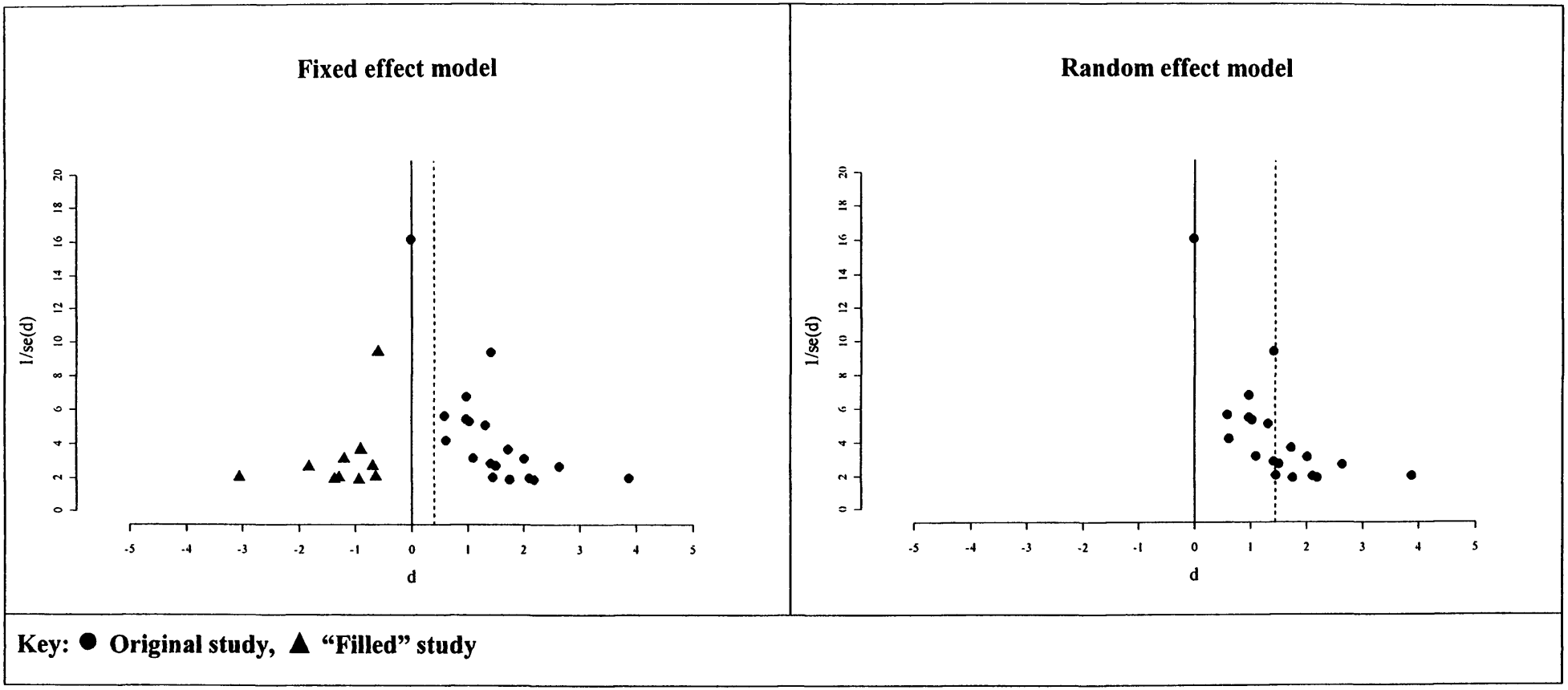
[1] Personal communication with Sue Duval

Trim and fill only assesses the equality of ranks of studies either side of the pooled estimate and does not consider their size. For this reason, the funnel considered around the random effect estimate has a balance in the location of the studies with respect to the x-axis, but not the y-axis. This potential weekness of trim and fill requires further consideration, perhaps a modification of the method to include consideration of the appearance of the funnel with respect to the precision (y) as well as the effect size (x) axis.

Another possibility which is intuitively appealing, and conceptually easy to justify, is the use of fixed and random effect models in combination. Since the fixed effect estimate is less influenced by asymmetry of a funnel than the random effect estimate, this could be used to estimate the number of studies "missing" and hence the centre of the funnel. When these missing studies are imputed, the new pooled estimate could be calculated using the random effect model. In this way, the pooled estimate is not as influenced by publication bias as the random effect one, however the heterogeneity in the studies can still be accounted for in the filled dataset.

If this procedure is used for the Before & After studies in the EFM example, the adjusted risk difference becomes –0.43 (-2.12 to 1.64), which halves the intervention benefit from the random effect estimate. The validity of such an approach is the subject of ongoing research, and simulation studies (similar to those used to validate trim and fill originally (Duval and Tweedie, 2000c, Duval and Tweedie, 2000b)) are required to examine the properties of this modified estimator. Further, the relationship between statistical heterogeneity and publication bias needs further exploration generally.

Figure 6.10 Trim and fill analysis of test of Fecal leukocytes vs. Stool culture



6.6.2 Artificially narrow confidence intervals? Bootstrapped confidence intervals for Trim and Fill

A further criticism which could be levelled at the Trim and Fill method is that the adjusted confidence interval is too narrow because precision in the pooled estimate is increased due to information included in the filled studies (which are only speculated and not known to exist). It would be desirable for the confidence interval not to reduce in magnitude due to the inclusion of these "fictional" studies. Clearly achieving this is not straightforward as one would want to include the filled studies, to properly estimate the standard error of the pooled estimate and the between study variance in a random effects model. Interestingly, this problem is also related to the heterogeneity and publication bias relationship issue. Since the filled studies are the mirror image of the most extreme ones on the unbiased side of the funnel, heterogeneity will increase in the filled dataset which works 'against' the reduction in confidence interval width induced by including the imputed studies.

One possible solution to this problem is to construct a bootstrap type confidence interval for the pooled estimate (bootstrapped confidence intervals for meta-analysis are considered further in Section 7.6 for use with sparse data). The proposal here is to sample the number of studies in the *original* dataset from the dataset including the filled studies (with replication). Hence, this is a slightly modified bootstrap procedure because usually samples of the size of the original dataset are used. In this way it is possible to consider the full set of studies (including the filled ones), and incorporating the heterogeneity associated with these, while basing precision on the fact that only the number in the original dataset are known about. This solution is not perfect as each of the studies has a different weight associated with it, and hence the actual 'quantity' of evidence included in each bootstrap sample will vary and not be exactly equal to that included in the original – non-'filled' dataset.

This procedure is carried out for the observational studies in the EFM dataset, and the results presented in Table 6.9. In all cases 2000 bootstrap samples were taken to construct confidence intervals, and the bias corrected estimates are reported. Both

bootstrapped results sampling the original number of studies and the original plus the filled number of studies is included for comparison.

Table 6.9 Results of using Bootstrapping to generate confidence intervals for trim and fill estimates

	Before & After	**Comparative Cohort**
Fixed Effect		
Dataset asymptotic 95% CI	-0.72 (-1.03 to –0.42) Width: 0.61	-0.46 (-0.67 to –0.25) Width: 0.42
Dataset bootstrapped 95% CI	(-2.15 to –0.29) Width: 1.86	(-2.86 to –0.27) Width: 2.59
Trim and Filled asymptotic 95% CI	-0.52 (-0.82 to –0.22) Width: 0.60	-0.42 (-0.63 to –0.21) Width: 0.42
Trim and Filled Bootstrapped 95% CI (sample number: original + filled studies)	(-1.60 to –0.06) Width: 1.54	(-1.45 to 0.42) Width: 1.87
Trim and Filled Bootstrapped 95% CI (sample number: original studies)	(-1.78 to 0.07) Width: 1.85	(-2.21 to 1.16) Width: 3.37
Random Effect		
Dataset asymptotic 95% CI	-1.86 (-2.84 to –0.89) Width: 1.95	-0.68 (-1.25 to –0.11) Width: 1.14
Dataset bootstrapped 95% CI	(-3.36 to –0.92) Width: 2.44	(-3.00 to –0.29) Width: 2.71
Trim and Filled asymptotic 95% CI	-1.63 (-2.63 to –0.64) Width: 1.99	-0.43 (-1.12 to 0.27) Width: 1.39
Trim and Filled Bootstrapped 95% CI (sample number original + filled studies)	(-3.22 to –0.75) Width: 2.47	(-1.75 to 0.71) Width: 2.46
Trim and Filled Bootstrapped 95% CI (sample number original studies)	(-3.06 to –0.66) Width: 2.40	(-2.31 to 1.28) Width: 3.59

This table is rather difficult to interpret since all the bootstrapped confidence intervals appear to have very different coverage from the corresponding asymptotic ones. This echoes the problems of applying the bootstrap to the meta-analysis datasets of rare outcomes in Chapter 7. Secondly, although the widths of each interval are given, the unadjusted and the publication bias 'adjusted' interval widths would not be expected to be the same for the random effect model due to the differences in the estimation of the between study variability parameter and its influence on the pooled estimate. It would appear that the use of the bootstrap method for meta-analysis is rather unpredictable, and before it can be applied in combination with Trim and Fill, further investigation into its performance is required. This approach is not pursued further here, due to the improvements offered by the use of the Bayesian bootstrap approach described in the next section.

6.6.3 Adapted Bayesian bootstrapped confidence intervals for Trim and Fill

Rubin (Rubin, 1981) first described the Bayesian bootstrap, and since it is not widely used a non-technical summary of the procedure follows. In the normal bootstrap each of the original observations is sampled with a weight/frequency. This is 0 or 1 or 2 . . . . or n, where n is the original sample size. The Bayesian bootstrap does the same except that the weight is not an integer but may take any value between 0 and n.

Use of such fractional weights results in a smoother distribution (now a *posterior* distribution) for the statistic of interest, especially when the sample size is small. (Rubin, 1981) However, the main appeal for using this method in combination with trim and fill is that the total study weighting pertaining to the *original* set of studies, rather than the 'filled' dataset, can be sampled over the studies in the 'filled' dataset. In this sense the Bayesian bootstrap method is modified (in the same way as the standard bootstrap above) because normally sampling equal to the total weight of the dataset is used. Hence, unlike the classical bootstrap approach, exactly the desired total study weight is sampled at every iteration.

This approach is illustrated using the Before and After EFM studies. Considering the concerns raised in Section 6.6.1, the number of studies estimated as missing by the fixed

effect approach (four) is used in combination with a random effect estimate. Since the adapted Bayesian bootstrap is used, a Bayesian meta-analysis model is also employed where all prior distributions are specified as vague. The results of this together with the Classical results for completeness are reported in Table 6.10

Table 6.10 Results of using adapted Bayesian Bootstrapping to generate estimate for Trim and Fill augmented before & after study dataset

Model Used	Estimate (95% CI/CrI)
Classical Random Effect	
Original dataset asymptotic 95% CI	-1.86 (-2.84 to –0.89) Width: 1.95
Original dataset bootstrapped 95% CI	-1.86 (-3.36 to –0.92) Width: 2.44
Trim and Filled asymptotic 95% CI	-0.74 (-1.76 to 0.28) Width: 2.04
Trim and Filled Bootstrapped 95% CI (sample number original + filled studies)	-0.74 (-1.93 to 0.80) Width: 2.73
Trim and Filled Bootstrapped 95% CI (sample number original studies only)	-0.74 (-2.33 to 1.07) Width: 3.40
Bayesian Random Effect	
Original dataset MCMC	-1.83 (-3.26 to –0.74) Width: 2.53
Original dataset Bayesian bootstrapped MCMC	-1.48 (-3.03 to –0.46) Width: 2.57
Trim and Filled MCMC	-0.65 (-2.39 to 1.19) Width: 3.58
Trim and Filled Bayesian bootstrapped MCMC (weight of original + filled studies)	-0.71 (-2.22 to 0.76) Width: 2.97
Trim and Filled Bayesian bootstrapped MCMC (weight of original studies only)	-0.73 (-2.21 to 0.82) Width: 3.03

This table (like Table 6.9) is difficult to fully interpret. There are certain inconsistencies, for example, the wide credible interval from the Trim and Filled MCMC approach, which is wider than either of the Bayesian bootstrapped estimates on the filled studies. As with the first modification in the previous section, research is being carried out to assess the performance of the Bayesian bootstrapped Trim and Fill estimate, but it does present potential advantages. A further modification that is also being considered is that it would be no longer necessary to round the number of studies estimated as missing (by Trim and Fill) up to a round number (as done currently) if a Bayesian bootstrap confidence interval is being constructed.

It should be noted that although the fractional weighting idea described above is phrased in terms of the Bayesian bootstrap, the same method could be considered from a classical perspective also.

6.7 Discussion

This chapter has considered the possibility of dissemination bias, in particular bias causing funnel plot asymmetry in studies with differing designs. In the introduction, reasons were given why different types of studies may be susceptible to differing dissemination biases. The EFM example, where evidence is available from different sources, was examined, and on the risk difference scale at least there was strong evidence of more extreme funnel plot asymmetry for the observational studies, which could partly explain the more favourable results observed in the observational studies.

Clearly, other reasons exist why asymmetry of the funnels is observed. Study quality (Petticrew et al. 1999), systematic heterogeneity of effects between small and large studies (Sterne et al. 2000b) and chance may be explanations. In addition, reasons associated with outcome scale choice would appear to be an issue here and the use of continuity correction factors, necessary when outcome data are sparse, has also been demonstrated as having an effect. Both these factors need further investigation in relation to the assessment of dissemination bias in meta-analysis.

Concerns are often expressed about basing treatment efficacy decisions on non-randomised evidence, largely being due fear of bias in estimated inherent in observational study designs. The concern with the EFM studies, and similar datasets is that even if the observational studies individually are unbiased, a meta-analysis of such evidence may be biased due to large levels of dissemination bias. If dissemination bias is generally a bigger (or indeed just as importantly a different) problem in the non-randomised study literature, this has important implications for meta-analysis and research synthesis more generally, including the use of observational studies in economic evaluation (Freemantle and Mason, 1997).

For RCTs long term measures are being put in place to try and alleviate the problem of dissemination bias. Steps are also being made to encourage the prospective registration of trials via trial registries.(Simes, 1986) Such alleviative measures are not in place, or even in prospect, for observational studies. It is difficult to envisage what measures could be taken to reduce the problem of dissemination bias in observational studies, where, as discussed above, the definition of the primary analysis of a study may be ill-defined, making prospective registration an impossibility. Additionally, some types of observational study do not require ethical approval, further reducing the feasibility of tracing them. It would appear that dissemination bias in observational studies will not diminish in either the short or the medium term.

Some technical issues with the analysis require comment. Firstly, as for the cholesterol analysis in Chapter 5, it may appear counter intuitive that uncertainty in results increased when the observational evidence was included in the three level synthesis model. Although this is due to heterogeneity between the results of the different study types and is considered desirable, since often evidence will be available from a limited number of sources, the between study heterogeneity term will often be estimated with considerable uncertainty which is reflected in the pooled result. The degree to which the number of sources of evidence influences the results requires further investigation, together with the robustness of the results to the assumption of Normality at the top level of the model.

In the process of carrying out this assessment potential problems with the method of Trim and Fill were identified. These are due to the influence of dissemination bias on the random effect estimate, and a too precise confidence interval due to the added information included from the 'filled' studies. The possibility of estimating the number of missing studies using a fixed effect model, and then producing the 'adjusted' estimate using a random effect model when heterogeneity is present (and in may only become apparent when the 'filled' studies are included) would seem entirely sensible, but a further simulation study to assess its performance is required. The second modification explored is the use of bootstrap methods to reduce the total weighting in the meta-analysis to that of the original studies. The Bayesian bootstrap would appear most suited for this, although discrepancies between bootstrapped confidence intervals from those derived by more traditional methods, indicates that the properties of bootstrapped confidence intervals for meta-analysis require further investigation before such a method can be recommended. Further, the Bayesian evidence synthesis framework is conditioned on the results of the Trim and Fill method. Hence, the uncertainty in the Trim and Fill analysis is not reflected in the final synthesis results. More desirable would be an approach which replaced Trim and Fill with a Bayesian approach to dealing with dissemination bias that could be implemented simultaneously with the evidence synthesis. This is the also the subject of ongoing research.

There are sections of the 'meta-analysis community' who are generally opposed to combining RCT and non-RCT evidence. I believe that this work goes some way to demonstrate that, with proper adjustment, use of observation studies and combination of observational and RCT evidence can at least in principle be sound. Researchers should be aware that different sources of evidence may be affected by differential levels of dissemination bias. This chapter has explored how to incorporate an assessment of dissemination bias into a generalised synthesis context using pre-existing methods which are relatively simple to implement in practice. While assessments such as those described in this thesis are recommended, awareness to their lack of robustness to outcome scale, model choice and correction factors is necessary. Sparse outcome data, such as that present in the EFM dataset, would appear to amplify this lack of robustness. Further research into methods to assess dissemination bias - ideally, methods that are invariant to the factors identified above - are needed. Some of these problems will be

apparent in meta-analyses restricted just to RCT evidence as well as in generalised synthesis of evidence.

Chapter 7 Synthesis of studies reporting rare outcomes incorporating the case studies of adverse effect of hormone replacement therapy and breast implants

7.1 Introduction

It has frequently been stated that one of the benefits of meta-analysis is that through the combination of multiple studies, it is capable of yielding statistically significant effects not possible in individual studies due to their limited power. This is particularly true in instances where the event of interest is rare such as side effects of new interventions or epidemiology of rare diseases. Further, RCTs are usually not powered to examine any but the most common of side effects, and often valuable information is collected in post marketing surveillance schemes outside a trial setting. In such situations data from different studies exist making pharmaco-epidemiology a very relevant topic to address using cross-design synthesis, and one in which general synthesis of evidence models can make a valuable contribution. The EFM data synthesised in the previous chapter provides an example of sparse outcome data where the RCT evidence is inconclusive.

Specific problems are encountered when synthesising sparse outcome data, not least that some traditional meta-analysis methods fall down altogether; these are described in the next section. Following this, a large proportion of this chapter explores the limitations of current meta-analysis methods for sparse data, both classical and Bayesian; these are outlined in Section 7.2.6. Although these explorations are not strictly in a generalised synthesis framework, such research was felt necessary as a foundation before synthesis of different study types was attempted, and hence is directly relevant to the main topic of this thesis. In the latter sections of this chapter, two examples of the generalised synthesis of rare events are considered.

7.2 Problems encountered when pooling rare outcomes

There are a number of related statistical problems encountered when attempting to pool studies with rare outcomes, in particular when there are zero events in one or more arms of a study. Some of these are general, while others are specific to certain outcome measures; these are considered below.

7.2.1 General problems - validity of methods based on normal assumptions

The assumption that effect sizes from each study are normally distributed is frequently made in a meta-analysis model. This assumption is questionable if studies have small numbers of events. (Carlin, 2000) Additionally, the overall pooled estimate is also usually assumed to be normally distributed, and confidence intervals for the pooled estimate are constructed on the basis of this assumption. (Adams et al. 1997) These normal approximations are likely to be less accurate when sample sizes are small, when there are large differences in sample size between the experimental and control groups, and with very large effect sizes. (Hedges and Olkin 1985)

7.2.2 Issues with the odds ratio

Several methods are available for combining studies using the odds ratio scale as described in Section 2.2. (Fleiss, 1993) An odds ratio cannot be calculated directly if there are zero events in either group being compared using the Mantel Haenszel, inverse-variance weighted method, or a classical random effect model without modification. The method of Peto, can however, produce pooled odds ratios from the original 2 x 2 table without the need to calculate odds ratios for individual studies. A continuity correction factor, usually adding a half to each cell of the 2 by 2 table from which the odds ratio is being calculated, is usually applied.

In instances where there are severe imbalances between the numbers in each group, such a procedure can produce misleading results. For example, consider the comparative

cohort studies of EFM presented in Table 6.2. In the second study in this table (C2) there are 15 deaths out of 6836 births on the group not given EFM, while there were no deaths out of the 150 births for which EFM was used. Adding 0.5 to each cell of the 2 x 2 table for this study produces odds of death not on EFM of 0.0023 (15.5/6836.5), compared to 0.0033 (0.5/150.5) for those not on EFM. This produces an odds ratio of 1.4, suggesting a harmful effect of EFM - even though there were no deaths observed in the EFM group.

Such a distortion can happen when there are differing numbers of persons in each arm of the study, but even if both arms contain similar numbers of persons, a distortion in comparative effect occurs if the number of events in the non-zero arm is small. The Mantel-Haenszel method has better frequentist properties than the inverse-variance weighted estimates, especially when data are sparse in some of the studies being combined, because the inverse variance method relies on approximations to the binomial likelihood. (Carlin, 2000) It is also possible to combine odds ratios using more complex methods including exact methods, (Fleiss, 1993) Bayesian methods utilising MCMC simulation, and logistic/Poisson regression. Some of these are explored below.

7.2.3 Issues with the relative risk

No special methods exist for calculating relative risks, so the standard inverse-variance weighted method is often used for a fixed effect model, and the standard extension for incorporating a random effect. As noted in the odds ratio section above, a continuity correction is required if there are zero (or all) events in either arm, which can produce the same problems highlighted there. Less attention has been given to applying more complex methods such as exact estimation, or MCMC simulation to the relative risk scale, compared with the odds ratio.

7.2.4 Issues with the risk difference

The risk difference is used less frequently for meta-analyses than either the odds ratio or relative risk scale. It has come under criticism due to statistical constraints - the range of variation is limited by the magnitudes of the rates in each group being compared. (Fleiss, 1994) The practical implications of this for meta-analysis are that artificial heterogeneity may be introduced if the rates in the control groups in the studies combined varies. The risk difference does have benefits when dealing with rare outcomes, however. Unlike an odds ratio or relative risk, a risk difference can still be calculated if there are zero events in either group, and hence no correction factor is required. However, a correction factor is still necessary to calculate the standard error of a risk difference. Additionally, if there are no events in either group, information is still included in a meta-analysis using the risk difference scale even without the use of a continuity correction factor (this is not the case for the odds ratio or the relative risk where studies with zero events in both arms are usually excluded). This issue was considered further in Section 6.5.2 where the effect of continuity correction factors on the appearance of funnel plots is considered. On a practical level, the risk difference scale is easier to interpret to consumers of meta-analysis than a relative scale such as the odds ratio, especially when rare events are being considered. For example, a trial which randomises 1000 patients to each of two arms and observes two adverse event in one group, and one in another would have an odds ratio of approximately 2, which is difficult to interpret without knowing the low absolute rate of adverse events in the two groups. In such a situation reporting a risk difference of 0.001 is more useful.

7.2.5 Previous work on methods for combining rare outcomes

Several investigations into the best ways of pooling studies with rare events have been carried out. (Sankey et al. 1996; Mengersen et al. 1995; Deeks et al. 1999; Adams et al. 1997; Carlin, 2000) Sankey et al. (Sankey et al. 1996) examined the effect of using the continuity correction factor of 0.5 on the Mantel-Haenszel odds ratio, the DerSimonian and Laird (random effect) odds ratio and the rate difference. They found that the uncorrected method performed better only when using the Mantel-Haenszel

odds ratio with very little heterogeneity present, and in all other situations the corrected methods performed better and were recommended. These simulations were limited to a ratio of sample sizes in of the control and treatment arms of between 1:1 to 2:1, and hence the effect of a continuity correction factor on severe group imbalance was not examined.

Deeks et al. (Deeks et al. 1999) examined a much wider group of methods for combining studies with rare events. Methods examined include Mantel-Haenszel odds ratio, Peto odds ratio, inverse variance weighted method, DerSimonian and Laird random effect odds ratio, and the MH and D&L risk difference methods. In addition, non-standard methods including fixed and random effect Poisson regression and exact methods were also examined. No detailed account of this work is currently available, however preliminary reports suggest that the Peto method provided the least biased and most powerful method of pooling study results with rare events among the commonly used methods except when trial group sizes are severely imbalanced.

Adams et al. (Adams et al. 1997) discuss the use of re-sampling methods for meta-analysis. Confidence intervals based on bootstrapping methods are described for both fixed and random effect models, which present a non-parametric alternative to the standard ones based on normal approximations. These were developed for combining ecological data, and have not been used in the combination of medical data. Investigations have found that such confidence intervals are sometimes more conservative than those based on normality assumptions, but these were based on meta-analyses of *continuous* outcomes only. The application of bootstrap methods to meta-analysis datasets with sparse binary outcomes has never been reported and is considered below.

Carlin (Carlin, 2000) recently considered the use of MCMC methods implemented using BUGS to fit an 'exact' Bayesian model which does not require the binomial likelihood to be approximated by a Normal distribution for the risk difference scale. This approach is also considered in more detail below.

7.2.6 Further work addressed here

Pervious research has begun to tackle the problem of sparse data in meta-analysis. From the work of Deeks et al. (Deeks et al. 1999) it would appear that the Peto method is the method of choice for carrying out a fixed effect model, where there are no imbalances between groups being compared. However no comparison with bootstrap or Bayesian simulation methods has been carried out, and less is known concerning the use of random effect models.

Several issues related to meta-analysis of rare events are thus explored further in the remainder of this chapter:

1) The use of continuity correction factors. These are generally used to allow studies to be included in a meta-analysis that would otherwise be excluded. Although several authors have advocated their use, (Sutton et al. 1998) their specific form is rarely discussed. For instance, usually 0.5 is added to all cells of a studies 2 by 2 table. Other possibilities exist, including adding a smaller correction fraction, a correction factor proportional to the size of the study arm, or a correction factor corresponding to no effect (e.g. an odds ratio of one etc).
2) The use of Bayesian methods to overcome the problems traditional methods encounter with sparse binary data, such as the need for a continuity correction factor, or normality assumptions discussed above. This work builds on that published previously (Smith et al. 1995b; Carlin, 2000) utilising the use of MCMC 'exact' sampling methods. The use of data inflation methods in this context is also explored. Data inflation is a currently un-documented technique yet to be validated and is described in detail later in the chapter.
3) The use of re-sampling bootstrap methods to overcome the need to make distributional assumptions about outcomes which may not hold for sparse binary data when using frequentist methods.
4) Combining sparse data from studies with different designs.

In addition, this chapter considers two examples in which outcomes are rare and considers how the studies should be combined. In the first example, a re-analysis of a

meta-analysis dataset consisting exclusively of data from RCTs, investigating the risk of breast cancer and cardiovascular events following hormone replacement therapy, is presented. This analysis is extended to include data from a large case-control study.

The second example examines the association between breast implants and risk of connective tissue disease. This adverse event is very rare, and data are combined from observational studies with different designs, where the data are reported in different formats.

7.3 Example 1: RCT evidence on the impact of postmenopausal hormone therapy on cardiovascular events and cancer

The first example is a meta-analysis, published previously, (Hemminki and McPherson, 1997) which considers the impact of postmenopausal hormone therapy on cardiovascular events and cancer. This meta-analysis included only data from clinical trials, and was designed to explore further findings in surveys which suggested that such hormone therapy may decrease the incidence of cardiovascular diseases and increase the incidence of hormone dependent cancers. In light of such findings, RCTs have been undertaken to verify the findings, but results of these will not be available for some years. Evidence is currently available from 23 trials designed to examine other, short term, aspects of postmenopausal hormone therapy. Data from these trials are presented in Table 7.1 (Hemminki and McPherson, 1997). Note that several of the trials had more than one active treatment arm in which different regimens and doses were administered; these have been merged for the purposes of this analysis. There are a number of features which are striking about this table. Firstly, the sparseness of the outcome reporting. A dash in the table indicates no information was given for that outcome in the primary trial report. In the original analysis this was assumed to mean no events actually happened (i.e. the dashes were turned into zeros). Secondly, there are many "true" zeros in the events, indicating extremely sparse data, which presents problems described above. There are 16 studies with 0 or unreported cardiovascular events, and similarly 14 studies for breast cancer; these studies will only be included in a meta-analysis if the risk

difference scale is used. Of the remaining studies, five have zero events in one arm for both outcomes. Thirdly, largely due to certain studies having multiple treatment arms which have been merged, there are large imbalances in the groups being compared, e.g. study number two has approximately four times as many women in the active treatment arm. Hence, this dataset has both problems of great data sparcity and imbalance in the study groups (as analysed here).

Table 7.1 Data from 23 trials of postmenopausal hormone therapy and cardiovascular events and breast cancer

Study number	women allocated		cardiovascular disease		breast cancer	
	Control	treatment	events in control	events in treatment	events in control	events in treatment
1	137	1128	-	-	0	6
2	174	701	0	5	1	7
3	78	39	-	-	1	0
4	42	40	0	0	-	-
5	32	46	-	-	-	-
6	14	15	1	0	-	-
7	51	100	-	-	0	2
8	39	36	-	-	1	1
9	25	50	0	0	-	-
10	19	41	-	-	1	0
11	40	116	-	-	1	1
12	16	15	0	1	-	-
13	19	21	-	-	-	-
14	20	20	1	1	-	-
15	39	61	-	-	-	-
16	54	60	-	-	-	-
17	24	76	-	-	-	-
18	48	44	-	-	-	-
19	26	29	0	1	-	-
20	121	56	-	-	1	2
21	84	84	3	1	4	0
22	30	120	0	0	-	-
23	66	68	0	3	-	-

The analysis carried out by the original investigators uses a "non-standard" meta-analysis technique. A simple marginal analysis was carried out summing totals across all individual trial 2 by 2 tables to produce one large one. For example there are a total of 5 cardiovascular disease events out of 1041 women in the control group, and 12 such events out of 1818 women in the treatment group. Calculating the odds ratio for this single table produces an odds ratio of 1.38 ($(12 \times 1036)/(5 \times 1806)$). Calculating a 95% confidence interval by standard methods produces an interval of 0.48 to 3.92 (which differs fractionally from that calculated in the original paper, due to discrepancies in data between sections of the original report (Various authors 1997)). In a similar manner, the odds ratio calculated for breast cancer events is 1.09 (0.50 to 2.35) (Note this is based on 10 breast cancer events in the control groups, as suggested by the data table, rather than 9 used in the calculation in the original paper (Various authors 1997)). Although, such a method of pooling data may seem appealing, there are severe problems with it, especially when event rates are heterogeneous across groups. This issue has been well documented (Simpson, 1951) and is often referred to as Simpson's or Yule's paradox. It does however allow data from all studies to be included in an analysis on an odds ratio scale.

When the meta-analysis investigating the impact of postmenopausal hormone therapy on cardiovascular events and cancer using data from clinical trials was originally published in 1997, (Hemminki and McPherson, 1997) it generated a lot of correspondence to the British Medical Journal. The meta-analysis was criticised for a number of reasons including:

1) Omission of epidemiological and other direct intervention studies
2) Drawing of strong conclusions despite being based on small numbers of women (4000), who had very few events.
3) Concern that the reporting of adverse events could have been biased.
4) Consideration of generally only the short term effect of HRT
5) Use of flawed marginal method of analysis
6) Lack of accounting for dosing regimes and duration of follow-up
7) The strong influence of one trial (number two in Table 7.1)

8) Inclusion of one trial (number 23) in which there were 3 events in the treatment group related to a high dose of the oestrogen mestranol, which has not been used for over 20 years. Removing this study changes the pooled estimate considerably.

Several of these shortcomings are addressed in the sections which follow. The fundamental issue of the method of analysis (point five) is considered first. This is followed by sections addressing the concern of biased reporting (point 3) through the use of sensitivity analysis, and combining further sources of evidence (point 1).

7.3.1 Comparison of classical methods to pool odds ratios

Their published results are compared with results obtained with a range of other classical statistical methods available to pool odds ratios. In all these analyses it is assumed that all the dashes in Table 7.1 are zeros (which was also assumed by the original authors in the marginal analysis summarised above). The fixed effect methods used are the Mantel-Haenszel, Peto, and inverse-variance weighted methods. The standard DerSimonian and Laird estimator was used as the random effect estimator. Continuity correction factors of 0.1, 0.5 and 0.5 split over the two arms, the split being proportional to the size of the arm were all applied to the inverse-variance weighted method and the random effect model. The latter correction factor corresponds to an odds ratio of one (before any events are taken into consideration). These corrections were applied using two strategies: 1) only when absolutely necessary (i.e. when there are zero events in one arm); and 2) to all studies in the dataset. The effect of excluding all studies with zero events in both arms, and zero events in either arm was also investigated. The results of applying these different pooling approaches are presented in Table 7.2.

Table 7.2 Results of pooling odds ratios for Cardiovascular and Breast cancer risk using a range of simple classical approaches

	Cardiovascular risk		Breast cancer risk	
Odds ratio combine method	Number of studies included	Pooled estimate (95% CI)	Number of studies included	Pooled estimate (95% CI)
FIXED EFFECT ANALYSES				
Marginal analysis (1 summary table)	23	1.377 (0.484 to 3.919)	23	1.089 (0.504 to 2.351)
Inverse variance (no continuity correction)	2	0.505 (0.085 to 2.998)	4	1.473 (0.427 to 5.085)
Inverse variance (including studies with at least one event & 0.5 continuity correction when necessary)	7	1.309 (0.436 to 3.929)	9	0.942 (0.376 to 2.360)
Inverse variance (including studies with at least one event & 0.5 continuity correction to all these studies)	7	1.239 (0.446 to 3.438)	9	0.916 (0.402 to 2.087)
Inverse variance (0.5 continuity correction to all studies)	23	0.881 (0.433 to 1.789)	23	0.845 (0.441 to 1.617)
Inverse variance (including studies with at least one event & 0.1 continuity correction when necessary)	7	0.980 (0.216 to 4.450)	9	1.215 (0.390 to 3.783)
Inverse variance (including studies with at least one event & 0.1 continuity correction to all these studies)	7	0.972 (0.224 to 4.218)	9	1.185 (0.396 to 3.543)
Inverse variance (0.1 continuity correction to all studies)	23	0.854 (0.252 to 2.891)	23	1.089 (0.403 to 2.937)
Inverse variance (including studies with at least one event & proportional continuity correction when necessary)	7	1.157 (0.304 to 4.407)	9	1.010 (0.346 to 2.946)
Inverse variance (including studies with at least one event & proportional continuity correction to all these studies)	7	1.131 (0.320 to 3.998)	9	1.019 (0.374 to 2.778)
Inverse variance (proportional continuity correction to all studies)	23	1.075 (0.410 to 2.819)	23	1.014 (0.438 to 2.346)
Mantel-Haenszel (including studies with at least 1 event & 0.5 continuity correction where necessary)	7	1.423 (0.559 to 3.619)	9	0.846 (0.388 to 1.841)
Peto	7	1.679 (0.610 to 4.621)	9	0.909 (0.386 to 2.144)

Only the fixed effect estimates are shown in Table 7.2 because in all instances the between study variation was estimated to be zero, so all random effect results were

exactly as the corresponding fixed effect ones presented. Although none of the odds ratios come close to statistical significance, there is considerable variation between the point estimates and 95% confidence intervals obtained with the various methods. For both outcomes, estimates both larger and smaller than unity occur.

In section 6.5.2 it was noted that trials in which no events are observed do not contribute to a meta-analysis on a relative scale such as the odds ratio. They can only be included if a continuity correction factor is applied, and the inverse-variance weighted method is used. If a constant correction factor is used, such as 0.5 then the odds ratio produced is determined simply by the relative size of the two treatment groups. This is the reason a proportional continuity correction was used – this ensures an odds ratio of 1 is produced for every trial with 0 events, irrespective of the size of the two patient groups. A problem with adding continuity correction factors to such a sparse dataset is that they actually add a substantial amount of information, resulting in a narrowing of confidence intervals, which can be observed in Table 7.2.

In summary, Table 7.2 demonstrates that the particular method used, and the nature of any continuity correction used can make substantial differences to the overall effect size and corresponding confidence interval. In the following sections more advanced pooling methods are applied to the dataset and the results of these are compared to these simple approaches.

7.3.2 The use of Bayesian MCMC methods to pool odds ratios from sparse data

Bayesian models estimated using MCMC methods implemented in WinBUGS are considered in this section. Focus is restricted to the cardiovascular endpoint to keep the number of analyses to a manageable size. Fixed and random effect models are both implemented using both exact methods (making no distributional assumptions about the outcome from each study), and methods where the log odds ratios are assumed to be normally distributed. In all these models the intention is to use non-informative priors, as interest is focused on the different model specifications. However, after fitting a few preliminary models, it became apparent that the pooled estimate was often dependent on

the prior placed on the between study variance. Because of this, several models using the same formulation were fitted using a range of priors to investigate the robustness of each to their specifications.

Hence, convergence was assessed informally using graphical techniques such as examining long chains. The MCMC simulations from the posterior distributions of all these models were computed very quickly so several tens of thousands of iterations were used for both the burn in and the run on which parameters were estimated. Autocorrelation was minimal in all models. A brief exposition of the different models used follows.

The fixed normal model - The simplest model fixed effect (analogous to the inverse-variance weighted method) applying a continuity correction of 0.5 to all studies. This model is given by equation (2.5) and reproduced below.

$$T_i \sim N[\theta, \sigma^2{}_i] \qquad i = 1 \ldots . k$$
$$\theta \sim N[-,-], \tag{2.5}$$

where θ is the estimate for the underling effect size, and [-,-] indicates a prior distribution to be specified.

The fixed logit model - A modification to the fixed normal model, which removes the requirement that the log odds ratio from each study is normally distributed, is to model the proportions of persons who have events in both arms of each study directly using a binomial distribution and models the logits of these proportions. No continuity corrections factors are required when using these models. This model is given in equation (2.12) and reproduced below.

$$a_i \sim Bin[p_{1i}, (a_i + b_i)] \qquad c_i \sim Bin[p_{2i}, (c_i + d_i)] \qquad i = 1 \ldots . k$$

$$\text{logit}(p_{1i}) = \mu_i \qquad \text{logit}(p_{2i}) = \mu_i + d \qquad (2.12)$$

$$\mu_i \sim [-,-] \qquad d \sim [-,-]$$

$$OR = exp(d),$$

The fixed proportions model - A further variation models the proportions directly and constructing an odds ratio using the posterior distributions for the proportions. The model is given in (2.13) and is reproduced below.

$$a_i \sim Bin[p_1,(a_i + b_i)] \quad c_i \sim Bin[p_2,(c_i + d_i)] \quad i = 1.....k$$

$$p_1 \sim Beta[-,-] \qquad p_2 \sim Beta[-,-] \qquad (2.13)$$

$$OR = (p_1 \times (1 - p_2))/(p_2 \times (1 - p_1))$$

Where p_{1i} and p_{2i} are the probabilities of events in the two groups being compared for the *i*th study. μ_i is the estimated ln(odds) of an event in group one, and d is the ln(odds ratio) between groups; priors are required for these parameters. This model makes the added assumption that the odds in the treatment and control arms are identical to each other across studies; rather than their ratio. This assumption may not be justified if the baseline risk varies between studies. This model also allows all 23 studies to be included without the need for continuity correction factors.

The fixed Poisson model – This model is similar to the fixed logit model, but it assumes events within the trial arms are distributed according to a Poisson rather than binomial distribution. This may be more appropriate when events are rare. This model estimates the relative risk rather than the odds ratio.

$$a_i \sim Poisson(rate_{1i} \times (a_i + b_i)) \qquad c_i \sim Poisson(rate_{2i} \times (c_i + d_i))$$

$$\log(rate_{1i}) = \mu_i \qquad \log(rate_{2i}) = \mu_i + d \qquad i = 1.....k \qquad (7.1)$$

$$\mu_i \sim [-,-] \qquad d \sim [-,-]$$

$$RR = exp(d)$$

The random normal model – This model extends the fixed normal model by introducing a random effect for the between study variance and is described by equation (2.17) in Section 2.33 and reproduced below.

$$T_i \sim \mathrm{N}[\theta_i, \sigma^2_i] \quad \sigma^2_i \sim [-,-] \qquad i = 1,..,k$$

$$\theta_i \sim \mathrm{N}[\mu, \tau^2] \qquad (2.17)$$

$$\mu \sim [-,-] \qquad \tau^2 \sim [-,-].$$

The random logit model (priors on between study precision parameter) – This extends the fixed effect logit model, in much the same way as the normal model. This is exactly the same model as that used by Smith et al. (Smith et al. 1995b) described by equation (2.18) and reproduced below

$$a_i \sim Bin[p_{1i}, (a_i + b_i)] \qquad c_i \sim Bin[p_{2i}, (c_i + d_i)] \qquad i = 1.....k$$

$$\mathrm{logit}(p_{1i}) = \mu_i \qquad \mathrm{logit}(p_{2i}) = \mu_i + delta_i \qquad (2.18)$$

$$delta_i \sim \mathrm{N}[\phi, \tau^2]$$

$$\mu_i \sim [-,-] \qquad \phi \sim [-,-] \qquad \tau^2 \sim [-,-],$$

where ϕ represents the overall pooled effect, on a log odds ratio scale, and τ^2 is a measure of the between-study heterogeneity. A similar extension of (2.13) has also been implemented, (Byar, 1980) however this is not pursued here.

The random logit model (prior on between study standard deviation parameter) – This model is the same as above except for the prior placed on the between study variance term. Here it is placed on the standard error rather than the precision. This formulation was used previously by Thompson et al. (Thompson et al. 1997)

The random Poisson model – Another straightforward extension of the related fixed effect model into hierarchical model with a normally distributed random effect.

$$a_i \sim Poisson(rate_{1i} \times (a_i + b_i)) \qquad c_i \sim Poisspn(rate_{2i} \times (c_i + d_i))$$

$$\log(rate_{1i}) = \mu_i \qquad \log(rate_{2i}) = \mu_i + delta_i \qquad i = 1 \ldots . k \qquad (7.2)$$

$$\mu_i \sim (-,-) \qquad delta_i \sim Normal(\phi, \tau^2)$$

$$\phi \sim [-,-] \qquad \tau^2 \sim [-,-]$$

The random proportions model – Although a random effect extension to the fixed effect proportions model is possible it is not considered in Table 7.3. The extension is less straightforward than the others presented here because hyper-parameters are required for the two parameters of a beta distribution. Specifying priors for such parameters is not intuitive. It should be noted however that such a model does have similarities with the exact bivariate approach to meta-analysis described by VanHowleingen et al. (Van Houwelingen et al. 1993) A more natural bivariate random effect meta-analysis model is considered under the risk difference section below.

The results of using these models to combine the cardiovascular endpoint are presented in Table 7.3

Table 7.3 Bayesian MCMC meta-analysis models for cardiovascular endpoint using odds ratio scale

	Num of studies included	Prior for baseline rates	Prior for (ln(OR)) pooled	Prior for $1/\tau^2$	Estimate for OR (95% CrI)	Estimate for τ^2 (95% CrI)
Fixed normal model (CC = 0.5 all studies)						
	7	NA	$N(0.0,10^5)$	NA	1.236 (0.441 to 3.443)	NA
	23	NA	$N(0.0,10^5)$	NA	0.878 (0.430 to 1.789)	NA
Fixed exact Logit model		*Prior for baseline logit odds*	*Prior for pooled ln(OR)*			
	7	$N(0.0,10^5)$	$N(0.0,10^5)$	NA	1.828 (0.6265 to 6.093)	NA
	23	$N(0.0,10^5)$	$N(0.0,10^5)$	NA	1.809 (0.6116 to 6.039)	NA
Fixed proportions model		*Prior for proportion of controls*	*Prior for proportion on treatment*			
	7	Beta(1,1)	Beta(1,1)	NA	0.962 (0.372 to 2.810)	NA
	23	Beta(1,1)	Beta(1,1)	NA	0.904 (0.350 to 2.625)	NA
	7	Beta(0.5,0.5)	Beta(0.5,0.5)	NA	1.011 (0.381 to 3.104)	NA
	23	Beta(0.5,0.5)	Beta(0.5,0.5)	NA	0.952 (0.360 to 2.91)	NA
	7	Beta(0.01)	Beta(0.01)	NA	1.071 (0.391 to 3.481)	NA
	23	Beta(0.01)	Beta(0.01)	NA	1.008 (0.369 to 3.258)	NA
Fixed poisson model (estimating RR)		*Prior for logit baseline rates*	*Prior for pooled ln(OR)*			
	7	$N(0.0,10^5)$	$N(0.0,10^6)$	NA	1.777 (0.631 to 5.893)	NA
	23	$N(0.0,10^5)$	$N(0.0,10^6)$	NA	1.800 (0.628 to 5.972)	NA
Random normal model (CC = 0.5 to all studies)						
	7	NA	$N(0.0,10^6)$	G(0.001,0.001)	1.258 (0.415 to 3.872)	0.0453 (0.0008 to 2.684)
	23	NA	$N(0.0,10^6)$	G(0.001,0.001)	0.876 (0.414 to 1.832)	0.0206 (0.0008 to 0.643)

Random Logit model (prior on precision)		*Prior for logit baseline rates*				
	7	$N(0.0,10^5)$	$N(0.0,10^6)$	Gam(0.001,0.001)	2.596 (0.281 to 5802.2)	2.713 (0.002 to 300.329)
	7	$N(0.0,1.0^2)$	$N(0.0,10^2)$	Gam(0.01,0.01)	2.132 (0.331 to 44.124)	1.580 (0.014 to 45.846)
	7	N(0.0,10)	N(0.0,10)	Gam(0.1,0.1)	1.121 (0.299 to 4.39)	0.564 (0.052 to 8.283)
	7	$N(0.0,10^5)$	$N(0.0,10^6)$	Gam(1,1)	2.770 (0.428 to 68.306)	2.039 (0.312 to 29.59)
	7	$N(0.0,10^5)$	$N(0.0,10^6)$	Gam(3,1)	1.964 (0.544 to 8.602)	0.410 (0.145 to 1.869)
	7	$N(0.0,10^5)$	$N(0.0,10^6)$	Par(1,0.01)	17.94 (0.073 to 37049.1)	46.14 (3.95 to 97.08)
	7	$N(0.0,10^5)$	$N(0.0,10^6)$	Par(1,0.2)	2.779 (0.417 to 24.953)	2.911 (0.311 to 4.899)
	7	$N(0.0,10^5)$	$N(0.0,10^6)$	Par(0.5,0.1)	5.812 (0.125 to 22247.84)	18.26 (0.281 to 93.65)
	7	$N(0.0,10^5)$	$N(0.0,10^6)$	N(0,0.001)I(0,)	1.736 (0.572 to 8.602)	0.055 (0.014 to 4.143)
	7	$N(0.0,10^5)$	$N(0.0,10^6)$	N(0,0.1)I(0,)	2.167 (0.505 to 19.240)	0.589 (0.149 to 11.97)
	23	$N(0.0,10^5)$	$N(0.0,10^5)$	Gam(0.01,0.01)	3.246 (0.2258 to 666700)	0.0473 (0.000004 to 4975.9)
	23	$N(0.0,10^5)$	$N(0.0,10^5)$	Gam(0.1,0.1)	3.224 (0.233 to 13220.0)	4.787 (0.095 to 292.07)
	23	N(0.0,100)	N(0.0,100)	Gam(0.1,0.1)	1.805 (0.421 to 10.44)	0.958 (0.063 to 11.587)
	23	N(0.0,10)	N(0.0,10)	Gam(0.1,0.1)	1.139 (0.462 to 2.894)	0.319 (0.045 to 2.450)
Random Logit model (Prior on standard deviation)		*Prior for logit baseline rates*	*Prior for pooled ln(OR)*	*Prior for* τ		
	7	$N(0.0,10^5)$	$N(0.0,10^5)$	N(0,10)I(0,)	4.016 (0.301 to 666.0)	6.88 (0.051 to 48.4)
	7	$N(0.0,10^5)$	$N(0.0,10^5)$	N(0,1)I(0,)	2.259 (0.516 to 17.81)	0.855 (0.0026 to 6.119)
	7	$N(0.0,10^5)$	$N(0.0,10^5)$	N(0,0.1)I(0,)	1.884 (0.622 to 6.747)	0.055 (0.0001 to 0.567)
	23	$N(0.0,10^5)$	$N(0.0,10^5)$	N(0,10)I(0,)	3.117 (0.311 to 220.4)	4.913 (0.03 to 38.13)
	23	$N(0.0,10^5)$	$N(0.0,10^5)$	N(0,1)I(0,)	2.167 (0.500 to 15.65)	0.830 (0.002 to 6.172)
	23	$N(0.0,10^5)$	$N(0.0,10^5)$	N(0,0.1)I(0,)	1.834 (0.603 to 6.355)	0.056 (0.00012 to 0.583)
Random Poisson model				*Prior for* $1/\tau^2$		
	7	$N(0.0,10^5)$	$N(0.0,10^6)$	Gam(0.001,0.001)	3.007 (0.310 to 2676445.055)	3.010 (0.002 to 499.076)
	7	$N(0.0,10^2)$	$N(0.0,10^2)$	Gam(0.01,0.01)	3.084 (0.319 to 3442.661)	3.298 (0.013 to 189.063)
	7	N(0.0,10)	N(0.0,10)	Gam(0.1,0.1)	3.053 (0.618 to 43.164)	1.414 (0.068 to 38.950)

In a similar manner to the Classical methods, while statistical 'significance' conclusions have not changed depending on which model is used, the parameter estimates and confidence intervals vary considerably. The simple fixed effect model which assumes normality of the log odds ratios from each study and adds a continuity correction factor of 0.5 produces very different results from the fixed exact logit model, which highlights the need for exact calculations in instances such as these. These differences appear to propagate over into the random effect versions of these models. Since no exact random effect models exist for classical methods, the Bayesian derived MCMC models may have a distinct advantage in situations such as these.

One interesting aspect of this investigation is the fact that the posterior distributions for the pooled log odds ratio appear to be skewed for the random effect models. Figures 7.1a and b show two typical posterior distributions for the log odds ratios from a logit fixed effect and random effect model respectively, indicating the skewness in the random effect model is extreme. This asymmetry implies very high odds ratios are being sampled, and credible intervals calculated using asymptotic methods assuming normality may be erroneous (centiles of the samples from the posterior distribution are used to construct the credible intervals for each model in Table 7.3). This finding has important implications for the classical methods, since these confidence intervals are always based on asymptotic normality assumptions.

Table 7.2 indicates that these models are all sensitive to the prior distributions specified. Most critical appears to be the prior distribution for the between study variance parameter, τ^2, in the random effect models. It would appear that it is very difficult to specify a completely non-informative prior. Indeed, often when a prior with very small precision is specified, this places some probability mass on extremely high values of τ^2, in turn allowing extreme odds ratios to be sampled. Because there is very little information in the data to estimate τ^2, the mass on these extreme values is non-negligible, allowing the distributions for the log(OR) and τ^2 to be skewed by extreme values, and explaining the large right hand tail to the posterior odds ratio distribution. This phenomenon would appear to hold over several distributional forms when the prior is placed on $1/\tau^2$, or when a truncated normal is used for τ. When a more informative

prior is placed on τ^2, the OR estimate reduces, as the extreme values are sampled much less frequently. The problem is that it is difficult to specify a prior that does not allow sampling of extreme results, but is still reasonably non-informative. In the analysis of this dataset this issue appears particularly critical. Obviously, the priors specified do not need to be non-informative, for example Higgins and Whitehead (Higgins and Whitehead, 1996) consider empirical prior distributions for τ^2 using information from historical meta-analyses. Such an approach may be particularly valuable in situations such as this.

Figure 7.1a Histogram of sample from the posterior distribution of the pooled log odds ratio for the fixed effect logit model

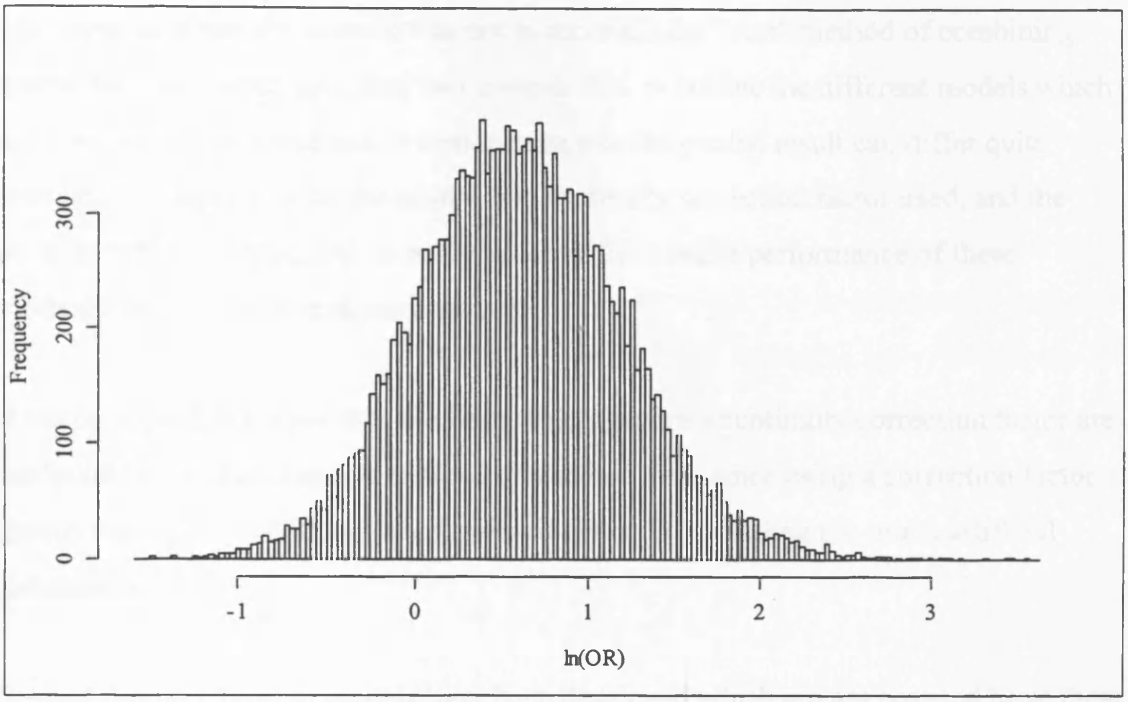
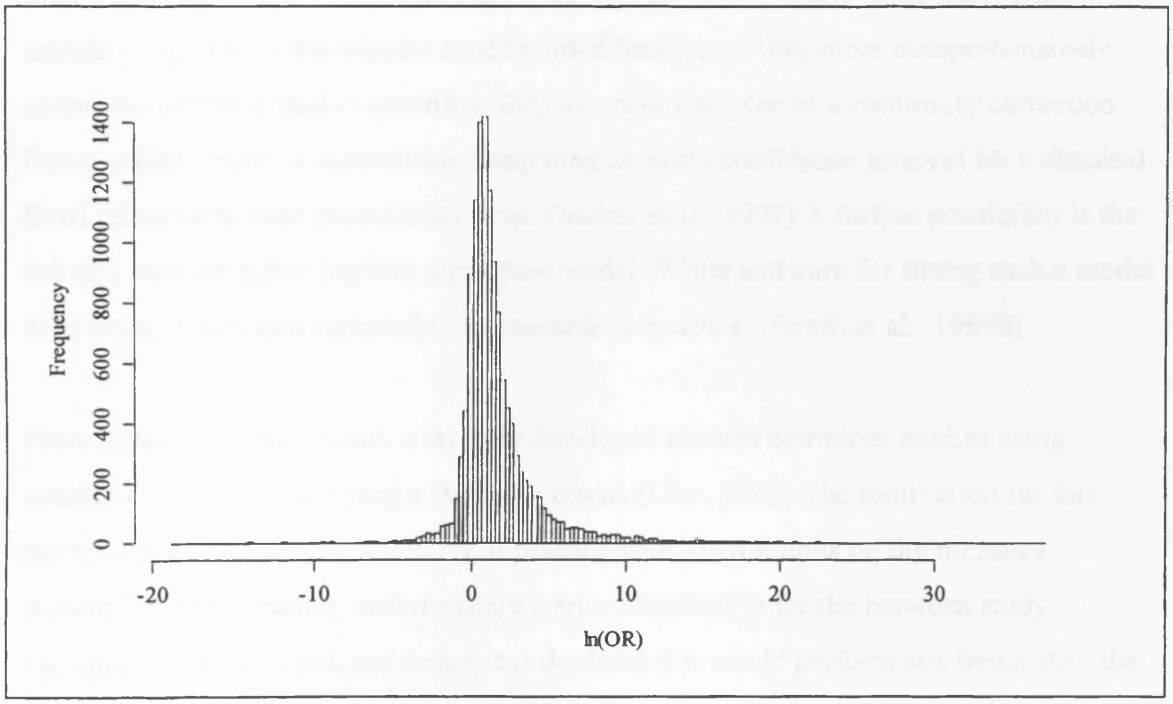


Figure 7.1b Histogram of sample from the posterior distribution of the pooled log odds ratio for the random effect logit model



7.3.3 Discussion of combining odds ratios for sparse data

The purpose of the above work was not to ascertain the "best" method of combining sparse data, but rather, as a first step towards this, to outline the different models which are theoretically possible and to demonstrate that the pooled result can differ quite dramatically depending on the model, any continuity correction factor used, and the prior distributions specified. In order to assess the relative performance of these methods, simulation studies are required.

It would appear, however, that methods which require a continuity correction factor are not suitable for data as sparse as those considered here, since using a correction factor greatly reduces the width of the confidence intervals (by adding too much artificial information).

Several further Classical models have been developed which are not covered here; these include methods for incorporating uncertainty in estimating τ^2 into the model, (Hardy and Thompson, 1996; Biggerstaff and Tweedie, 1997) and a bivariate random effect model (see Section 2.3.2). (Van Houwelingen et al. 1993) While these models are arguably superior to the simpler models fitted here, since they more comprehensively model the between study variability, they all require the use of a continuity correction factor. Additionally, a method for computing an exact confidence interval for a classical fixed effect odds ratio model does exist. (Mehta et al. 1997) A further possibility is the use of a random effect logistic regression model. While software for fitting such a model does exist, it has been reported to be unstable in practice. (Smith et al. 1995b)

From a Bayesian standpoint, a recently developed method combines studies using conditional likelihoods using a Bayesian model.(Liao, 1999) The motivation for this model was to remove the necessity of placing prior distributions on the nuisance parameters in the model; unfortunately a prior distribution for the between study variance is still required, and hence it is doubtful if it would perform any better than the Bayesian models described above.

An interesting point regarding studies with no events in either arm is that results are affected slightly by their inclusion in the Bayesian model. This appears to be in contrast to classical analyses where it has been stated that such trials add no information when combination takes place on the odds ratio scale. (Liao, 1999)

To overcome some of the problems encountered above, some novel methods are discussed below. Firstly, the use of data inflation to stabilise the Bayesian methods, by reducing the influence of prior distributions are examined. This is followed by sections exploring the effect of combining on the alternative risk difference scale, and the use of re-sampling methods to combine studies.

7.4 Data inflation methods

The previous section showed how sensitive the pooled estimate was to the prior distributions for the parameters, and particularly to that placed on the between study variance term. It thus appeared to be impossible to produce a consistent pooled estimate across any range of priors. A novel method which aims to overcome problems with prior distributions unintentionally influencing the estimation of variance parameters for hierarchical models estimated using MCMC methods, is known as data inflation. (Scurrah et al., 2000, Burton et al., 1999) The concept behind such a method is simple. The likelihood generated by the data is made n times as large, thus making the prior distribution less informative by a factor of n. Sampling then can proceed as normal, but the posterior parameters will be too precise due to the inflated amount of data fitted in the model. An adjustment is required to compensate for this. If distributions are symmetric, then inflating the standard deviations of parameters so obtained by a factor of $\sqrt{n}$ will provide correct 95% credible intervals. Alternatively, if inferences are being based on posterior distribution percentiles, and the posterior is symmetric, then the distance from the median to each centile can be multiplied by $\sqrt{n}$. A slightly more sophisticated approach is required if posterior distributions are not symmetric, and no transformations will make them symmetric. Under these circumstances the posterior distribution has to be raised to the power of $1/n$ and then made proper again (i.e.

dividing by a constant so it sums to one). This allows an approximate 95% credible interval to be estimated by examining the cumulative distribution function of this new density.

7.4.1 Initial application to meta-analysis of the effects of diuretics on pre-eclampsia

Since the application of data inflation in a meta-analysis setting is novel, it is first applied to a less extreme dataset, to illustrate how it works and examine its applicability for meta-analysis. The dataset chosen is a meta-analysis of nine trials of effects of diuretics on pre-eclampsia; (Collins et al. 1985) the data are provided in Table 7.4. This dataset is chosen for a number of reasons. Firstly, since it contains only nine trials, there will be considerable uncertainty in the estimation of the between study variance term in the meta-analysis model, and hence the prior placed on this parameter will be particularly influential. Secondly, this dataset has been used to illustrate classical methodology for incorporating the uncertainty in the estimation of the between study heterogeneity in the meta-analysis model, (Hardy and Thompson, 1996; Biggerstaff and Tweedie, 1997) and hence results for this dataset are available for comparison from several sophisticated methods.

Table 7.4 Meta-analysis of nine trials of effects of diuretics on pre-eclampsia

Study Number	Cases of pre-eclampsia/total number of patients		Odds ratio (95% CI)
	Treated	Control	
1	14/131	14/136	1.04 (0.48, 2.28)
2	21/385	17/134	0.40 (0.20, 0.78)
3	14/57	24/48	0.33 (0.14, 0.74)
4	6/38	18/40	0.23 (0.08, 0.67)
5	12/1011	35/760	0.25 (0.13, 0.48)
6	138/1370	175/1336	0.74 (0.59, 0.94)
7	15/506	20/524	0.77 (0.39, 1.52)
8	6/108	2/103	2.97 (0.59, 15.07)
9	65/153	40/102	1.14 (0.69, 1.91)

A random effect logit model, with the prior for the between study heterogeneity placed on the precision parameter (equation (2.18)) was fitted to this dataset. Normal(0,10^5) priors are placed on both individual trial baseline rates, and the pooled log odds ratio. A gamma prior was placed on the between study precision, and the effect of changing values given to its two parameters from 0.0001 to 1 was explored. Data were inflated by factors of three and ten.

Problems were encountered with the estimation of credible intervals when inflating the data by a factor of 10. The problem is illustrated diagrammatically in Figure 7.2. This figure displays the sampled posterior distribution (the circles) for the pooled log odds ratio, and the corrected distribution (line), using 8000 samples. The problem which this plot clearly shows is the poor estimation of the tails of the corrected distribution. This problem occurs because the tails of the corrected distribution correspond to the very extremes of the inflated distribution. Because these extremes are rarely sampled from, the tails of the corrected distribution are very poorly estimated. In this case, the poor estimation encroaches well into the central 0.95 probability mass, and since it is the 2.5 and 97.5 centiles of the distribution which we are trying to estimate, this creates problems. Several measures can be taken to improve this situation. Firstly, a larger sample will improve the evenness of coverage at the extremes of the distribution. However there are limits to the number of iterations which are feasible to work with, and how well the very extremes of the distribution are sampled from will depend on accuracy of the algorithm within WinBUGS, and the number of decimal places it uses in the internal calculations. In this instance, using a thinning factor of 5 (i.e. only keeping every 5th value of the MCMC chain generated to maximise the information obtained since the number of iterations used is limited by computer memory since this removes autocorrelation – the correlation between successive values sampled)
and increasing the number of used iterations to 36,000 did not greatly improve the situation.

Figure 7.2 Original sample distribution and distribution raised to the power 1/10 for log odds ratio. (data inflation factor 10 – sampled iterations 8000)

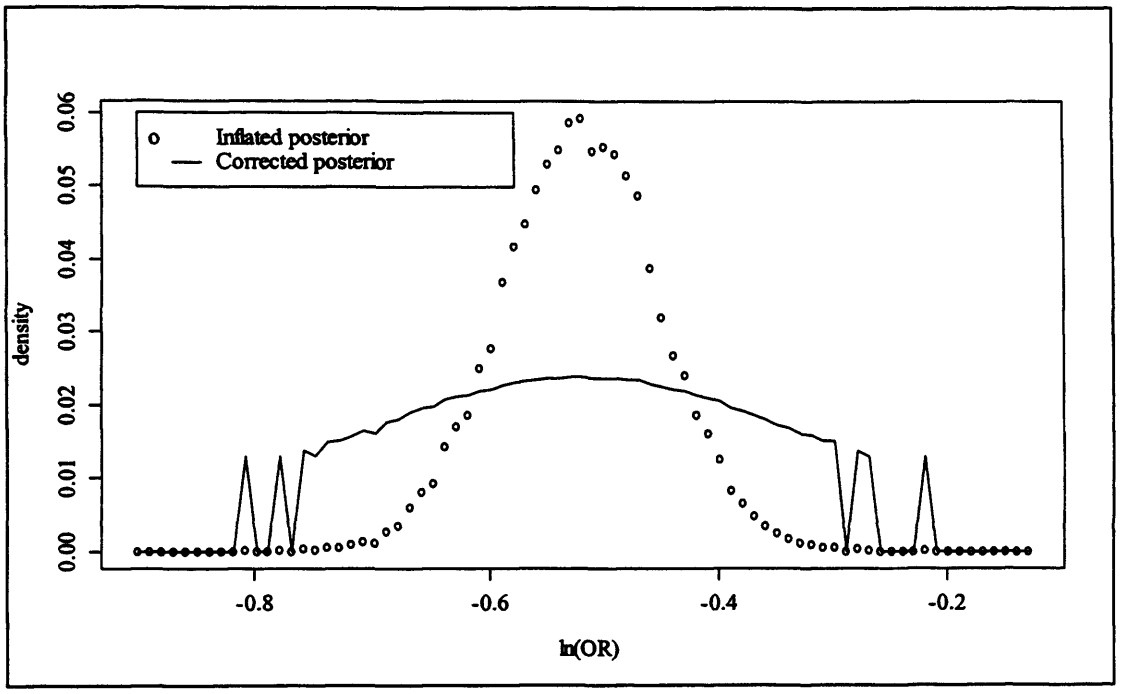
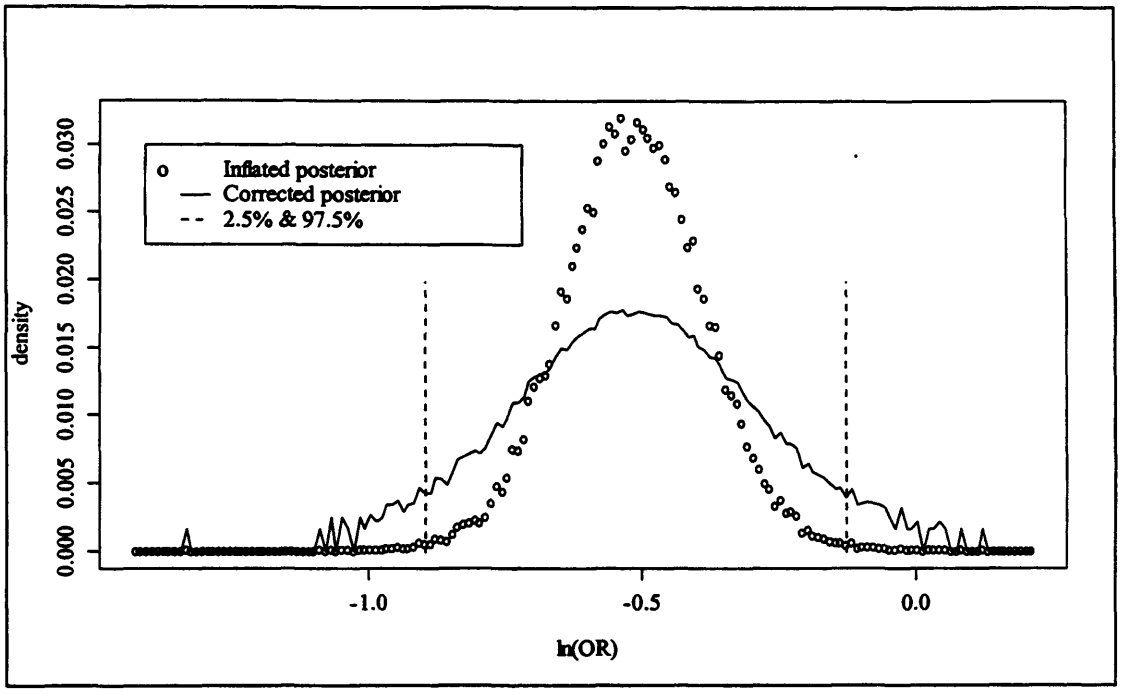


Figure 7.3 presents the same posterior plot as Figure 7.2, but based on data inflation factor of 3, and 36 000 iterations with a thinning factor of 25. It can be clearly seen that the 2.5 and 97.5 centiles of the distribution are estimated more satisfactorily here.

Figure 7.3 Original sample distribution and distribution raised to the power 1/3 for log odds ratio. (data inflation factor 3 – using 36000 iterations after thinning by a factor of 25)



If a larger inflation factors than this are required, and the number of iterations used has reached practical limits, but the tail areas of interests are still poorly estimated, approaches such as applying a smoothing algorithm to the distribution, or fitting parametric distributions to the data may help the estimation, but these are not pursued here. Results of these investigations using runs with burn in of 2000 and a further 36000 iterations, thinned by a factor of 5 from 180 000 iterations to calculate confidence intervals for each combination of prior distributions for an inflation factors of 3 and 10 (asymptotic only). are presented in Table 7.5.

Table 7.5 Results of data inflation on the posterior parameters for the meta-analysis of effects of diuretics on pre-eclampsia

Prior for logit baseline rates	Prior for (ln(OR)) pooled	Prior for $1/\tau^2$	Median estimate for OR (95% CrI)	Median estimate for τ^2 (95% CrI)	Mean estimate for OR (95% CrI)	Mean estimate for τ^2 (95% CrI)
			No Inflation – CrI's based on centiles of the posterior distribution		Inflation x 10 – CrI estimation by multiplying estimated standard errors by $\sqrt{10}$	
$N(0,10^5)$	$N(0,10^5)$	Gam(0.0001,0.0001)	0.601 (0.360 to 1.005)	0.316 (0.012 to 1.721)	0.595 (0.389 to 0.907)	0.274 (0.056 to 1.257)
$N(0,10^5)$	$N(0,10^5)$	Gam(0.001,0.001)	0.599 (0.360 to 1.014)	0.320 (0.027 to 1.751)	0.595 (0.389 to 0.907)	0.274 (0.056 to 1.258)
$N(0,10^5)$	$N(0,10^5)$	Gam(0.01,0.01)	0.599 (0.355 to 1.010)	0.336 (0.044 to 1.778)	0.595 (0.389 to 0.908)	0.274 (-0.056 to 1.270)
$N(0,10^5)$	$N(0,10^5)$	Gam(0.1,0.1)	0.598 (0.353 to 1.022)	0.381 (0.087 to 1.718)	0.596 (0.390 to 0.914)	0.276 (0.060 to 1.291)
$N(0,10^5)$	$N(0,10^5)$	Gam(1,1)	0.596 (0.327 to 1.107)	0.583 (0.219 to 1.890)	0.596 (0.377 to 0.937)	0.315 (0.076 to 1.299)

Prior for logit baseline rates	Prior for (ln(OR)) pooled	Prior for $1/\tau^2$	Mean estimate for OR (95% CrI)	Mean estimate for τ^2 (95% CrI)	Estimate for OR (95% CrI)	Estimate for τ^2 (95% CrI)
			Inflation x 3 – CrI estimation by multiplying estimated standard errors by $\sqrt{3}$		Inflation x 3 – CrI estimation by raising posterior to power of 1/3	
$N(0,10^5)$	$N(0,10^5)$	Gam(0.0001,0.0001)	0.597 (0.379 to 0.938)	0.285 (0.053 to 1.403)	0.597 (0.38 to 0.94)	0.285 (0.04 to 1.15)
$N(0,10^5)$	$N(0,10^5)$	Gam(0.001,0.001)	0.597 (0.381 to 0.943)	0.283 (0.053 to 1.406)	0.597 (0.38 to 0.95)	0.283 (0.04 to 1.22)
$N(0,10^5)$	$N(0,10^5)$	Gam(0.01,0.01)	0.596 (0.380 to 0.940)	0.290 (0.056 to 1.391)	0.596 (0.38 to 0.94)	0.290 (0.04 to 1.21)
$N(0,10^5)$	$N(0,10^5)$	Gam(0.1,0.1)	0.596 (0.374 to 0.937)	0.306 (0.067 to 1.329)	0.596 (0.38 to 0.93)	0.306 (0.06 to 1.17)
$N(0,10^5)$	$N(0,10^5)$	Gam(1,1)	0.595 (0.363 to 0.998)	0.400 (0.122 to 1.435)	0.595 (0.36 to 0.97)	0.400 (0.14 to 1.39)

Now it would appear for the random effect logit model that the data inflation method has succeed in making the pooled estimates robust to prior specifications. Table 7.6 compares the results obtained using the Bayesian random logit model, with and without using data inflation, and other classical methods of combining the data reported previously.

Table 7.6 Comparison of methods for combining trials of diuretics on pre-eclampsia

Method	Odds ratio (95% CI/CrI)	τ^2 (95% CI/CrI)
Fixed – inverse variance	0.67 (0.56 to 0.80)	-
Random – DerSimonian & Laird	0.60 (0.40 to 0.89)	0.23
Random – Hardy & Thompson (Hardy and Thompson, 1996)	0.60 (0.37 to 0.95)	0.24 (0.03 to 1.13)
Random – Biggerstaff & Tweedie (Biggerstaff and Tweedie, 1997)	0.62 (0.41 to 0.96)	0.23 (0.04 to 2.35)* (0.03 to 1.13)** [0 to 0.57)***
Random Bayesian**** – logit model	0.60 (0.36 to 1.01)	0.32 (0.03 to 1.75)
Random Bayesian**** – logit model data inflation ×3*****	0.60 (0.38 to 0.95)	0.28 (0.04 to 1.22)

* Method of moments based confidence interval

** Approximate likelihood ratio based confidence interval

*** Approximate maximum likelihood based confidence interval

**** Priors used $N(0.0,10^5)$ for Prior for logit baseline rates & pooled (ln(OR)). Prior for $1/\tau^2$ - Gam(0.001,0.001).

***** Credible intervals calculated by raising the posterior to power 1/3

All the methods broadly agree, but there are some interesting similarities and differences. Firstly, only the non-inflated Bayesian credible interval contains the value one. This slight increase in width over comparable classical methods could be due to the influence of informative prior distributions, because the data inflated credible interval falls back in line exactly with the classical methods which take the uncertainty in estimating τ^2 into account. Hence, it would appear that data inflation has removed the unintentional influence of the prior distribution placed on $1/\tau^2$. It should be stressed that the use of data inflation needs careful fuller validation in other meta-analysis context. For illustration, the method is applied to the oestrogen replacement therapy trials in the next section.

7.4.2 Applying data inflation to the oestrogen replacement therapy trials

Data inflation is now applied to the cardiovascular endpoint of the oestrogen replacement trials in an effort to see if improvements in the stability of the estimates over various intended vague prior distributions for the between study variance parameter can be achieved. The random effect logit model (equation 2.18), placing the prior on the precision was used.

Histograms of the samples from the posterior distributions for a preliminary run of 20000 iterations, using a data inflation factor of 10 are presented in Figure 7.4a and 7.4b. It can clearly be seen that both the posterior distributions for the log odds ratio and the log of the between study variance are considerably skewed. This implies that it will be necessary to raise the posterior distributions to a power of $1/n$, rather than inflating the estimated standard error to provide credible intervals. Having observed the problems with using large inflation factors when using this method, an inflation factor of three is used. Runs using an over-relaxed sampler, (Neal, 1998) taking a burn in of 4000, followed by a further 36 000 samples, thinned to 18 000 were used to produced the results in Table 7.7.

Figure 7.4a Histogram for samples from the posterior density for the pooled log odds ratio for the cardiovascular event endpoint for the oestrogen replacement studies

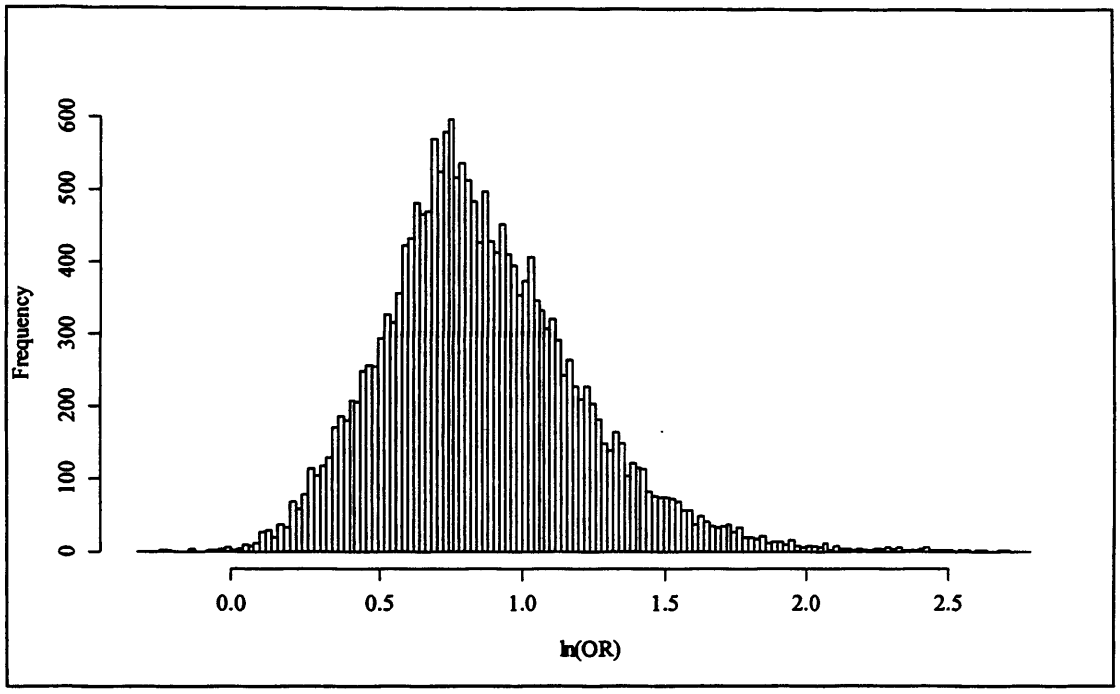


Figure 7.4b Histogram for samples from the posterior density for the log of the between study variance, for the cardiovascular event endpoint for the oestrogen replacement studies

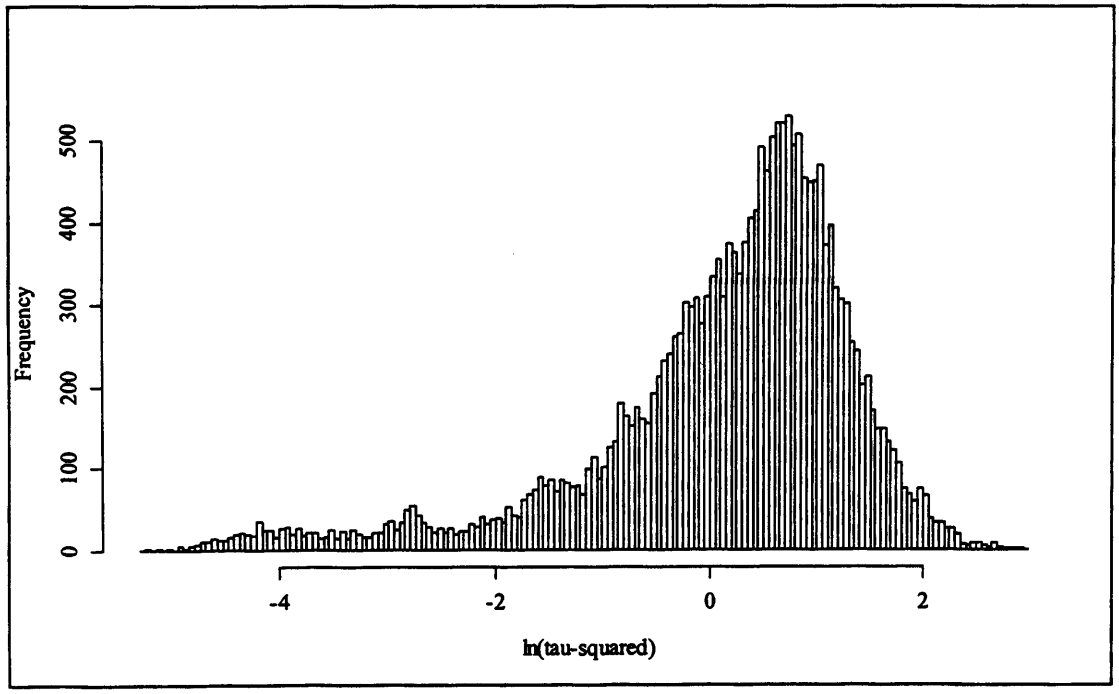


Table 7.7 Bayesian MCMC random effect meta-analysis using a data inflation factor of 3 for cardiovascular endpoint using odds ratio scale – random effect logit model

Number of studies included (x 3)	*Prior for logit baseline rates*	*Prior for (ln(OR)) pooled*	*Prior for $1/\tau^2$*	*Estimate for OR (95% CrI)*	*Estimate for τ^2 (95% CrI)*
23	$N(0,10^5)$	$N(0,10^5)$	Gam(0.001,0.001)	2.37 (0.47 to 154.2)	1.70 (0.00 to 33.95)
23	$N(0,10^5)$	$N(0,10^5)$	Gam(0.01,0.01)	2.42 (0.44 to 576.51)	1.81 (0.01 to 53.68)
23	$N(0,10^5)$	$N(0,10^5)$	Gam(0.1,0.1)	2.56 (0.45 to 435.28)	2.03 (0.03 to 66.02)
23	$N(0,10^5)$	$N(0,10^5)$	Gam(1,1)	2.48 (0.48 to 62.74)	1.60 (0.19 to 26.74)

Table 7.7 indicates although there is still considerable variation between estimates when different priors are specified for $1/\tau^2$, the estimates are more stable than those using no data inflation factor, reported in Table 7.3. Although there is a lot of variation in the upper limit of the credible interval for the pooled odds ratio, all results consistently suggest that the risk of cardiovascular events could be very high indeed. These confidence intervals are much wider than any produced by classical methods, suggesting such methods may seriously underestimate the uncertainty in the data.

Again, it would appear that data inflation is a potentially useful tool for removing the influence of priors, and in particular the prior placed on the between study variance in random effects meta-analysis model. This will be particularly valuable when there is very little information in the meta-analysis, for example when events are rare and/or there are very few studies.

A final point of interest is that in their original paper (Hemminki and McPherson, 1997) Hemminki and McPherson calculated the probability of obtaining their pooled odds ratio for cardiovascular events, when the true odds ratio is 0.7 or 0.5. This is another way of asking, 'what is the probability that oestrogen replacement therapy has moderate to large protective effects given the data?', which is necessarily cumbersome to enable it to be formulated using classical statistics. Such questions are much more natural to formulate under a Bayesian framework. The equivalent question, framed from a

Bayesian standpoint is, 'what is the probability the true odds ratio is less than 0.7 or 0.5', and this can be answered directly from the posterior distribution for the pooled effect size. Using the results from Table 7.7 above, where the prior on $1/\tau^2$ is Gamma(0.001,0.001), the probability that the odds ratio is less than 0.7 is 0.05, and the probability the odds ratio is less than 0.5 is 0.02. These are broadly comparable to the classical estimates in the paper of a probability of 0.10, and 0.03.

7.5 Combining postmenopausal hormone therapy adverse event data on the risk difference scale

7.5.1 Results using classical meta-analysis methods

The choice of methods to combine binary data using the risk difference scale is somewhat more limited than the array of methods for combining odds ratios because no methods specific to this scale have been developed. Results of using standard meta-analysis methods, using the various continuity correction factors described in Section 7.3.1, are reported in Table 7.8.

Table 7.8 Results of pooling risk differences (per 1000 women) for Cardiovascular and Breast cancer risk using classical approaches

	Cardiovascular risk			Breast cancer risk		
Risk Difference (per 1000 women) pool method	*Number of studies included*	*Pooled estimate (95% confidence or credibility interval)*	*Between study variance (95% confidence or credibility interval)*	*Number of studies included*	*Pooled estimate (95% confidence or credibility interval)*	*Between study variance (95% confidence or credibility interval)*
FIXED EFFECT ANALYSES						
Inverse variance (continuity correction 0.5 when required for s.e.)	23	2.441 (-4.452 to 9.334)	NA	23	2.537 (-4.182 to 9.257)	NA
Mantel-Haenszel (continuity correction 0.5 when required for s.e.)	23	6.919 (-2.307 to 16.146)	NA	23	-0.691 (-9.261 to 7.878)	NA
RANDOM EFFECT ANALYSIS						
Inverse variance or M-H estimate of Q	23	2.441 (-4.452 to 9.334)	0 (NA)	23	2.537 (-4.182 to 9.257)	0 (NA)

Heterogeneity was not statistically significant for either outcome, and hence the between study variance was estimated as zero in both instances, which means the random effect and the inverse variance weighted estimates are identical. Although confidence intervals all include zero, there are quite large discrepancies between the two methods of estimation. These results are compared to those derived using Bayesian methods below.

7.5.2 Results using Bayesian meta-analysis methods

As for the odds ratio models, all methods were implemented in WinBUGS with the intention of using non-informative priors. Focus was restricted to the breast cancer endpoint in this section. Again, the priors placed on the between study heterogeneity term is given considerable attention. Generation of the MCMC chains was very quick for all models considered, so several tens of thousands of iterations were used for both the burn in and the run on which parameters were estimated. Autocorrelation was also minimal in all models. A brief exposition of the different models used follows.

The fixed normal model – This model is given in equation (2.5). Continuity corrections are applied to calculate standard errors, where required.

The fixed proportions model – This model is essentially the same as that of equation (2.12), the only difference is that the risk different ($p_1 - p_2$) is constructed on the last line instead of the odds ratio.

The random normal model – This model is given in equation (2.17). Versions putting the prior on the precision and on the standard deviation of the between study variance term are both explored.

The random logit bivariate exact model – This model differs from the random logit model because dependent random effects are specified for both the baseline and the treatment effects. Such a model has been suggested previously by VanHowleingen et

al. (Van Houwelingen et al. 1993); a recently described Bayesian formulation using WinBUGS removes the need for a continuity correction. (Carlin, 2000) This model is somewhat more complex than those previously described because a multivariate distribution is specified for the random effects. A prior with a Wishart distribution is a natural choice for such random effects. Such priors are not intuitive to work with, however, and two variants removing the requirement of a Wishart prior, by allowing priors to be placed on the precision and standard deviation parameters respectively were also investigated. The model using the Wishart prior is set out below.

$$a_i \sim Bin[p_{1i},(a_i+b_i)] \qquad c_i \sim Bin[p_{2i},(c_i+d_i)] \qquad i=1.....k \qquad j=1,2$$

$$\text{logit}(p_{1i}) = \mu_i - delta_i/2 \qquad \text{logit}(p_{2i}) = \mu_i + delta_i/2$$

$$\begin{bmatrix}\mu_i \\ d_i\end{bmatrix} \sim Multivariate Normal\left(\begin{bmatrix}\theta_1 \\ \theta_2\end{bmatrix},\begin{bmatrix}\omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22}\end{bmatrix}\right) \qquad (7.3)$$

$$RD = e^{(\theta_1+\theta_2/2)}/\left(1+e^{(\theta_1+\theta_2/2)}\right) - e^{(\theta_1-\theta_2/2)}/\left(1+e^{(\theta_1-\theta_2/2)}\right)$$

$$\begin{bmatrix}\omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22}\end{bmatrix} \sim Wishart\begin{bmatrix}- & - \\ - & -\end{bmatrix}$$

Re-parameterising the model using a product normal formulation to remove the need for a Wishart prior is described elsewhere (*http://www.mrc-bsu.cam.ac.uk/bugs/faqs/modelling.shtml*).

It is not possible to construct models analogous to the fixed or random logit models for the risk difference outcome. Poisson, and random proportions models are possible, but are not pursued here.

The results of fitting these models with various prior distribution combinations are presented in Table 7.9

Table 7.9 Results of pooling risk differences (per 1000 women) for Breast cancer risk using Bayesian methods

	Number of studies included	*Prior for (RD)*	*Prior for $1/\tau^2$*	*Estimate for RD (per 1000 women) (95% CrI)*	*Estimate for τ^2 (95% CrI)*
Fixed Normal Model (CC* for SE where required)					
	23	N(0,0.1)	NA	2.525 (-4.19 to 9.25)	NA
Fixed Proportions Model		Prior for control/trt proportion			
	23	Beta(0.01,0.01)	NA	-1.763 (-8.333 to 3.474)	NA
Random Normal model prior on precision (CC* for SE where required)					
	23	N(0,1)	Gamma(1,1)	-3.372 (-128.7 to 121.5)	0.087 (0.052 to 0.164)
	23	N(0,0.1)	Gamma(1,1)	-3.383 (-129.0 to 121.7)	0.087 (0.052 to 0.164)
	23	N(0,0.01)	Gamma(1,1)	-3.385 (-129.1 to 121.7)	0.087 (0.052 to 0.164)
	23	N(0,0.1)	Gamma(10^{-6},10^{-6})	2.023 (-5.608 to 9.495)	6.97×10^{-6} (5.8×10^{-7} to 1.18×10^{-4})
	23	N(0,0.1)	Gamma(10^{-5},10^{-5})	1.664 (-6.758 to 9.714)	2.41×10^{-5} (4.22×10^{-6} to 1.82×10^{-4})
	23	N(0,0.1)	Gamma(10^{-4},10^{-4})	0.763 (-9.164 to 10.38)	8.65×10^{-5} (2.63×10^{-5} to 3.43×10^{-4})
	23	N(0,0.1)	Gamma(0.001,0.001)	-0.466 (-13.51 to 12.3)	3.42×10^{-4} (1.5×10^{-4} to 8.72×10^{-4})
	23	N(0,0.1)	Gamma(0.01,0.01)	-1.622 (-22.34 to 18.92)	0.00160 (8.60×10^{-4} to 0.0033)
	23	N(0,0.1)	Gamma(0.1,0.1)	-2.502 (-47.97 to 43.33)	0.010 (0.0060 to 0.0204)
	23	N(0,0.1)	Gamma(1,1)	-2.63 (-127.5 to 123.9)	0.087 (0.051 to 0.164)

Random Normal model – prior on s.d. (CC* for SE where required)			*Prior for τ*		
	23	N(0,1)	N(0,1)I(0,)	1.912 (-6.627 to 9.625)	1.32×10^{-5} (2.62×10^{-8} to 2.04×10^{-4})
	23	N(0,0.1)	N(0,1)I(0,)	1.912 (-6.627 to 9.626)	1.32×10^{-5} (2.62×10^{-8} to 2.04×10^{-4})
	23	N(0,0.001)	N(0,1)I(0,)	1.912 (-6.627 to 9.626)	1.32×10^{-5} (2.62×10^{-8} to 2.04×10^{-4})
	23	N(0,1)	N(0,0.001)I(0,)	1.912 (-6.627 to 9.626)	1.32×10^{-5} (2.62×10^{-8} to 2.04×10^{-4})
	23	N(0,1)	N(0,0.01)I(0,)	1.912 (-6.627 to 9.626)	1.32×10^{-5} (2.62×10^{-8} to 2.04×10^{-4})
	23	N(0,1)	N(0,0.1)I(0,)	1.912) (-6.627 to 9.626)	1.32×10^{-5} (2.62×10^{-8} to 2.04×10^{-4})
	23	N(0,1)	N(0,10)I(0,)	1.915 (-6.616 to 9.613)	1.32×10^{-5} (2.59×10^{-5} to 2.02×10^{-4})
	23	N(0,1)	N(0,100)I(0,)	1.919 (-6.617 to 9.605)	1.32×10^{-5} (2.59×10^{-8} to 2.01×10^{-4})
Bivariate logit (Wishart prior)		*Priors for log odds in the trt and contrl groups*	*Multivariate prior for between study variation ($1/\tau^2$) in trt and contrl groups*		*Covariance matrix for τ^2*
	23	$N(0,10^6)$	$Wishart\left(\begin{pmatrix}0.1 & 0\\ 0 & 0.1\end{pmatrix},2\right)$	1.050 (-4.604 to 7.123)	0.125 (0.018 – 1.167) 0.025 (–0.383 – 0.870) 0.172 (0.019 – 2.222)
	23	$N(0,10^4)$	$Wishart\left(\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix},2\right)$	0.261 (-5.217 to 7.000)	0.538 (0.140 – 2.581) 0.125 (–0.933 – 1.973) 0.842 (0.168 – 4.840)
	23	$N(0,10^2)$	$Wishart\left(\begin{pmatrix}10 & 0\\ 0 & 10\end{pmatrix},2\right)$	0.260 (-4.239 to 6.056)	2.369 (0.911 – 8.608) 0.160 (–3.721 – 4.096) 3.640 (1.198 – 13.35)
	23	$N(0,10^2)$	$Wishart\left(\begin{pmatrix}0.1 & 0\\ 0 & 0.1\end{pmatrix},2\right)$	0.779 (-4.433 to 7.622)	0.121 (0.018 – 1.102) 0.026 (–0.375 – 0.764) 0.160 (0.019 – 2.136)
Bivariate logit (Priors placed on precisions)		*Prior placed on 3 parameters relating to log odds*	*Prior for between study variation in treatment and control groups (placed on precision)*		
	23	$N(0,10^6)$	Gamma(0.0001,0.0001)	2.038 (-4.091 to 9.257)	0.001 (0.000 – 0.138) 0.012 (–0.080 – 0.291) 0.658 (0.015 – 4.367)

	23	$N(0,10^4)$	Gamma(0.001,0.001)	1.794 (-4.003 to 8.747)	$\begin{pmatrix} 0.009 & -0.013 \\ (0.001-0.310) & (-0.382-0.335) \\ & 1.034 \\ & (0.022-6.273) \end{pmatrix}$
	23	$N(0,10^2)$	Gamma(0.01,0.01)	0.716 (-5.656 to 8.709)	$\begin{pmatrix} 0.187 & 0.147 \\ (0.036-1.801) & (-0.872-1.506) \\ & 1.771 \\ & (0.192-7.832) \end{pmatrix}$
	23	N(0,10)	Gamma(0.1,0.1)	0.233 (-5.584 to 8.224)	$\begin{pmatrix} 0.617 & 0.211 \\ (0.201-2.382) & (-1.155-1.751) \\ & 1.631 \\ & (0.387-6.369) \end{pmatrix}$
Bivariate logit (Priors placed on standard deviations)			*Prior for between study variation in treatment and control groups (placed on standard deviation)*		
	23	$N(0,10^6)$	N(0,0.0001)I(0,)	1.908 (-4.338 to 9.557)	$\begin{pmatrix} 0.005 & -0.003 \\ (0.000-1.003) & (-0.643-0.673) \\ & 2.02 \\ & (0.107-0.233) \end{pmatrix}$
	23	$N(0,10^4)$	N(0,0.001)I(0,)	1.376 (-4.986 to 9.050)	$\begin{pmatrix} 0.026 & 0.018 \\ (0.000-1.646) & (-0.607-1.344) \\ & 1.958 \\ & (0.102-9.948) \end{pmatrix}$
	23	$N(0,10^2)$	N(0,0.1)I(0,)	1.180 (-5.150 to 9.048)	$\begin{pmatrix} 0.103 & 0.011 \\ (0.000-2.27) & (-0.921-1.517) \\ & 1.542 \\ & (0.058-8.118) \end{pmatrix}$
	23	N(0,10)	N(0,1.0)I(0,)	1.100 (-5.209 to 8.682)	$\begin{pmatrix} 0.153 & 0.012 \\ (0.000-1.476) & (-0.506-1.027) \\ & 0.710 \\ & (0.018-3.838) \end{pmatrix}$

* CC – Continuity correction factor

There is a lot of variation between the results in the table. The first striking feature is the gross difference in the results from the two fixed effect methods. Great variation is also observed in the random effect models, but there is less variation in pooled estimates to the specification of the prior for the fixed normal model which placed a prior over the standard deviation scale for the between study variance term. These results appear much more stable than those which placed Gamma priors on the variance scale. Although the more complex bivariate models fitted and could be updated in WinBUGS without problem, results were not stable. Interestingly, reducing the bivariate variance model to a function of two univariate parameters, and placing priors on the standard deviation scale did appear to improve the stability considerably.

7.5.3 Discussion of combining sparse data on the risk difference scale

Great variation between estimators is evident for both classical and Bayesian methods. In a similar fashion to that observed for the odds ratio scale, Bayesian methods are hampered by the difficulty of placing non-informative priors, particularly on the variance parameters. Unlike the odds ratio scale, no classical exact methods exist for either fixed or random effect methods. The Bayesian bivariate random effect model described by Carlin does use "exact" simulation methods, and has the added advantage that one can perform the modelling on the most appropriate scale from a statistical point of view, and then generate inferences for whatever derived parameters might be desired, by simulation. (Carlin, 2000) This does remove the objections from a statistical point of view of using the risk difference scale for meta-analysis. Unfortunately, due to the stability problems above it cannot be recommended for routine use with datasets containing sparse data. It would be possible to use data inflation in conjunction with these models for estimating the risk difference. This is not pursued here; problems with tail area estimation might still remain. Another potentially promising approach is the use of re-sampling methods for estimating the pooled treatment effect on the risk difference scale. This is considered in the next section.

7.6 Application of re-sampling methods to sparse event meta-analysis

Assumptions made by many of the methods of combining studies have been considered throughout this chapter. It has been stressed that part of the theoretical appeal of Bayesian methods is that they require fewer modelling assumptions. However, it has been shown in Sections 7.3.2 and 7.5.2 that there are sometimes problems with the stability of these models, particularly those incorporating one or more random effect term. This section considers the use of re-sampling methods in a Classical framework as an alternative method for constructing confidence intervals for the pooled effect size which requires fewer assumptions than more standard approaches.

7.6.1 Applying Bootstrap methods to meta-analysis

An alternative Classical approach which avoids the need for the assumption that the pooled effect size is normally distributed is the use of the Bootstrap re-sampling method. (Efron and Tibshirani 1993) To the author's knowledge such a technique has only been used twice before in a meta-analysis context. The only application to a medical meta-analysis dataset was described by Smith et al (Smith et al. 1995a), where it had been used to estimate the standard errors of a regression slope coefficient from individual studies to be combined in a meta-analysis (coincidentally) investigating the risk of breast cancer risk and duration of oestrogen use. (Steinberg et al. 1994) The method was compared to standard parametric estimates, and results obtained were very similar. Smith et al. (Smith et al. 1995a) comment that bootstrap methods are an attractive alternative to parametric methods, especially in evaluating heterogeneity by examining the histogram of sampled means, but they need further investigation for application in meta-analysis.

Adams et al. (Adams et al. 1997) use re-sampling methods on ecological datasets combining standardised mean differences. Bootstrapped estimates for fixed and random effect models were compared to standard methods. In addition, re-sampling methods were used to obtain significance levels for a randomisation test, which is analogous to the standard heterogeneity test. The results of this previous work (Adams et al. 1997)

indicated that, in the examples examined, the bootstrap confidence intervals were generally wider for the fixed effect model, but in broader agreement under the random effect model. Tests for heterogeneity were in broad agreement, but the bootstrap result was more conservative in a small proportion of cases. The motivation for the use of these methods by Adams et al. (Adams et al. 1997) was a concern that distributional assumptions of standard meta-analysis methods were not valid for sparse data. It should be noted, however, that the datasets they examined considered continuous outcomes only, and hence the issue of sparse binary data was not addressed.

7.6.2 Pooled estimates, confidence intervals, and heterogeneity statistics using the Bootstrap

The premise behind the Bootstrapped confidence intervals is simple, but quite computer intensive; (Efron and Tibshirani 1993) it is outlined here as it is applied to the postmenopausal hormone therapy trials. The 23 studies to be meta-analysed are sampled with replacement to produce a new dataset including 23 studies (i.e. each original study may appear once, multiple times, or not at all). The statistics of interest - the fixed and random effect pooled estimates and the Q statistic - are calculated from this new dataset. This procedure is replicated several thousand times, and the results from each replication are combined to form a dataset from which inferences about the statistics of interest can be made. For instance, calculating the proportion of times the Q statistic was more extreme than the one calculated for the original dataset produces a p-value which is not dependent on the test statistic being approximately chi-square distributed. Confidence interval estimates for the pooled estimates can also be constructed using the distribution of effect sizes from the sampled data. Three different estimates are often calculated, namely: 1) those based on normality assumptions but using the bootstrap estimate of standard error, 2) the percentile confidence interval, where the $\alpha/2$ and $100-\alpha/2$ centiles of the sampled distribution are calculated; and 3) the bias-corrected confidence interval - a modification on 2) which adjusts for the fact that the bootstrap distribution may not be centered around the observed estimate. The latter has been recommended for general use. (Efron and Tibshirani 1993)

Bootstrap homogeneity statistics for the postmenopausal hormone therapy trial meta-analysis based on risk differences

Classical estimates of Q for cardiovascular and breast cancer mortality on the risk difference scale (inverse variance method) are 10.85 (p=0.977) and 7.67 (p=0.998) respectively. When 2000 bootstrap replications of these test statistics were calculated, a more extreme result was observed 738 times for cardiovascular mortality and 833 times for breast cancer mortality. Hence p-values for these re-sampling tests are 0.37 (p = 738/2000) and 0.42 (p = 833/2000) for cardiovascular mortality and breast cancer mortality respectively. Although still non-significant, these are considerably lower than the standard estimates. Histograms of the distributions of the samples for the values of Q for CHD and breast cancer mortality are displayed in Figures 7.5 and 7.6 respectively. Certainly the distribution of Q for the breast cancer example looks unusual, and it is bi-modal. Chi-squared distributions on 22 degrees of freedom are superimposed over the histograms, clearly indicating the sample distributions are shifted along the x-axes from the theoretical test statistic distributions, although the distributional shapes are quite similar.

Figure 7.5 Distribution of the Q statistic from 2000 bootstrap samples for coronary heart disease outcome

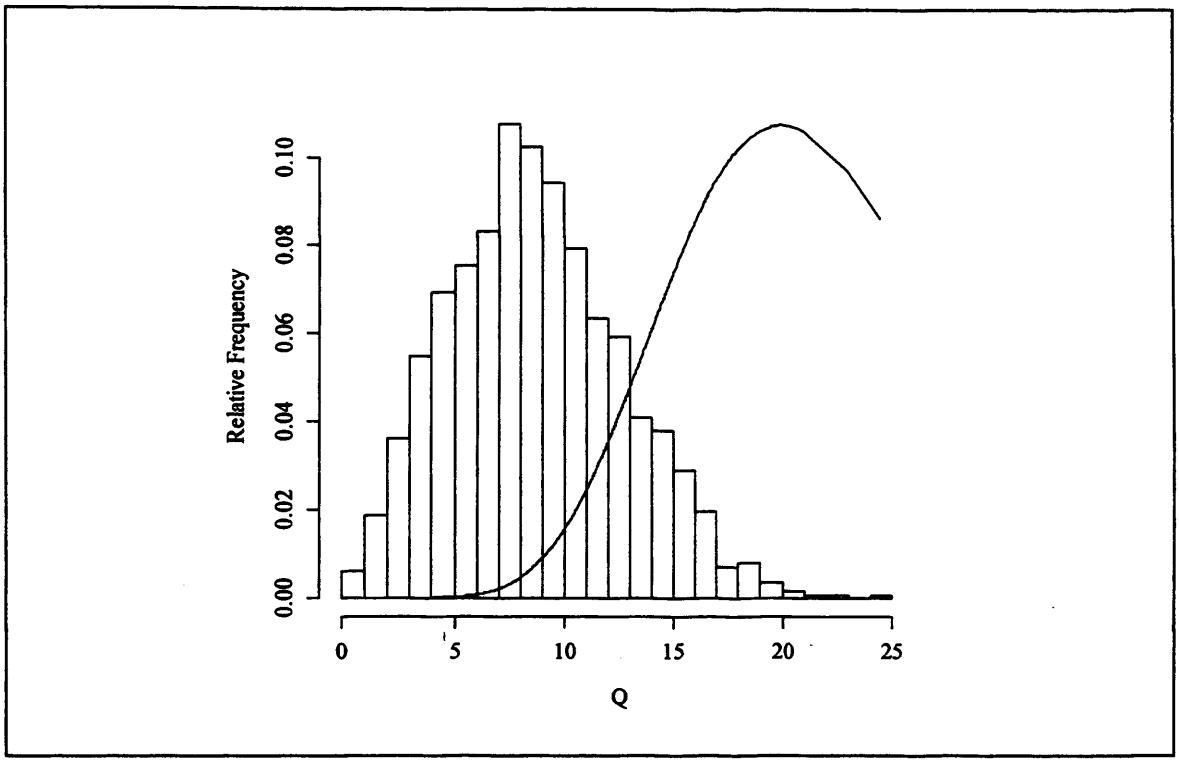
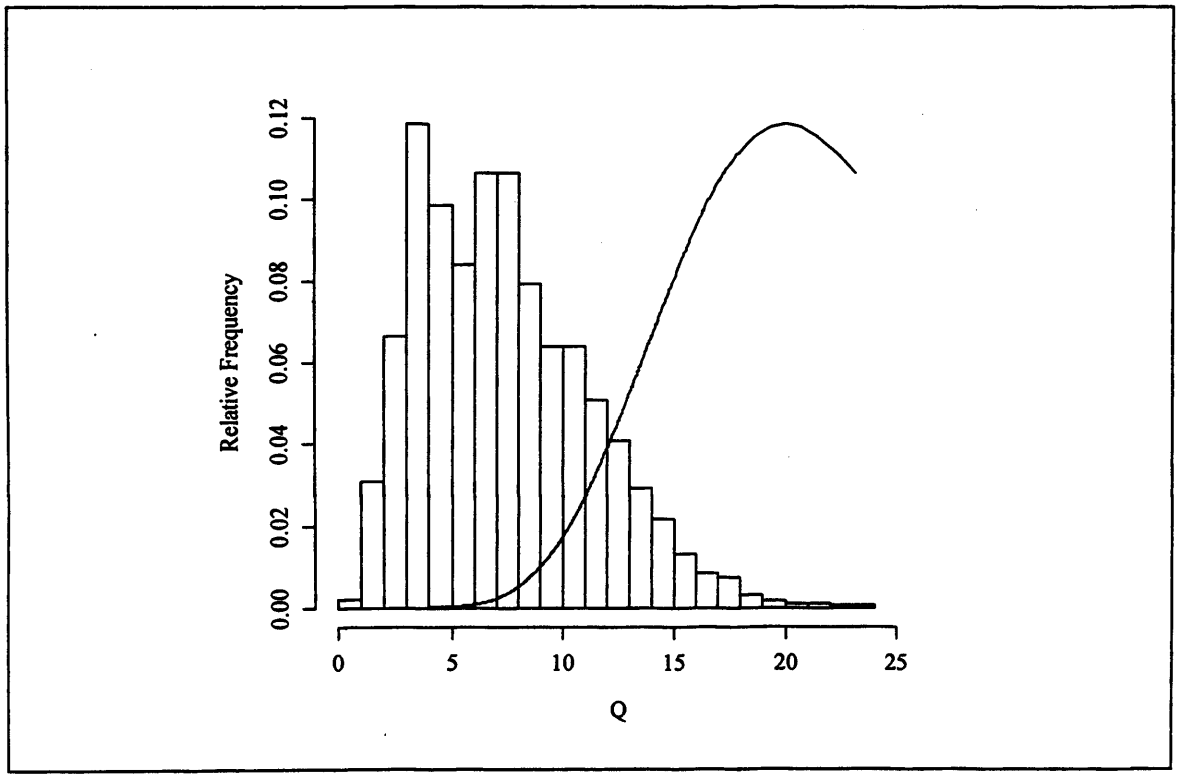


Figure 7.6 Distribution of the Q statistic from 2000 bootstrap samples for breast cancer outcome



Calculating bootstrap confidence intervals for the risk difference outcome measure

Longer bootstrap runs of 5000 samples were used to estimate confidence intervals for the pooled treatment difference, established from repeated runs using increasing numbers of samples until the estimates stabilised. Since τ^2 was estimated to be zero for both outcomes, the fixed and random estimates are identical. The confidence intervals produced by the three bootstrap methods are displayed in Table 7.10, where the classical fixed effect and a Bayesian random effect model are also included for comparison.

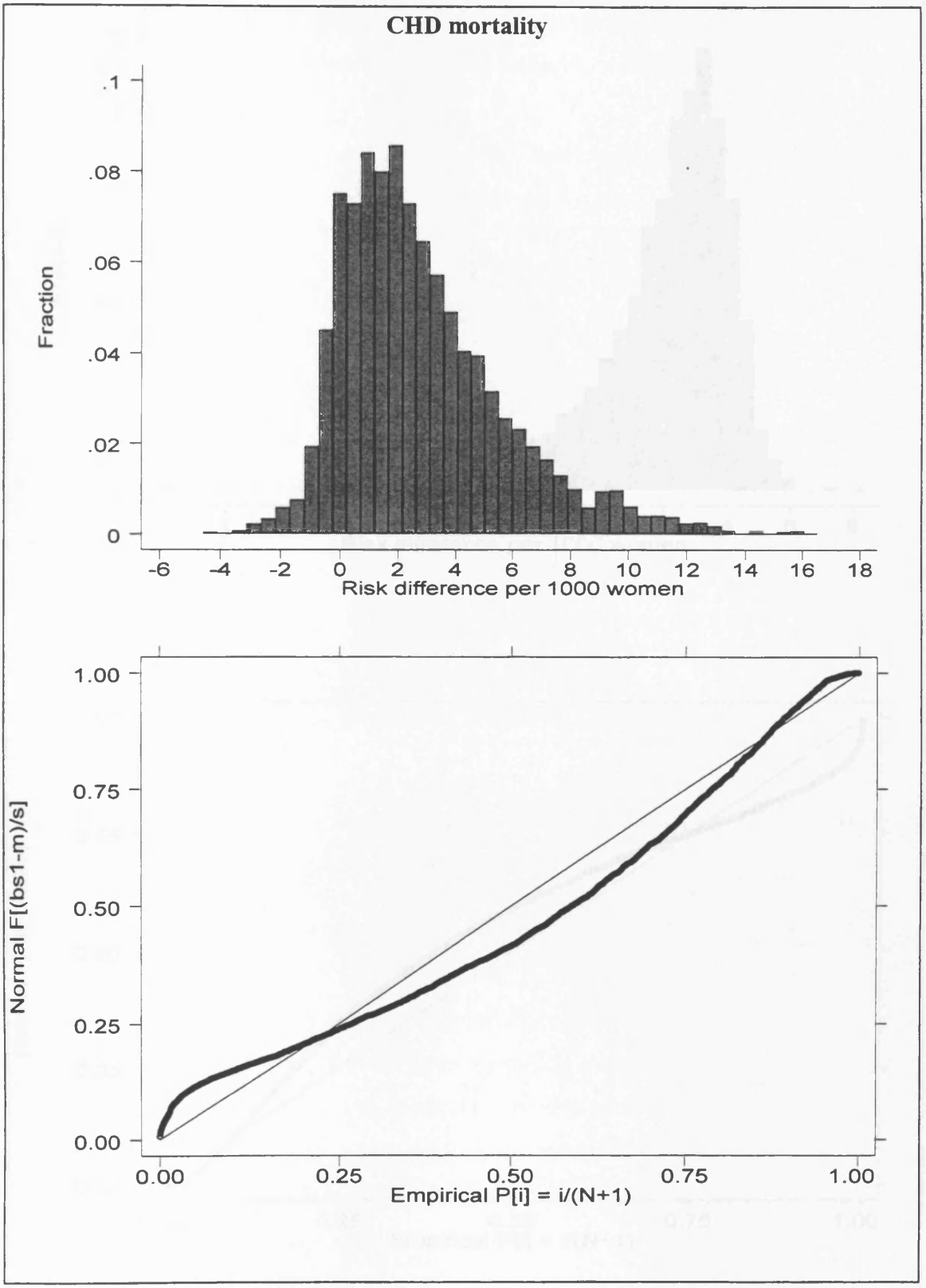
Table 7.10 95% Confidence/credible intervals for the pooled risk difference per 1000 women using different methods of estimation

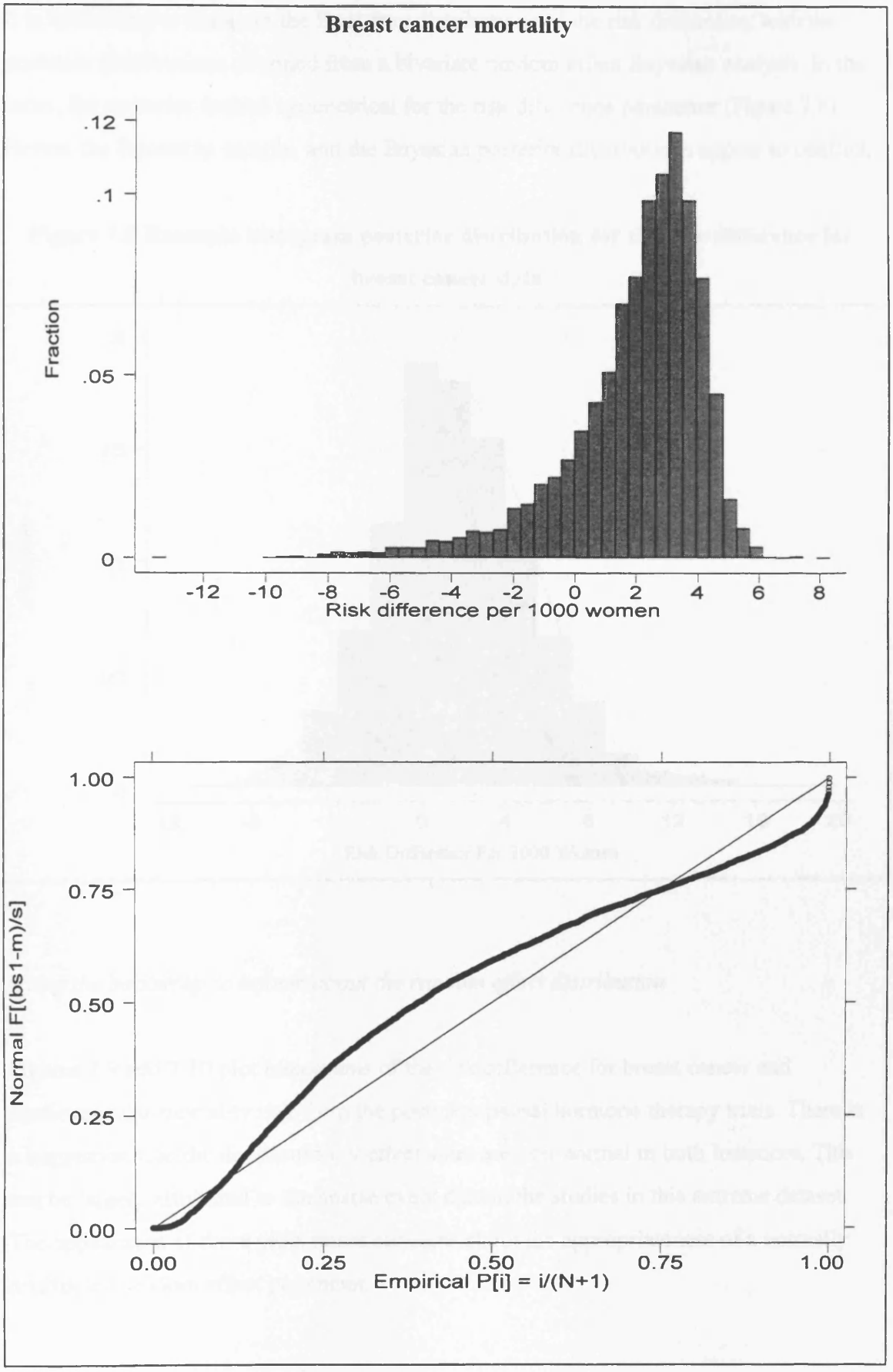
Method	CHD mortality		Breast cancer mortality	
	Interval	Interval width	Interval	Interval width
Classical Mantel-Haenszel	(-2.307 to 16.146)	18.453	(-9.261 to 7.878)	17.139
Classical inverse variance	(-4.452 to 9.334)	13.786	(-4.182 to 9.257)	13.439
Bayesian (Stable logit formulation)	(-6.876 to 8.948)	15.824	(-6.627 to 9.626)	16.253
Bootstrap normal	(-3.195 to 8.078)	11.273	(-1.916 to 6.991)	8.952
Bootstrap percentile	(-1.065 to 10.096)	11.161	(-4.172 to 4.899)	9.071
Bootstrap adjusted	(-0.832 to 10.830)	11.662	(-3.985 to 4.946)	8.931

Histograms and normal plots of the sample distribution of the effect size estimates the bootstrap confidence intervals are based on are provided in Figure 7.7. The distributions are highly skewed for both outcomes, although their skewness is in opposite directions. There are clearly big differences in both the width and location of the confidence intervals between the methods for both outcomes. In both instances the Mantel-Haenszel is the widest, followed next by the Bayesian interval, then the classical inverse variance interval, with the bootstrap based estimates being the narrowest. These results contradict those found using previous datasets, where Bootstrap confidence intervals were found to

be wider than those of standard methods. (Adams et al. 1997) Just as dramatic as the difference in the width of the intervals is the variability in their location. While no treatment difference is significant at the 5% level, the proportion of the interval either side of zero varies greatly, especially for the CHD mortality outcome.

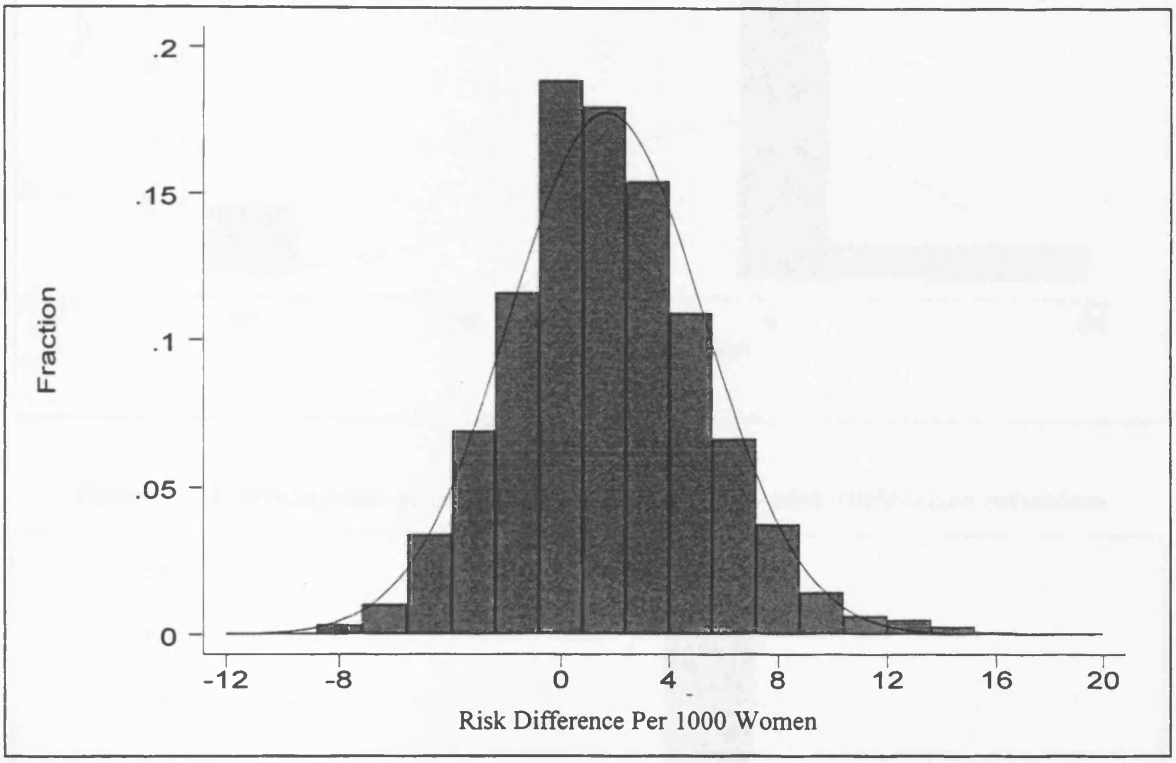
Figure 7.7 Histogram and normal plots for sampling distribution of the pooled estimate of the risk difference per 1000 women




Breast cancer mortality
Fraction
.12
.1
.05
0
-12 -10 -8 -6 -4 -2 0 2 4 6 8
Risk difference per 1000 women
Normal F[(bs1-m)/s]
1.00
0.75
0.50
0.25
0.00
0.00 0.25 0.50 0.75 1.00
Empirical P[i] = i/(N+1)

It is interesting to compare the Bootstrap distributions of the risk difference, with the posterior distributions obtained from a bivariate random effect Bayesian analysis. In the latter, the posterior looked symmetrical for the risk difference parameter (Figure 7.8). Hence, the Bootstrap sample, and the Bayesian posterior distributions appear to conflict.

Figure 7.8 Example histogram posterior distribution for the risk difference for breast cancer data



Using the bootstrap to inform about the random effect distribution

Figures 7.9 and 7.10 plot histograms of the risk difference for breast cancer and cardiovascular mortality risk from the postmenopausal hormone therapy trials. There is a suggestion that the distribution of effect sizes are non-normal in both instances. This can be largely attributed to the sparse event data in the studies in this extreme dataset. The appearance of these plots raises concerns about the appropriateness of a normally distributed random effect parameter.

Figure 7.9 Histogram of breast cancer mortality risk difference estimates

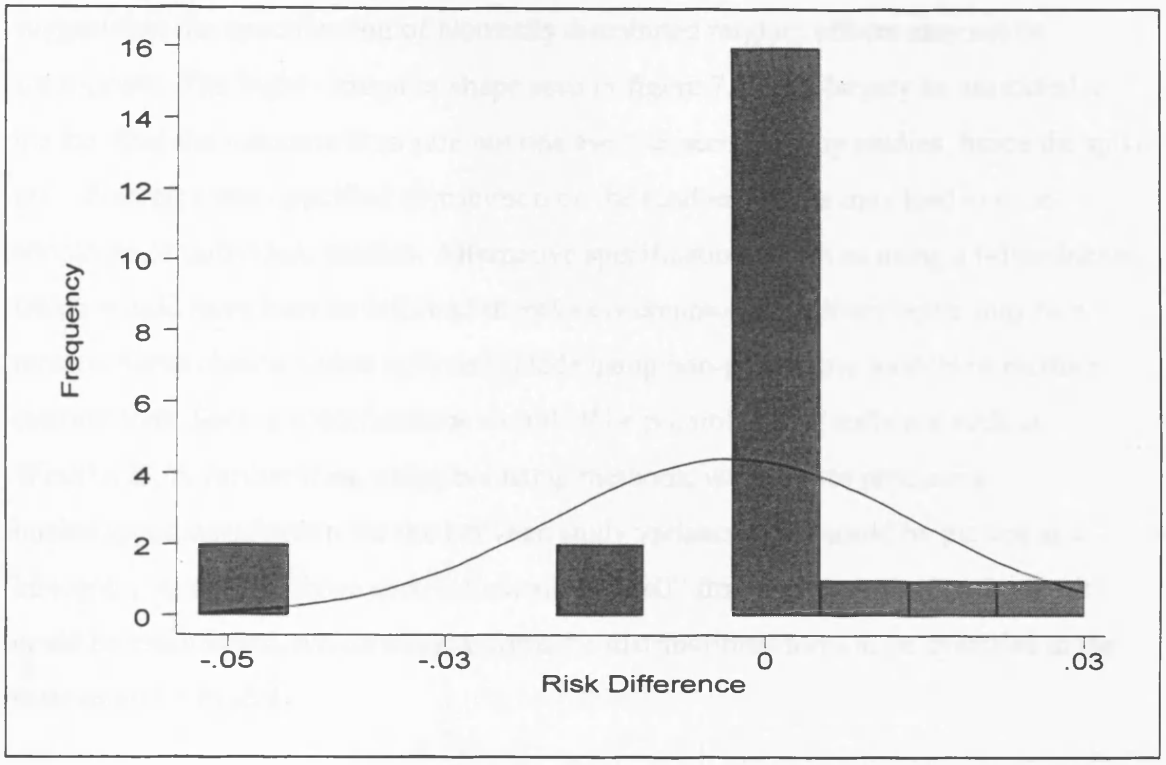
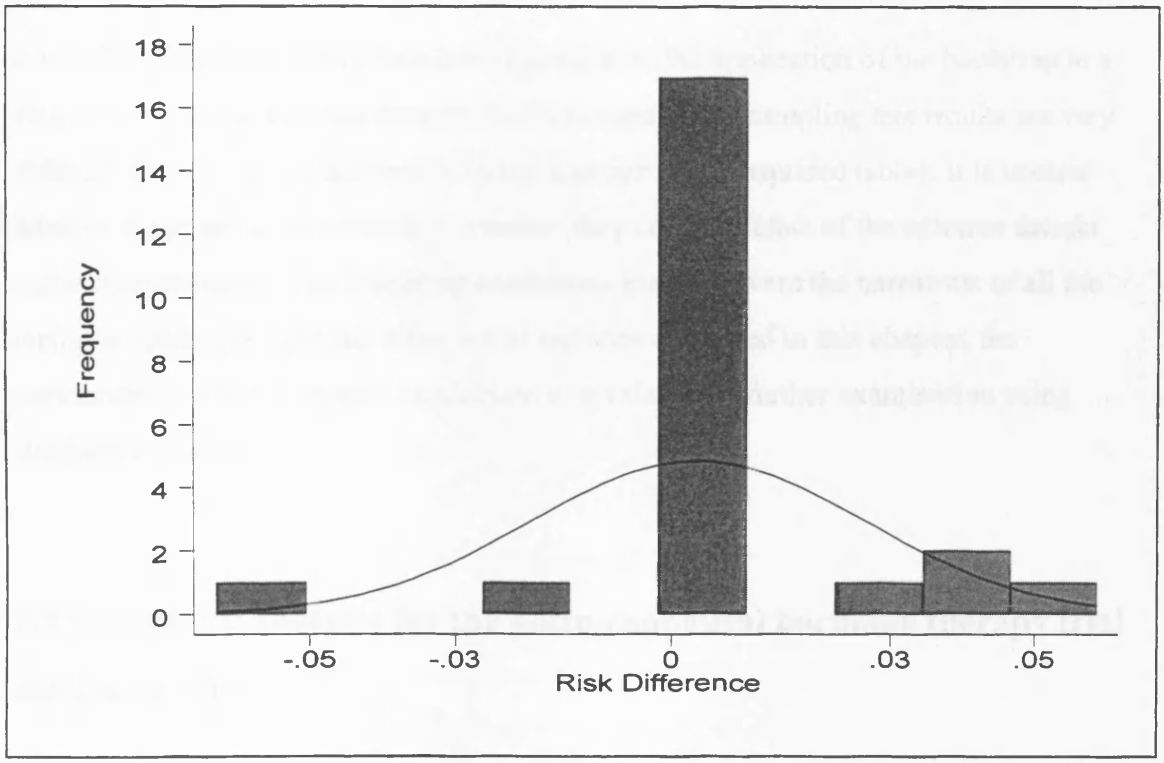


Figure 7.10 Histogram of cardiovascular mortality risk difference estimates



It is very difficult to consider the distribution of random effects for this dataset, due to the apparent lack of heterogeneity reported by the Q statistic, leading to a value for the

between study variance of zero for classical analyses. However, these histograms suggest that the specification of Normally distributed random effects may not be appropriate. The highly irregular shape seen in figure 7.10 can largely be attributed to the fact that the outcome is so rare not one event is seen in many studies, hence the spike at 0. Placing a mis-specified distribution on the random effects may lead to over-shrinkage of individual studies. Alternative specifications, such as using a t-distribution, which would have heavier tails and therefore accommodate outliers better may be a more suitable choice. Other options include using non-parametric models or mixture distributions. Such a specifications would all be possible using software such as WinBUGS. A further idea, using bootstrap methods, would be to produce a bootstrapped distribution for the between study variance. This could be plotted as a histogram, similar to those above, then and a "well" fitting parametric distribution(s) could be established, which could inform the distributional form to be specified in the meta-analysis model.

Discussion of the use of the Bootstrap

It is difficult to draw many firm conclusions from the application of the bootstrap to a single sparse meta-analysis dataset. The homogeneity re-sampling test results are very different from p-values derived from the appropriate chi-squared tables. It is unclear whether these are to be trusted, or whether they are an artefact of the extreme dataset under consideration. The bootstrap confidence intervals were the narrowest of all the methods observed. Like the other novel methods examined in this chapter, the performance of bootstrapped confidence intervals needs further examination using simulation methods.

7.7 Sensitivity analysis for the postmenopausal hormone therapy trial meta-analysis

Perhaps the strongest assumption made in the postmenopausal hormone therapy trial meta-analysis concerns not explicitly reported data. From Table 7.1 it can be seen that

outcomes were not explicitly mentioned for 13 and 14 of the trials for cardiovascular and breast cancer events respectively. In the original analysis, and in analyses in this chapter up to this point, it has been assumed that when no outcomes were reported then none occurred. This is a large assumption, and uncertainty concerning the data from over half the trials included has a potentially serious impact on the meta-analysis results.

Sensitivity analyses are underused in meta-analysis, though they are explicitly recommended in guidelines. (Sutton et al. 2000a) An approach which could be taken is to exclude trials for which there is doubt about the outcome data. However, since this would exclude more than half of the trials, this is not ideal, especially as there may be a relationship between outcomes and the probability of them being reported. Basing a meta-analysis only on studies where there is certainty under this scenario would produce a biased result, as this is a type of publication bias. A further approach would be to impute extreme values for the uncertain outcomes to assess the extent to which the pooled estimate is influenced. However, with so many uncertain data points, it is difficult to define what values could be considered extreme but still feasible.

Recently, a sensitivity analysis method to assess the impact of bias in reporting of outcome variables within studies on meta-analyses was described. (Hutton and Williamson, 2000) This method addressed the impact of the situation where outcomes with significant results are more likely to be reported, and hence can be considered a type of publication bias. Values were imputed for outcomes not reported in trials known to exist in the topic of interest.

Although a similar approach could be adopted here, due to the large number of studies which did not report explicitly the outcomes of interest, many permutations of possible true reporting patterns exist, which would require an extremely large number of analyses. A more general simulation approach is described and applied here. Because so many studies have uncertainty regarding outcomes it would be a very laborious task to re-run the meta-analysis with different values for the uncertain values individually, so the approach taken here is to generate simulated values for the uncertain values under different broad scenarios and examine the variability and range of the pooled estimates produced by these datasets.

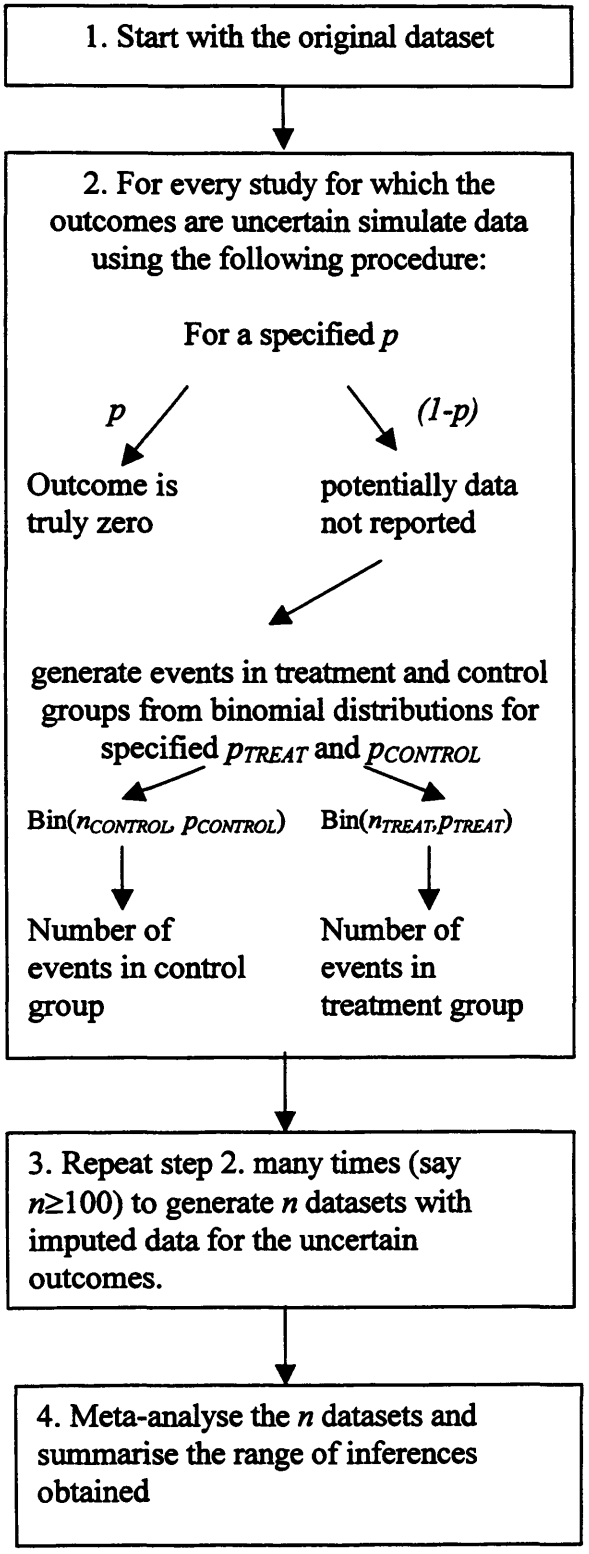
7.7.1 Simulation sensitivity analysis investigating potential impact of unreported outcomes

Although conceptually simple, the author is unaware of any previous attempts at a sensitivity analysis like the one described below. The process is illustrated using the cardiovascular outcome measure, about which there was uncertainty in 13 studies. For illustration, a fixed effect Bayesian exact model is used to combine studies, although the methods can be applied to any model. A fixed effect model was used because the computation time is reduced considerably, and since heterogeneity was estimated to be small from this dataset, a random effect model would give very similar answers.

It was decided to use a two stage random process to simulate data from the studies where there was uncertainty; the data in the other studies was always fixed at the observed values reported in the original meta-analysis. For each study where no outcomes were reported, it was assumed that outcomes would have been reported if they had occurred (i.e. there truly were no events) with probability, p, also there was a $(1\text{-}p)$ probability that outcomes may have occurred but not been reported. For the studies where it was deemed outcomes could have occurred, each individual in each arm of each study had a fixed probability of having the outcome in question, although the probability could be fixed at different values for treatment and control arms. Hence, the number of events in each arm were assumed to be drawn from a binomial distribution, with n equal to the number of people randomised to the arm, and probability of event equal to p_{treat} or $p_{control}$, depending on the arm.

Ideally, a new replicate of the simulated dataset could be made at every iteration of the Gibbs sampler within WinBUGS, but when a model requires the iterative results of a simulation solution (which the meta-analysis model does) this is not possible. Slightly more cumbersome, but still feasible, is to simulate the data using one WinBUGS program (although any other program which is able to simulate the data could be used), and then feed this simulated data into a meta-analysis routine in WinBUGS. These

Figure 7.11 Schematic representation of the simulation/analysis process



multiple simulated datasets can then be meta-analysed in one sampler run by setting the WinBUGS meta-analysis model code to loop over all datasets, thus greatly reducing the human time needed to carry out such an exercise. The whole simulation/analysis process is illustrated schematically in Figure 7.11.

Several sets of conditions for the simulated data are considered. The probability that a study would have reported the adverse outcomes had they occurred, p, is set to 0.7, 0.5 and 0.3. Two different values for the combination of p_{treat} and $p_{control}$ are explored. Both parameters' values are first set to 0.01, the approximate event rate in both arms for studies that reported outcomes explicitly. Then, while keeping $p_{control}$ fixed at 0.01, p_{treat} is set at 0. These second set of values were defined to explore the scenario where studies were more likely to report adverse events if they occurred in (at least) the treatment arm. This is a reasonable concern because original study investigators may have felt events observed in the control group may have been less cause for concern, because there is no possibility they could be due to the new treatment (This point was raised in the original correspondence to the BMJ following the original publication of the meta-analysis (Hemminki and McPherson, 1997)). Examining every combination of p, p_{treat} and $p_{control}$, specified above, resulted in the six simulations reported in Table 7.11.

SPECIAL NOTE

THIS ITEM IS BOUND IN SUCH A MANNER AND WHILE EVERY EFFORT HAS BEEN MADE TO REPRODUCE THE CENTRES, FORCE WOULD RESULT IN DAMAGE

Table 7.11 Results of simulation studies, for cardiovascular endpoint & odds ratio scale

P	p_{treat}	$p_{control}$	Total number of simulations	Mean median OR (s.d.)	Mean lower 2.5% (s.d.)	Mean upper 97.5% (s.d.)	Range lower 2.5%	Range upper 97.5%	Number of simulations 95% CrI included only values <1	Number of simulations 95% CrI included only values >1
0.7	0.01	0.01	100	1.59 (0.41)	0.63 (0.15)	4.51 (1.58)	0.29-1.16	1.70-9.82	0	2
0.7	0.00	0.01	100	1.30 (0.36)	0.48 (0.11)	3.88 (1.43)	0.19-0.66	1.14-6.78	0	0
0.5	0.01	0.01	100	1.50 (0.43)	0.64 (0.16)	3.94 (1.54)	0.32-1.11	1.41-9.14	0	2
0.5	0.00	0.01	100	1.05 (0.30)	0.41 (0.10)	2.91 (1.10)	0.19-0.64	1.03-6.64	0	0
0.3	0.01	0.01	100	1.47 (0.45)	0.66 (0.18)	3.58 (1.52)	0.30-1.11	1.24-10.54	0	5
0.3	0.00	0.01	100	0.89 (0.27)	0.35 (0.09)	2.33 (0.90)	0.16-0.61	0.82-5.64	1	0

It can be seen from Table 5.11 that 100 datasets were simulated for each of the six scenarios. The choice was largely arbitrary: the distribution of the odds ratios appeared reasonably smooth when histograms for each of the 100 simulations were constructed, and using 100 simulated datasets also meant the amount of computer processing time was reasonable. A burn in of 1000 followed by 4000 iterations was used to produce posterior distributions for each of the 100 simulated meta-analyses within each scenario.

The mean values over all the simulations for the pooled odds ratio and its 95% confidence are provided in the table for each scenario. These figures could be considered the sample mean parameter values given the conditions placed on the uncertain values. Table 7.11 would suggest that the original pooled estimate (1.81(0.61 to 6.04) using the same meta-analysis assuming all uncertain values are zeros) is reasonably robust over the conditions explored in the simulations. When p_{treat} and $p_{control}$ were both set to 0.01, two of the 100 simulations produced credible intervals for the treatment effect that included only values greater than 1 for postmenopausal hormone therapy for both $p = 0.7$ and $p = 0.5$.This increased to five when $p = 0.3$. Only one CrI that did not contain 1 was produced when p_{treat} was fixed at 0; this occurred when $p = 0.3$. Hence, since inferences changed in no more than 5% of simulations over the six scenarios explored, if it is believed that these scenarios cover the range of possibilities which could have actually occurred, then it can be concluded with a reasonable amount of certainty that inferences are robust to the uncertainty inherent in the data. However, it should be noted that there is considerable variation in the parameter estimates, and hence are not robust.

7.7.2 Extension to simulation sensitivity analysis

The previous section outlined a conceptually simple and powerful method of checking the robustness of a meta-analysis to uncertainty in outcome data from multiple studies. Such an approach reduced the dimensionality of the problem by eliminating the need to consider ranges of data for studies individually, by generating simulated data from

specified distributions. The approach can be made more general still, reducing the need to simulate multiple scenarios.

Consider again the cardiovascular endpoint. Table 7.11 provides results of simulations where p is the probability a study would publish the endpoint, if any events had occurred. In reality, it is difficult to know the likely range of values for p. In such a situation, a uniform distribution from 0 to 1 can be placed on p allowing p to vary between individual simulated datasets. In this way the sensitivity analysis considers all possible values for p, and considers them on an equal weighting. If prior information were available, perhaps from external sources such as an empirical investigation into the fullness of reporting of adverse events in trials, a different distribution could be placed, perhaps on a limited range of values, or giving more weight to values believed to be more likely.

Further, the probability of an event for any patient in any trial was assumed to be constant in the previous example, but this restriction can also be lifted by placing distributions on the probabilities of events in both arms. For example, below a N(0.01,0.0026) distribution is placed on both p_{treat} and $p_{control}$. This is derived by assuming the probability of an event for an individual in either group has mean 0.01, and lies between the range 0.005 and 0.015 95% of the time. In this example, 300 datasets were simulated, and the results provided in Table 7.12.

There is now a considerable increase in the numbers of simulated datasets that produced a significant odds ratio with a CrI that does not include 1. This is not surprising considering that a wider range of values for p were chosen, and that the rates in the two arms of a single trial were allowed to vary. Datasets leading to both harmful, and protective pooled odds ratios were generated. If the distributions placed on the simulation parameters are considered to yield only realistically possible values, then the conclusions of this meta-analysis are brought into doubt.

SPECIAL NOTE

THIS ITEM IS BOUND IN SUCH A MANNER AND WHILE EVERY EFFORT HAS BEEN MADE TO REPRODUCE THE CENTRES, FORCE WOULD RESULT IN DAMAGE

Table 7.12 Results of simulation studies, for cardiovascular endpoint & odds ratio scale placing distributions on data simulation parameters

P	p_{treat}	$p_{control}$	Number of simulat-ions	Mean median OR (s.d.)	Mean lower 2.5% (s.d.)	Mean upper 97.5% (s.d.)	Range lower 2.5%	Range upper 97.5%	No. simulations OR sig <1	No. simulations OR sig > 1
U(0,1)	N(0.01, 0.0026)	N(0.01, 0.0026)	300	1.45 (0.65)	0.72 (0.31)	3.30 (2.11)	0.16 – 2.54	0.44 – 19.94	9 (3%)	37 (12.3%)

7.7.3 Discussion of simulation sensitivity analysis

The previous sections have discussed a computer-intensive simulation approach to sensitivity analysis. Although the approach was developed here on an individual meta-analysis dataset, there is no reason why a similar approach could not be used in other applications where there is concern regarding biased reporting of outcomes.

Additionally, there is often uncertainty in data extracted from study reports, and differences in reporting practices mean data often have to be transformed often requiring assumptions before it is combinable, mean uncertainty regarding the data in a meta-analysis dataset will often exist. It may be desirable to include this uncertainty directly into the meta-analysis model, eliminating the need for a sensitivity analysis. Incorporating all uncertainty in a single stochastic model was one of the aims of the Confidence Profile method (Eddy et al. 1992) (Section 3.6). However, in examples such as the oestrogen replacement therapy one considered here, it is not obvious how this could be done satisfactorily. Re-running the meta-analysis many times imputing plausible/possible values for parameters for which data is uncertain, provides an appealing alternative. It should be noted that in principle a Bayesian model is not required to use this method, and indeed the use of Classical methods may make the process quicker. Nonetheless, in this particular example the problems associated with the sparse data make the Bayesian approach appealing. Spreadsheet programs allowing Monte Carlo type sampling could be used for this purpose.

Many extensions to the approach outlined above are possible, and can be tailored to the specific example in hand. For instance, the code generating the simulated data could be made more complex as required, so that multiple sources of uncertainty could be explored in a single simulation exercise. A further issue which needs clarification is the form the distributions for the parameters generating the simulated data should take. Two distinct approaches are possible. The first is to place uniform distributions over all theoretically possible values, and the second is to place essentially an informative prior distribution over the possible values.

The usefulness of such an approach largely depends on how reasonable and realistic the simulated data are, so careful thought should be given to the simulation structure and parameter values used.

7.8 Considering non-randomised evidence of adverse events associated with postmenopausal hormone therapy

One of the criticisms of the original analysis of the postmenopausal hormone therapy adverse event meta-analysis was that there was much observational evidence available which had not been considered. For example, the nurses' health study alone has (Grodstein et al. 1997; Colditz et al. 1995) followed over 120 000 women since 1976, in which over 3600 women died, and each of whom was matched with 10 controls. Interestingly, this observational evidence is considered as being superior to the randomised evidence by several of the authors of letters responding to the original RCT meta-analysis.

In this instance, there would be little point in performing a conventional meta-analysis which combined the observational evidence with the randomised evidence with the studies on an "equal footing" because the estimates from the observational evidence would completely "swamp" the evidence from the randomised studies. It would be possible to use a hierarchical model, as used in Chapter 5 to combine the cholesterol studies, putting a constraint on the weight given to the observational evidence, but since the observational evidence is questionably superior in this instance, it is questionable to whether this would be sensible.

For the purposes of illustration of how to proceed with including non-randomised evidence in the synthesis, attention is restricted to breast cancer evidence in the 23 RCTs trials described above, and the nurses' health study. (Colditz et al. 1995) The Nurses health study reported 923 cases of breast cancer over 344,942 person years of follow-up in those subjects who did not take hormone therapy. In the group which took conjugated estrogens alone, 270 cases of breast cancer were observed in 89,427

years of follow-up. This produces a relative risk, adjusted for many potential confounders, of 1.32 (1.14 - 1.54). The data being combined are summarised in Table 7.13. The RCT estimate used is the one produced by the exact logit fixed effect Bayesian model, pooling data from all 23 trials. The weights given in the last column of Table 7.13 correspond to those that would be allocated if a fixed effect inverse variance weighted model were applied to both sources of evidence. The RCTs are given 3% of the weight. Pooling these results "at face value" produces odds ratio of 1.31(1.13 to 1.52) (both fixed and random effect estimates are the same). Not surprisingly, the pooled estimate is very close to that from the Nurses Study.

Table 7.13 Summary data from 23 RCTs and Nurses' Study on risk of breast cancer and oestrogen therapy

Source	OR/RR (95% CI/CrI)	ln(OR/RR)	var(ln(OR/RR))	Fixed effect weight (%)
23 RCTs	0.91 (0.39 to 2.20)	-0.089	0.197	5.08 (3)
Nurses' Study (Prospective Cohort)	1.32 (1.14 to 1.54)	0.278	0.006	166.7 (97)

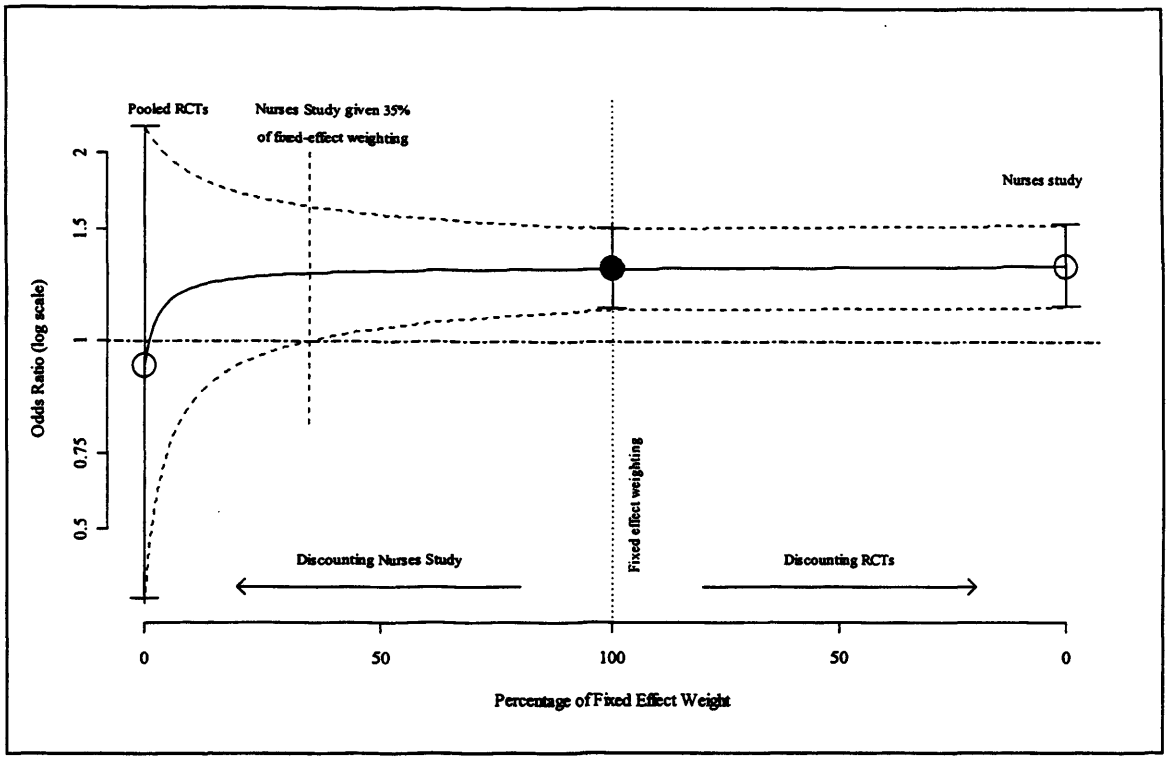
The relative weighting given to the types of evidence is only based on the precision of the estimates, and does not take into account many other relevant factors, many of which have been considered above and in previous chapters, including the uncertainty in the trial outcome data and bias due to confounding in the observational study. The next section considers simple plots which may aid the interpretation of the data, and any conclusions drawn from it.

7.8.1 Sensitivity plot to weighting of different sources of evidence

In the absence of any quantitative guidance on how much more/less reliable, or less/more biased the RCT evidence is compared to a large cohort study like the Nurses' study, it may be instructive to examine the effect of discounting the weight given to each source of evidence. A graphical plot which does this is presented in Figure 7.12. Here the pooled estimate and 95% confidence interval for the RCTs is displayed on the left hand side on the figure, and that for the Nurses' study on the right hand side. The pooled estimate produced by combining both sources using the standard fixed effect models is given in the middle. The solid line connects the point estimates for the different pooled estimates if each source of evidence was down weighted by a percentage of the original fixed effect weight, as indicated on the x-axis with down-weighting of the Nurses' study to the left, and the RCTs to the right, and the dotted lines connect the 95% confidence intervals for each of these estimates.

Using this plot, the effect of discounting sources of evidence can be assessed. If inferences remain constant over values one considers reasonable then the evidence is robust. For example, an extreme viewpoint would be to consider only RCT evidence as being at all reliable, and hence only the result considered would be that at the left hand side of the plot; conversely if the RCT evidence is considered totally unreliable, then the only result would be that on the right hand side of the plot. In general it is necessary to consider the robustness to inferences over some portion of the graph. Alternatively, the plot tells us under what weighting conditions inferences would change. A plot such as this may be particularly relevant when it is not clear if evidence from one source is superior to the other (such as for the oestrogen therapy data) and the effect of down weighting both sources may be of interest. Such an approach has a lot in common with the inverse inference approach considered by Glasziou et al. (Glasziou et al. 1990) In this example inferences change at the point where the cohort studies are given 35% of their original weight, down weighting them further produces a pooled estimate which is non-significant.

Figure 7.12 Sensitivity plot exploring the influence of down weighting sources of evidence



This plot is conceptually similar to the sensitivity plot advocated by Thompson (Thompson, 1993) for assessing the effect of weighing individual trials in a meta-analysis by varying the value of the between study variation. A further sensitivity analysis plot for meta-analysis which has been advocated, illustrates the impact on the pooled estimate of individual studies.(Tobias, 1999) The above plot extends this type of sensitivity analysis to a cross-design situation. In combination these plots provide a powerful way of exploring the robustness of the synthesis findings.

The idea of exploring the effect of discounting the weight given to sources of evidence based on the precision of the estimate alone is not new, and several schemes have been proposed for changing the weighting in a meta-analysis based on a quality assessment of the primary studies. (Tritchler, 1999; Berard and Bravo, 1998) The issue of study quality and its impact on a cross design synthesis is given further consideration in the discussion (Chapter 9).

7.8.2 A Bayesian approach to assessing the sensitivity of the weighting of different sources of evidence

The previous section described a plot to assess the robustness to a pooled result by discounting or down weighting sources of evidence. A fixed effect model was used to combine the results of an RCT meta-analysis with those of the largest observational study.

A natural Bayesian approach to this sort of sensitivity analysis is to derive a prior for the effect size in the RCT meta-analysis using the observational studies. For example, the estimate of the effect for the observational study (or pooled effect size if more than one observational study is been considered) could be specified as the mean for a normally distributed prior. By varying the variance specified for this normally distributed prior, the influence of down weighting the observational evidence can be ascertained. This has been illustrated recently for a synthesis of studies of electronic fetal heart rate monitoring (see Chapter 6) where the observational evidence was used to derive a prior distribution for a meta-analysis of the randomised evidence. (Sutton and Abrams, 2001) Three different standard errors were used; namely i) the 'naïve' one using the variance resulting directly from a meta-analysis of the observational evidence, ii) an 'equal' one using a prior with equal variance to the RCT meta-analysis and hence weighting each source of evidence equally, and iii) a 'sceptical' one where the variance of the observational studies is one quarter of the randomised evidence. Clearly, if once considered the observational evidence to be less susceptible to bias than the RCTs, (which although uncommon, may be the case for the HRT adverse outcome synthesis) then a further option would be to use the RCT evidence to derive a prior for the observational studies, allowing it to be down weighted by a degree.

7.9 Example 2: Breast implant side effects: Meta-analysis of sparse outcomes from observational studies with different designs

Epidemiological studies often examine the risks related to rare outcomes. If cohort studies are used to examine rare outcomes then they have to be large even to observe a small number of events. Case-control studies are often used in such situations due to greater ease of implementation.

As previous sections of this chapter have demonstrated, meta-analysis is a potentially valuable tool for examining risks associated with rare events because of the increase in statistical power gained from combining different studies' results. It has been shown that many of the traditional methods of meta-analysis produce questionable results, or break down altogether, when combining data from 2 by 2 tables with sparse events. Bayesian alternatives which use more exact methodology were also discussed, although problems with placing priors on the between study variance parameter(s) in random effect models were noted.

Further complications exist if not all the studies data can be represented in a 2 by 2 table. For example, a recent meta-analysis of breast implants and connective tissue disease (Perkins et al. 1995) included case-control studies, cohort studies with internal comparison groups, and a cohort study using an external comparison group. Due to the fact that data from different study designs is being combined and that the data cannot be represented in the same 2 by 2 table for all studies exact synthesis methods for 2 by 2 tables were not possible. The data available from each study is provided in Table 7.14.

Table 7.14 Epidemiological studies of breast implants and connective tissue diseases (reproduced from (Austin et al. 1997))

Case-control studies

Study Number	Cases		Controls		Relative risk (95% CI)
	Exposed	Unexposed	Exposed	Unexposed	
1	1	299	12	1,444	0.40 (0.01, 2,74)
2	12	857	23	2,038	1.24 (0.56, 2.61)
3	4	247	5	284	0.92 (0.18, 4.43)
4	2	272	14	1,170	0.61 (0.07, 2.70)
5	1	132	0	100	∞(0.2, ∞)

Follow-up studies (internal comparison)

Study Number	Cases		Woman-years		Relative risk (95% CI)
1	1	1	616	663	1.08 (0.01, 84.5)
2	5	10	5,847	12,361	1.05 (0.28, 3.39)
3	3	513	11,170	1,170,074	0.61 (0.31, 1.80)

Follow-up studies (external comparison)

1	Observed =3	Expected = 1.47	2.04 (0.42, 5.98)

For this meta-analysis Austin et al. (Austin et al. 1997) developed new methodology to combine these results across different study designs using 'exact' methods. Essentially, their method manipulates the cohort studies data into the 2 by 2 table format used to report the case-control studies, and then uses Mantel-Haenszel type formula to combine them. This is achieved by multiplying the denominators in the cohort studies by very large numbers. For example, in the first internal comparison study above there was 1 exposed and 1 unexposed case with person-time denominators of 616 and 663 woman-years, respectively. In the original analysis, the person-time denominators were multiplied by 10^4; converting the data to 1 exposed

and 1 unexposed case and 6,160,000 exposed and 6,630,000 unexposed controls. While this ingenious approach appears completely satisfactory for a fixed effect analysis, no random effect model is developed.

A random-effects model is possible, taking a Bayesian approach, with solutions derived using MCMC via the software package WinBUGS. This model, which uses the same assumptions about the data as the original fixed effect approach (Austin et al. 1997) (other than the implications due to the inclusion of a random effect between studies), is described below.

Each of the three types of studies is modelled distinctly. Firstly, the case-control studies are considered. It is assumed that the events in the diseased and non-diseased groups are derived from underlying binomial distributions. The difference between the logit proportions of events in the two groups provides an estimate of the log odds ratio. This part of the model is exactly the same as the standard random effects model for meta-analysis (equation 2.18). Arranging the data as in Figure 2.1b, the model specification is given below for completeness.

$$a_i \sim Bin[p_{1i},(a_i+b_i)] \qquad c_i \sim Bin[p_{2i},(c_i+d_i)] \qquad i=1.....5$$

$$\log it(p_{1i}) = \mu_i - d.cc_i/2 \qquad \log it(p_{2i}) = \mu_i + d.cc_i/2$$

$$\mu_i \sim N[0,10^5] \qquad d.cc_i \sim \mathrm{N}[\phi,\tau^2],$$

where $d.cc_i$ indicates the treatment effect in the ith case control study.

The cohort studies using an internal comparison group are considered next. The expected number of events in each group, in this example cases of connective tissue disease, is calculated as the underlying rate of disease multiplied by the number of person (woman) years observed. Hence,

$$expected_{1j} = \lambda.ic_{1j} \times py_{1j} \qquad j = 1, . . , 3$$

$$expected_{2j} = \lambda.ic_{2j} \times py_{2j}$$

The observed numbers of events in each group are assumed to be distributed Poisson with rate parameters defined by the expected number of cases,

$$events_{1j} \sim Poisson(expected.ic_{1j})$$

$$events_{2j} \sim Poisson\ (expected.ic_{2j})$$

The difference in the logs of these rates provide an estimate of the relative risk of having connective tissue disease, for those having breast implants against women who have not. These again are assumed distributed in an identical manner to the estimates from the case control studies.

$$\log(\lambda.ic_{1j}) = \mu_j \qquad \log(\lambda.ic_{1j}) = \mu_j - d.ic_j$$

$$\mu_i \sim N[0,10^5] \qquad d.ic_i \sim N[\phi, \tau^2]$$

In important point is that the case-control studies estimate an odds ratio, while the cohort studies estimate a relative risk. In this model both of these are equated to one another, and assumed to be the same quantity. Since the outcome of interest is so rare this should be a reasonable assumption.

Now for the cohort study which use an external comparison group. The observed number of cases is assumed to be distributed Poisson. There is no need to include a woman-year follow-up time term in the model as the expected number in an unexposed population assumes equal number of years follow-up, hence,

$$events.ec_k \sim Poisson(\lambda.ec_k) \qquad k = 1$$

An estimate of the log relative risk, *d.ec* is obtained by specifying the log of the rate in the exposed group is the sum of the baseline rate plus the excess

$$\log(\lambda.ec_k) = \log(expected.ec_k) + d.ec_k$$

$$d.ec_i \sim \mathrm{N}[\phi, \tau^2]$$

Finally, non-informative prior distributions are placed on the parameters of interest, ϕ and τ^2.

$$\phi \sim Normal(0,10^6)$$

$$\tau^2 \sim InverseGamma(0.001,0.001)$$

The pooled estimate of the relative risk for the breast implants and connective tissue disease using this model is 0.96 (95% CrI 0.59 to 1.49). This can be compared to the fixed effect estimate reported previously (Austin et al. 1997) of 0.98 (0.63 to 1.48). Although very similar, the random effect estimate has a slightly wider confidence interval which takes into account between study heterogeneity, and the uncertainty in its estimation. This example provides an excellent example of the power of WinBUGS for fitting flexible example specific models, here where the data of interest are in different formats, and is easier to program and more 'natural' than the classical solution suggested previously.

Several straightforward extensions to this model would be to add uncertainty via specification of a distribution for the external risk parameter (it is assumed known in the example), and to consider a three level model of the form considered in Chapter 5 allowing for heterogeneity between types of study.

7.10 Summary

This chapter has given attention to synthesis of evidence pertaining to rare outcomes. A generalised synthesis in areas such as adverse reactions to drugs, which are often rare, is appealing since evidence from trials alone is often underpowered. Hence, while parts of this chapter concentrate on a meta-analysis situation concerning evidence from one study type, much of the work is relevant to a generalised synthesis context and its potential application there is highlighted throughout the chapter.

There are issues specifically relating to meta-analysis of rare outcomes that have been considered in this Chapter including the impact of continuity correction factors, choice of outcome scale, use of exact methods, influence of prior distributions and the use of non-parametric confidence intervals. This chapter has outlined existing and several novel methods of estimating effect sizes when data are sparse. The novel methods include the use of bootstrap methodology and the use of data inflation in a binary meta-analysis context. It has been demonstrated that different conclusions can be obtained depending on the method used. Further, rigorous research (that was beyond the scope of this thesis) is required to study the performance of these methods before advice can be given on which are recommended. Specifically, simulation work i) investigating the performance of different continuity correction factors on different meta-analysis methods, ii) validating the use of data inflation in a meta-analysis context, and iii) investigating the performance of bootstrap methods for sparse meta-analysis data is required.

A novel simulation method to assess the potential impact of missing outcomes (potentially non-randomly) is described which could be adapted as a form of sensitivity analysis in other meta-analyses where there is uncertainty in the reported data. A further sensitivity analysis plot is described which can be used to assess the robustness to a synthesis to the relative weights given to the different sources of evidence. Finally, a Bayesian model for a synthesis of evidence from observational studies with different designs, using exact methods, is described which illustrates the flexibility of the Bayesian approach in specifying 'non standard' models.

Chapter 8 Meta-analysis of composite measures of benefit and harm: a Bayesian exposition of the Net Benefit model

8.1 Introduction

Chapters four and five considered a generalised synthesis of evidence which combined information from different study types in a single model. This chapter also considers how to combine data from studies with different designs; previously, however, all studies were essentially providing estimates of the same quantity, but here different study types provide estimates for different parameters in the net-benefit model described below. Section 3.11 outlined an approach to generalising RCT results using additional information called net benefit which can be summarised by equation 3.5 repeated below.(Glasziou and Irwig, 1995)

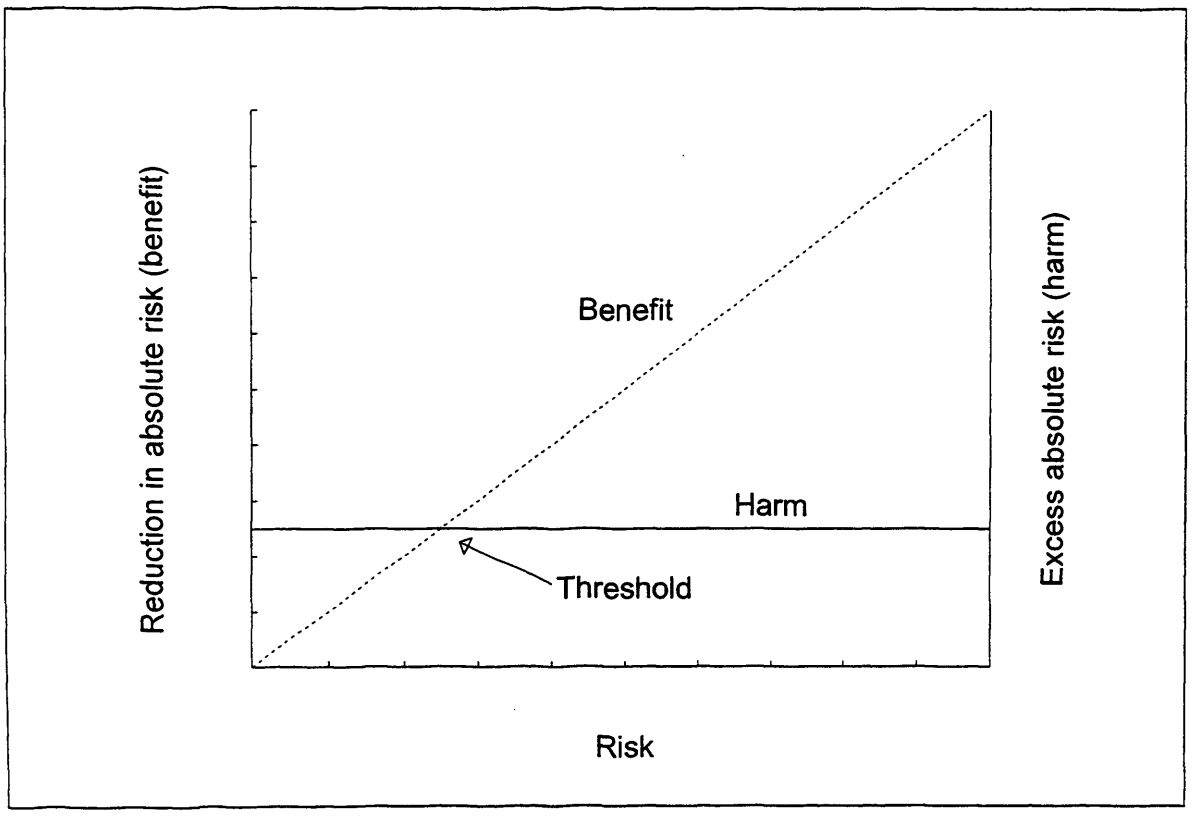
$$\text{Net Benefit} = (\text{Risk Level} \times \text{Risk Reduction}) - \text{Harm}.$$

In this chapter, the example originally used to illustrate the method - assessing whether low-dose warfarin should be given to patients with non-rheumatic atrial fibrillation - is implemented using a Bayesian model which attempts to take into account all sources of uncertainty. The analysis combines data from several sources relating to different quantities required by the model, including several different outcomes. The term net benefit as used here should not be confused with the recent usage of the term in the health economics literature. (Stinnett and Mulahy, 1998) Although there are some similarities in the modelling, economists equate effectiveness to costs to evaluate cost effectiveness in contrast with the different clinical outcomes equated here to evaluate solely clinical benefit. However, consideration of how costs could be included in the modelling presented here is discussed in Section 8.6.

Specifically, this approach utilises separate assessments of the benefit and the harm of treatments, and is based on the premise that patient benefit increases with absolute risk from the disease. This is not to be confused with the situation considered in Chapter 4, where the relative as well as the absolute treatment effect of cholesterol reduction increased with risk. (Thompson et al. 1997) Here we are assuming, what is perhaps the

more common situation, where the relative risk is constant across levels of absolute risk. The method further assumes that the harm caused by the new treatment, through patients experiencing adverse events, is constant across levels of patient risk. Hence, for some treatments a threshold value of patient risk may exist where the risk of an adverse event outweighs the benefit of the treatment, even when the treatment may have been shown to be effective, on average in clinical trials but in high risk patients the benefit of risk reduction will be sufficient to outweigh harm. This is illustrated graphically in Figure 8.1

Figure 8.1 Graphical representation of the net-benefit model (Based on Figure 1 of Glasziou and Irwig (Glasziou and Irwig, 1995)). Benefit increases with risk, while harm is assumed constant. Net benefit occurs only when risk is above threshold.



Hence, the model aims to establish the net benefit of treatment for given levels of patient risk, opposed to producing a single pooled treatment effect. Although it is assumed here that harm is constant across levels of risk, it would be possible for it to take any functional form desired if this was thought to be too simplistic.

8.2 Method outline

Glasziou and Irwig (Glasziou and Irwig, 1995) outline the steps required to apply this model. Firstly, they state that an RCT, or a meta-analysis of RCTs, is the most appropriate method to estimate the relative risk for the benefit of the intervention. It is necessary to check that the assumption that the relative risk does not vary with patient risk is reasonable; this will not be the case if the intervention has both positive and negative effects on one outcome (e.g. this was potentially true when total mortality was considered for cholesterol lowering treatments in Chapter 4 since it has been suggested that lowering cholesterol increases rates of accidents/suicides due to drugs changing behaviour). Note, it would be possible to model outcome as a function of patient risk, using meta-regression models such as those applied to the cholesterol analysis in Chapter 4 if this were appropriate, however this is not pursued here. Further, the meta-analysis data can be used to check that absolute harm is independent of risk. Further data on side effects may be available from sources other than RCTs. If these two assumptions are satisfied, then the predicted benefit needs to be weighed up against the potential harm. In order to do this, both the benefit and harm outcome need placing on the same scale. This could be achieved by assessments of quality of life following different events. Finally, in order to apply the model usefully, and identify patients who should expect benefit to be greater than harm, we need to predict each patient's risk. In order to do this, the major risk factors need identifying and multivariate risk prediction equations constructing. This information may come from cohort studies or from RCTs; because eligibility criteria for trials are often narrow, population based cohorts studies are preferable.

8.3 Example: Re-analysis of anticoagulants and non-rheumatic atrial fibrillation

A re-analysis of the original example examining the use of anticoagulants for non-rheumatic atrial fibrillation is presented. This analysis uses essentially the same data used by Glaziou and Irwig, for reasons of clarity, and to enable comparisons with that paper. (Glasziou and Irwig, 1995) It is acknowledged that the analysis is not using the most up-to date evidence, and hence should not be considered as definitive.

Data on six trials comparing low dose warfarin for patients with non-rheumatic atrial fibrillation to placebo are presented in Table 8.1; references to the original RCTs are available elsewhere (Glasziou and Irwig, 1995)

SPECIAL NOTE

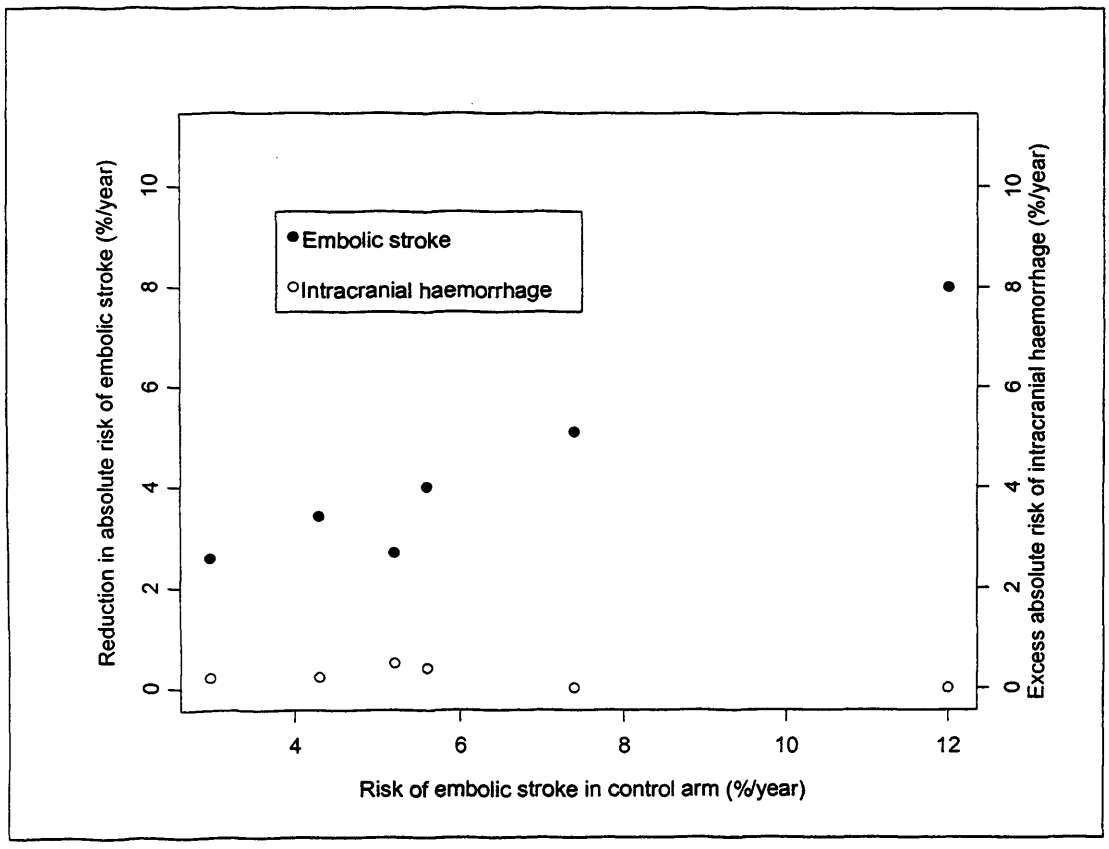
THIS ITEM IS BOUND IN SUCH A MANNER AND WHILE EVERY EFFORT HAS BEEN MADE TO REPRODUCE THE CENTRES, FORCE WOULD RESULT IN DAMAGE

Table 8.1 Randomised Trial outcome data for low dose warfarin RCTs

	Warfarin Group				Control group				
Trial	*No. of patients*	*No. of strokes*	*Annual rate of stroke (%)*	*No. of fatal haemo-rrhages (%)*	*No. of patients*	*No. of strokes*	*Annual rate of stroke (%)*	*No. of fatal haemo-rrhages*	*% reduction in relative risk (95% CI)*
Boston	212	2	0.41	1 (0.21)	208	13	3.0	0	86 (51 to 96)
Veterans Affairs	260	4	0.88	1 (0.22)	265	19	4.3	0	79 (52 to 90)
Canadian	187	5	2.5	2 (0.50)	191	11	5.2	0	52 (-36 to 87)
Atrial Fib, aspirin	335	4	1.6	1 (0.40)	336	21	5.6	0	71 (23 to 90)
Stroke prevention	210	6	2.3	0 (0)	211	18	7.4	0	69 (27 to 85)
European	225	20	4.0	0 (0)	214	50	12.0	0	66 (43 to 80)

The final column of Table 8.1 provides the % reduction in relative risk (i.e. (1-RR)%) of having an embolic stroke. From these studies, at least, there is little evidence to suggest the relative risk is not constant. Using the data in Table 8.1 Figure 8.2 can be constructed. This displays the absolute risk difference for strokes and the adverse event haemorrhage. This suggests that the reduction in absolute risk of thromboembolic strokes (annual % in the control group – annual % in the treatment group) rises linearly with the risk of stroke, while the rate of intacranial haemorrhage seems to be stable across varying risks of strokes. Hence both of the required assumptions appear to hold.

Figure 8.2 (Based on Figure 2 of Glasziou and Irwig (Glasziou and Irwig, 1995)) Trials of warfarin in atrial fibrillation showing that benefit (reduction in absolute risk of stroke) increases with increasing risk of stroke but that harm (intracranial haemorrhage) seems to be constant



The net benefit equation can now be derived. In the original analysis, a meta-analysis of five of six of the trials reported in Table 8.1 (the European trial had not been published in time for inclusion in the meta-analysis report used in the original analysis (Singer, 1993)), was used to calculate the reduction in relative risk of a stroke by taking warfarin. An odds ratio of 0.73 (0.57 to 0.83) was calculated using the Mantel-Haenszel technique in this meta-analysis. For the net benefit equation, it appears that this OR was assumed to be equivalent to the RR since 1 – 0.73 = 0.27 was used as the estimated reduction in RR in the net benefit equation (3.5).

For this parameter estimate, and the estimates for all the other quantities required for the model, the estimated values are assumed to be known, when in fact they were estimated from previous studies, and hence the uncertainty inherent in the estimates was not accounted for. A natural way to incorporate such uncertainty, which, among other advantages, allows confidence intervals for net benefit to be constructed, is through the use of simulation methods.

If the distribution of each input parameter required for the model were known then, evaluation would be straightforward using Monte Carlo simulation. However, MCMC methods (within WinBUGS) are adopted here, as this approach has several theoretical advantages over the simpler Monte Carlo simulation approach, which are outlined below.

8.3.1 Evaluating the efficacy of warfarin for preventing strokes

The original analysis result (OR/RR = 0.73 (0.57 to 0.83)) translates to a normal distribution on the log scale of

$$\ln RR \sim \text{Normal}(-0.315, 0.0092),$$

from which a relative risk reduction can be derived using the equation

$$RRR = 1 - \exp(\ln RR). \tag{8.1}$$

Although such a specification could be included in the net benefit model, since this would appear to accurately represents the findings of the original meta-analysis, a better approach is to incorporate all the data and repeat the meta-analysis within the model specification. This has several advantages. Firstly, a fixed effect analysis was utilised by Glasziou and Irwig, and a Bayesian random effect analysis is preferred for advantages outlined in previous sections of this thesis and elsewhere (Sutton and Abrams, 2001). Secondly, in instances when the ln(OR) is not (approximately) normally distributed (for instances when the distribution for the pooled effect size is not symmetric see Chapter 7) the correct (posterior) distribution for the pooled effect will be used. Third, in instances when data from the studies in this evaluation are included in other parts of the model (which is the case here), estimates will be correlated and no longer independent. The effects of doing so will automatically be accounted for in all parameter estimation.

Hence, an exact binomial random effect meta-analysis model is specified as described in equation (2.18), and the second line of equation (8.1) is included in order to calculate the reduction in the relative risk. Vague prior distributions are placed on the pooled log odds ratio (Normal(0,10^6)), and the between study variance (InverseGamma(0.001,0.001)).

8.3.2 Evaluating the risk of an intercranial haemorrhage (fatal bleed)

For the original analysis, a previously published meta-analysis specifically examining the risk for fatal haemorrhage following treatment by warfarin (Landefeld and Beyth, 1993) was used to derive an estimate of harm. This meta-analysis included data from both randomised trials and longitudinal studies of inception cohorts of patients followed from the start of warfarin therapy. Three of the RCTs used to evaluate the RRR for warfarin were included in this meta-analysis. Also included were trials and cohort studies of conditions other than atrial fibrillation where warfarin is used as an intervention; these included including cerebrovascular disease, prosthetic valve, myocardial infarction and deep vein thrombosis. The results of 21 RCTs and 4 cohort studies included are reproduced in Table 8.2. Note, the numbers in the first column indicate the reference number in the original meta-analysis, (Landerfeld and Beyth, 1993) although the three trials also included in Table 8.1 have been named for easy identification.

Table 8.2 Data for studies estimating the risk of fatal bleeding following treatment with warfarin

Indication for Warfarin	No of patients	Total length of treatment (y)	Fatal bleeds (No. (%))	Fatal bleeds (%/y)
Experimental studies				
Cerebrovascular disease				
21	78	45.3	4 (5)	8.8
22	52	96.8	3 (6)	3.1
23	95	93.4	2 (2)	2.1
Prosthetic valve				
24	65	121.8	1 (2)	0.8
25	210	52.5	0 (0)	0.0
26	247	857.0	2 (1)	0.2
Myocardial Infarction				
19	145	362.5	1 (0.7)	0.2
27	119	418.3	4 (3)	1.0
20	128	168.0	0 (0)	0.0
28	68	109.8	0 (0)	0.0
5	607	1,872.0	3 (0.5)	0.2
Atrial fibrillation				
Stroke prevention	212	444.0	1 (0.5)	0.2
Boston	210	260.0	1 (0.5)	0.4
Canadian	187	236.9	2 (1)	0.8
Miscellaneous indications				
29	50	8.3	0 (0)	0.0
30	156	139.0	0 (0)	0.0
Deep vein thrombosis				
31	24	6.0	0 (0)	0.0
32	33	8.3	0 (0)	0.0
33	53	113.3	0 (0)	0.0
34	96	24.0	0 (0)	0.0
35	198	49.0	0 (0)	0.0
Observational studies				
Prosthetic valve				
38	183	506.0	3 (2)	0.6
39	415	726.2	5 (1)	0.7
40	122	365.7	2 (2)	0.5
Miscellaneous indications				
41	565	904.0	10 (2)	1.1
Total	**4,318**	**7,988.1**	**44 (1)**	**0.6**

The figure in the bottom right hand corner of Table 8.2, 0.6% risk of a fatal bleed per year, was originally used in the harm model. Note, this is obtained by taking a simple un-weighted average of the rates in each of the studies. In the analysis here a more sophisticated random effects meta-analysis model, which assumes the events follow a Poisson distribution, is adopted

$$
\begin{aligned}
&bleeds_j \sim \text{Poisson}(rate_j) \\
&\ln(rate_j) <\text{-} beta_j + \ln(followup_j) \\
&beta_j \sim \text{Normal}(ln.harm, tau\text{-}sq) \qquad (8.2) \\
&ln.harm \sim \text{Normal}(0,10^6) \\
&tau\text{-}sq \sim \text{Inverse-Gamma}(0.001,0.001),
\end{aligned}
$$

where $bleeds_j$ is given in column four of Table 8.2, and $followup_j$ in column three for all studies except the three RCTs also included in the meta-analysis for treatment benefit as described above. Although the total follow-up time could be imputed directly into the model for these three RCTs, as for the remaining studies it is calculated as the number of persons in the treatment group (*nt* – as given in columns 2 of both Tables 8.1 and 8.2) multiplied by the average patient follow-up time (estimated here has total length of treatment/total number of patients, or column three divided by column two of Table 8.2). Specification in this format indicates that the assessments of benefit and harm are not independent since the same data (i.e. *nt*) is used in both treatment and harm models.

8.3.3 Evaluating the trade-off between a stroke and a haemorrhage event in terms of quality of life

Since only the benefit of reduction in non-fatal strokes is being considered, and being compared to the harm of fatal haemorrhages, it is necessary to equate how many strokes are equivalent to one fatal haemorrhage, or more precisely putting the two outcomes onto a single scale. In the original analysis a study of quality of life assessments of

patients following a thromboembolic stroke were used for this purpose. (Glasziou et al. 1994) They concluded that the average quality of life after thromboembolic stroke is between 0.7 and 0.8, on a scale of 0 (death) to 1 (normal good health), and hence a ratio of about 4 strokes to one haemorrhage is suggested (i.e. reduction of quality of life per stroke is 0.25 therefore four strokes is equivalent to a change from one to zero on the scale, or a death). Clearly, such reasoning assumes that measurements on the QoL scale are additive.(Drummond et al. 1997) Incorporating QoL in the model in this way does not account for uncertainty in the quantitative equating of strokes to haemorrhages, but more importantly it makes a fundamentally flawed assumption that

$$1/\text{Expectancy(QoL reduction)} = \text{Expectancy}(1/\text{QoL reduction}). \qquad (8.3)$$

This does not hold true for a skewed distribution, and since the QoL data is highly skewed this produces a very large error as is illustrated below. (This same mistake is often encountered in the analysis of cost data where, due to the fact that the distribution of costs is highly skewed, a transformation is often used before the data is analysed. Unfortunately, biased results are obtained if this is not accounted for when back transforming. (Thompson and Barber, 2000))

In this analysis, data from a time trade-off survey is used. (Glasziou et al. 1994) Persons having experienced a myocardial infarction were asked six months after the event how many years they would be willing to give up in exchange for returning to full health. A time trade-off index for life expectancy of 15 years was calculated as follows

$$\text{QoL time trade off index} = \frac{15 - \textit{years given up}}{15}. \qquad (8.4)$$

Although data for individuals were not tabulated a graph was included in the original paper. (Glasziou et al. 1994) From this it appeared that people generally responded in whole years, implying only a maximum of 15 different values were actually given for the index, the exception to that being for responses under a year where it would appear fractions were reported. Unfortunately, the graph was not detailed enough to accurately distinguish these fractions of one year, and hence Table 8.3 represents the most accurate data extraction possible for the empirical distribution function for the quality of life

time-trade off index. (The author did request the individual patient data from the original study investigators, but unfortunately they were no longer available.)

Table 8.3 Empirical distribution for the Quality of Life time trade off index following stroke

QOL time trade off index	Proportion of patients	Cumulative proportion of patients
0	0.00	0.00
0.07	0.12	0.12
0.13	0.00	0.12
0.20	0.00	0.12
0.27	0.00	0.12
0.33	0.10	0.22
0.40	0.00	0.22
0.47	0.00	0.22
0.53	0.00	0.22
0.60	0.00	0.22
0.67	0.13	0.35
0.73	0.10	0.45
0.80	0.00	0.45
0.87	0.00	0.45
0.98	0.54	0.99
0.99	0.01	1.00

In order to take into account the distributional shape and the uncertainty in the parameter estimates, the full data in Table 8.3 are included in the net-benefit model. The probability that an individual responds to each of the 16 index states is modelled using a multinomial distribution, and hence

$$rcat_i \sim \text{Multinomial}(p_i, N) \qquad i = 1 \ldots\ldots 16$$

where $rcat_i$ represents the number of persons in each of the $i = 16$ categories, p_i the probability of being in the ith category, and N the total number of people who responded in the survey. Vague priors (Beta(1,1)) are placed on each of the p_is, for $i = 1$ to 15, and p_{16} is defined as $1 - \text{sum}(p_1 \ldots . p_{15})$. An alternative approach would be to fit a skewed continuous distribution to the data, however this is not pursued here.

Unfortunately, the total number of persons who experienced a stroke is not provided in the original report, hence an estimate was made. It is known that a total of 647 individuals responded to the question in the study and they were categorized into stroke, re-infarction or neither. For purposes of modeling, after considering the event rates in the control arms of the six warfarin RCTs, it was assumed that around 15%, or $N = 100$ of those individuals had had a stroke within six months of experiencing a myocardial infarction. Note, in the modeling it would be possible to place a distribution on the number N to represent our uncertainty about it.

The 16 index scores were assigned to the 16 categories of the model, and the distribution of index scores evaluated. From this the sample posterior of QoL reduction (defined as 1-index score) was calculated allowing the distribution for the ratio of the two outcomes (1/QoL reduction) to be derived. Using this function, the mean QoL score is 0.74, but because of the skewed nature of this distribution, the median and 2.5 and 97.5 centiles are (0.07 to 0.98).

A potentially important limitation of using these data is that those persons who had a fatal stroke are not been included in the calculation of the outcome ratio, since only patients surviving to six months were interviewed. This potentially makes the stroke outcome less 'serious' than it is as only the non-fatal ones have been considered, and hence warfarin less beneficial than it truly is. This is a limitation of the data available rather than of the modelling framework. Since this chapters' primary aim is to demonstrate the potential of Bayesian methods to implement models such as the one described here, rather than to be a definitive assessment of the warfarin for atrial fibrillation literature, this limitation is noted but not resolved.

8.3.4 Evaluation of an individual's risk of a stroke

The combined effect of the factors which influence the risk of stroke in atrial fibrillation were examined by the Stroke Prevention in Atrial Fibrillation Investigators in a multivariate risk model. (Glasziou et al. 1994; Thacker et al. 1997) This study suggested that three clinical features – hypertension, recent congestive cardiac failure, and previous thromboembolism – and two echocardiographic features – left ventricular dysfunction and atrial size – are important. Table 8.4 summarises the risks and prevalence associated with these characteristics as found in the study cohort.

Table 8.4 Risk factors for stroke derived from a cohort study

	No. of patients	*Percentage of Cohort*	*Thromboembolism rate (% per year (95% CI))*	*Log Thromboembolism rate (standard error)*
No. Clinical risk factors				
0	241	42	2.5 (1.3 to 5.0)	-3.69 (0.34)
1	259	46	7.2 (4.8 to 10.8)	-2.63 (0.21)
2 or 3	68	12	17.6 (10.5 to 29.9)	-1.74 (0.27)
No. Clinical risk factors + echocardiographic features				
0	147	26	1.0 (0.2 to 4.0)	-4.61 (0.76)
1 or 2	336	60	6.0 (4.1 to 8.8)	-2.81 (0.19)
≥ 3	78	14	18.6 (11.6 to 30.1)	-1.68 (0.24)

The net benefit for these risk levels, expressed as percentage per year can be evaluated, including the uncertainty in the levels of risk. These rates were estimated using Poisson regression; hence the log risks should be normally distributed, with the appropriate standard errors derived from the log risk confidence intervals. (Greenland, 1987) Note,

unlike the meta-analyses, which were re-analysed in the model specification from a Bayesian standpoint, the results of the original classical Poisson regression model are used directly in the net-benefit model. Although, theoretically the Poisson regression model could be included as part of the net benefit model, this was not possible here as individual patient level data are required, and they were not easily available.

8.3.5 Specifying the net benefit equation

The previous sections have described how each segment of the analysis necessary to estimate all parameters in the net benefit equation (3.5) can be specified from a Bayesian standpoint using WinBUGS. This approach is very 'neat' since all the analysis required is contained in one single integrated concise program. Additionally, it also (automatically) includes the uncertainty associated with estimating each parameter of interest, which propagates across into the estimation of net benefit itself. Hence, all that remains is to create a node in WinBUGS to estimate the net-benefit for the risk level(s) of interest. The generic equation is

Net benefit = (risk × relative reduction in risk of stroke) – (risk of fatal bleed × outcome ratio).

An absolute theoretical level of risk, or a level estimated with uncertainty such as those described above for clinical and echocardiographic risk factors can be specified. In this example the net benefit is evaluated on a stroke equivalent scale. The annotated WinBUGS code used for the whole analysis is provided in Appendix B.

8.4 Results

A burn-in of 5,000 iterations followed by a further monitored 15,000 produced the following results.

8.4.1 Effectiveness of warfarin for preventing strokes

As to be expected, the synthesis of the warfarin v placebo trials for preventing strokes produced similar results to the previous meta-analysis. In this analysis the pooled estimate and confidence interval for the odds ratio are 0.77 (0.59 to 0.87) compared to 0.73 (0.57 to 0.85) reported previously. (Singer, 1993) (Note: the European trial was excluded from this analysis to allow direct comparison with the results of Glasziou and Irwig.) This leads to a relative risk reduction of 0.23 (0.13 to 0.41). A small amount of between study heterogeneity was estimated by the model (median for the between study variance is 0.02). The simulated posterior distributions for these parameters are provided in Figure 8.3.

8.4.2 Risk of a fatal haemorrhage when taking warfarin

The posterior distributions for the meta-analysis estimating the risk of a fatal hemorrhage including both the observational studies and the RCTs are described above. The analysis performed here is somewhat more sophisticated than the original, where a simple un-weighted average was computed, but the results are in broad agreement with the risk of a fatal bleed in one year being 0.52% (0.27% to 0.84%) from this model compared with 0.6%. There is considerable heterogeneity between the study results. Further details of these parameters are provided in Figure 8.4.

8.4.3 The trade-off between a stroke and a haemorrhage

Figure 8.5 displays the posterior probabilities of being in the 16 quality of life states after suffering a stroke. The small distributions indicate the posterior distributions for each of the 16 probabilities. This distribution is highly irregular with a large proportion of persons in good health (state 2), however quality of life is much reduced for considerable numbers of patients.

Figure 8.6 summarises the posterior distributions for the intermediate parameters required to estimate the ratio between outcomes. The proportions in each of the QoL states are translated into QoL scores from which the reduction in QoL distribution can

be estimated. From this the number of strokes equivalent to one fatal bleed can be evaluated. This distribution is clearly bi-modal with a large spike at around fifty and the body of the distribution much closer to zero. The large spike results from the large proportion of persons making an almost full recovery after stroke, and hence for whom the ratio is high. For persons who make poorer recoveries the ratio is much lower, hence the mass near zero.

Figure 8.3 Posterior densities for the effectiveness of warfarin for preventing strokes

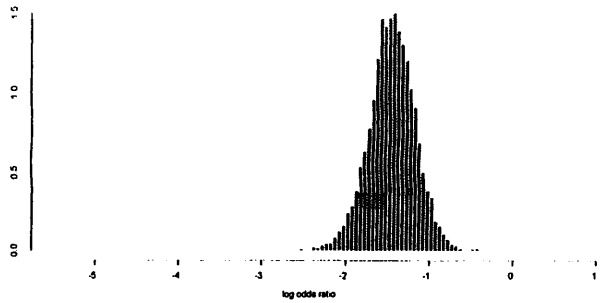
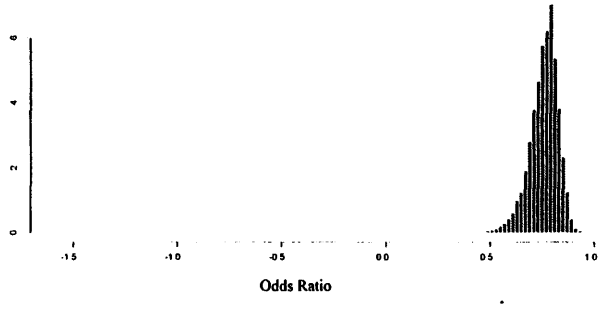
Parameter	Mean (s.e.)	Median (95% CrI)	Simulated PDF
Log odds ratio Warfarin v placebo	-1.45 (0.29)	-1.45 (-2.03 to –0.90)	
Odds ratio Warfarin v placebo	0.76 (0.08)	0.77 (0.59 to 0.87)	

Figure 8.3 Continued

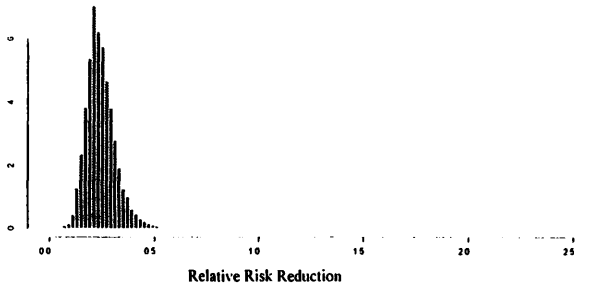
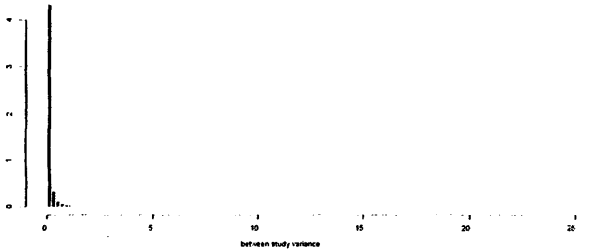
Relative risk reduction Warfarin v placebo	0.24 (0.08)	0.23 (0.13 to 0.41)	
Between study variance	0.13 (0.60)	0.02 (0.0007 to 0.86)	

Figure 8.4 Posterior densities for the risk of a fatal haemorrhage when taking warfarin

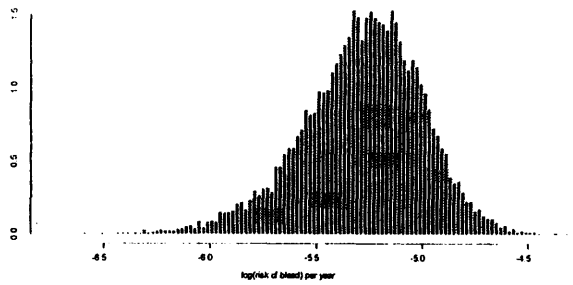
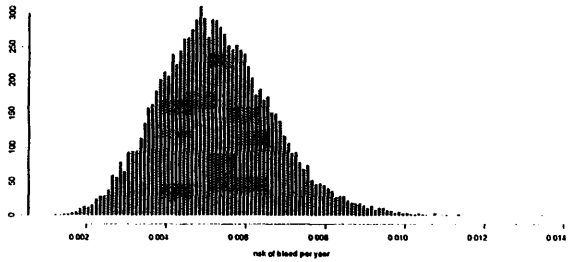
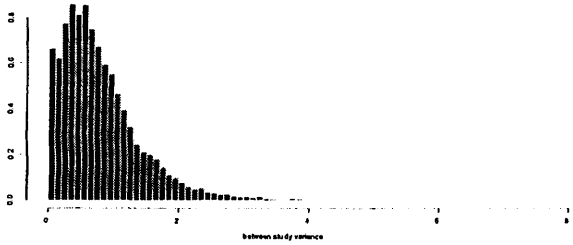
Parameter	Mean (s.e.)	Median (95% CrI)	Simulated PDF
Log risk of fatal bleed per year	-5.29 (0.29)	-5.27 (-5.92 to –4.77)	
Risk of fatal bleed per year	0.53% (0.15)	0.52% (0.27 to 0.84)	
Between study variance	0.8 (0.65)	0.66 (0.02 to 2.45)	

Figure 8.5 Posterior multinomial distribution for proportion of patients in each of the 16 quality of life states

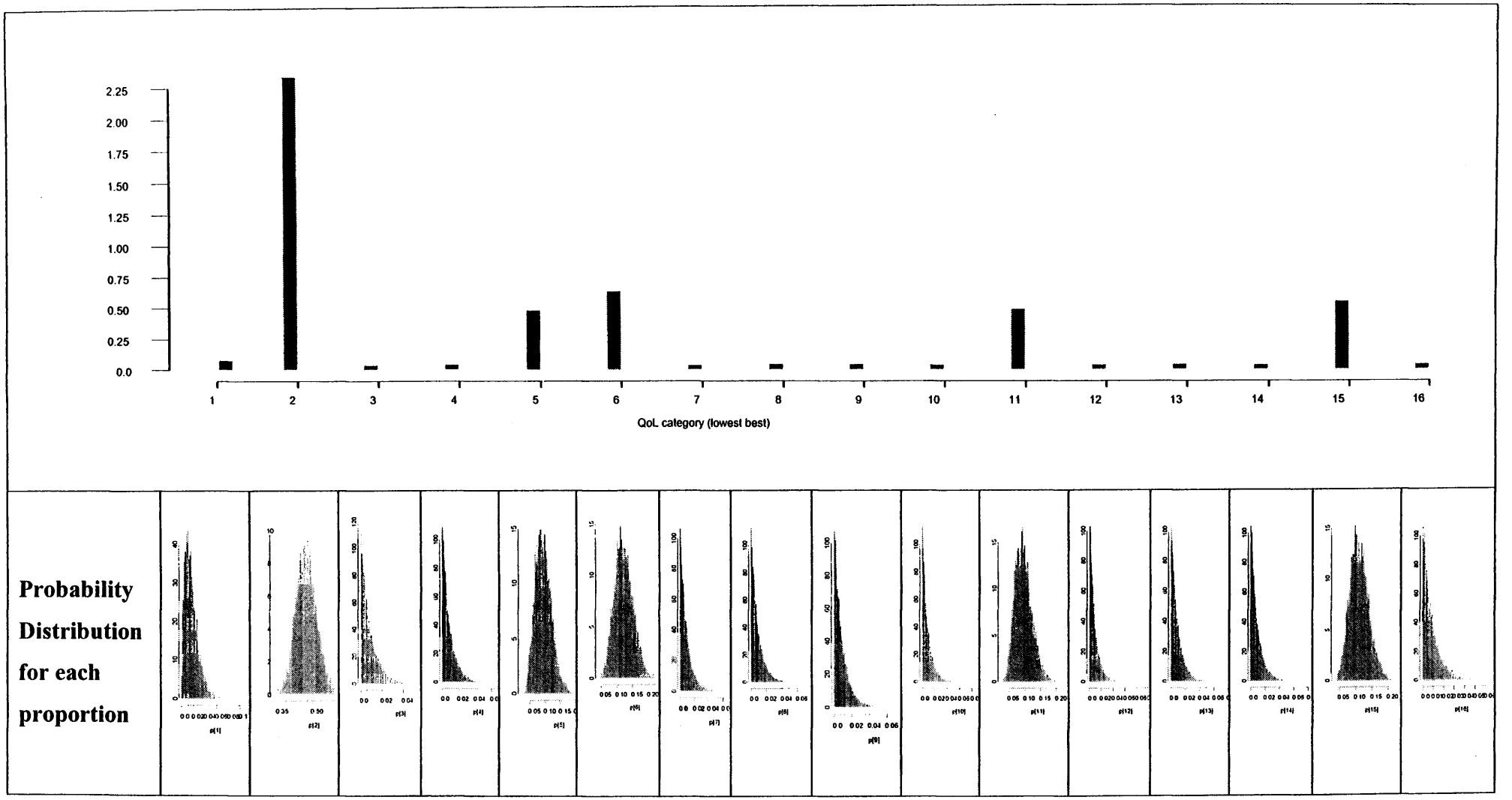


Table 8.6 Posterior distributions for quality of life variables

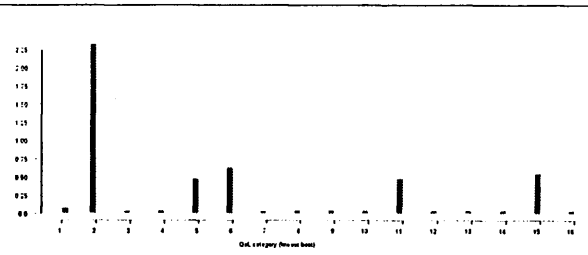
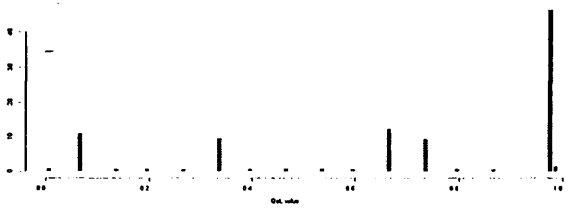
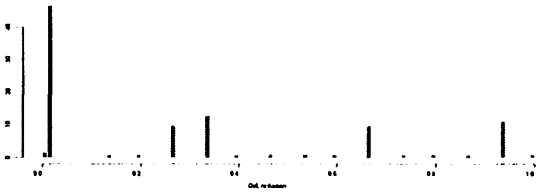
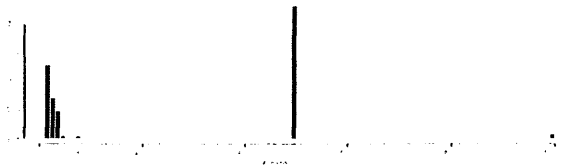
Parameter	*Mean (s.e.)*	*Median (95% CrI)*	*Simulated PDF*
QoL category	5.75 (4.67)	5.0 (2.0 to 15.0)	
QoL score	0.70 (0.33)	0.73 (0.07 to 0.98)	
QoL reduction	0.30 (0.33)	0.27 (0.02 to 0.93)	
Outcome ratio	26.14 (25.40)	3.75 (1.07 to 50)	

Figure 8.7 Net benefit for different levels of risk of stroke

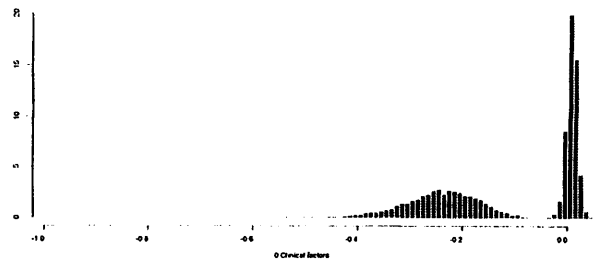
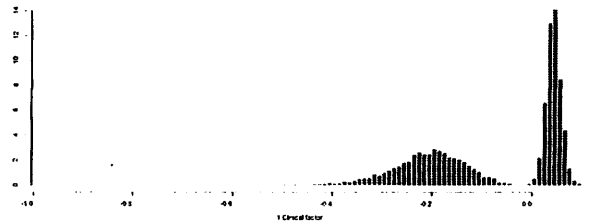
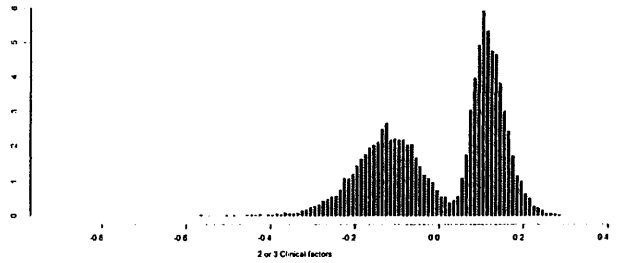
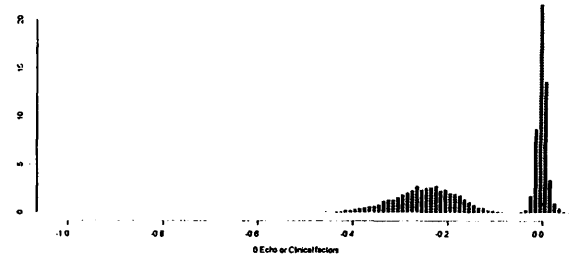
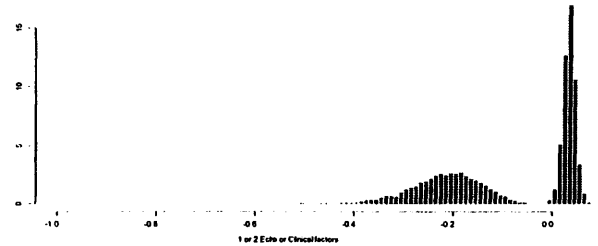
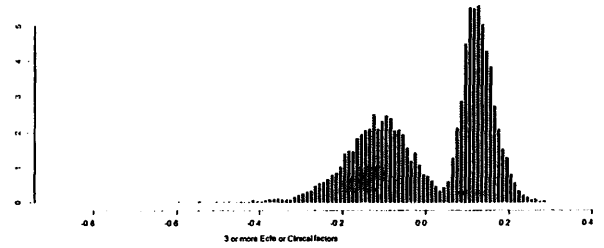
	Mean (s.e.)	*Median (95% CrI)*	*Probability of Benefit > 0*	*Simulated PDF*
No. Clinical risk factors				
0	-0.12 (0.14)	-0.01 (-0.39 to 0.02)	40.8 %	
1	-0.08 (0.14)	0.02 (-0.36 to 0.07)	51.6 %	
2 or 3	-0.0004 (0.15)	0.06 (-0.29 to 0.20)	54.2 %	

Figure 8.7 Continued

No. Clinical risk factors + echocardiographic features				
0	-0.13 (0.14)	-0.02 (-0.40 to 0.02)	19.0 %	
1 or 2	-0.09 (0.14)	0.01 (-0.37 to 0.05)	51.3 %	
≥ 3	0.01 (0.15)	0.07 (-0.28 to 0.20)	54.8 %	

8.4.4 Estimation of net-benefit

Figure 8.7 describes the posterior distributions of net-benefit for different risks based on clinical and clinical and echocardiographic risk factors. The bi-modality of the outcome ratio distribution propagates over into these net-benefit distributions making them highly irregular. This highlights the point that if other computational methods, which assumed more regular distributions for model parameters, had been used potentially misleading results could be obtained. Spiegelhalter et al. (Spiegelhalter et al. 2000) when revisiting the confidence profile method using WinBUGS have demonstrated this issue recently.

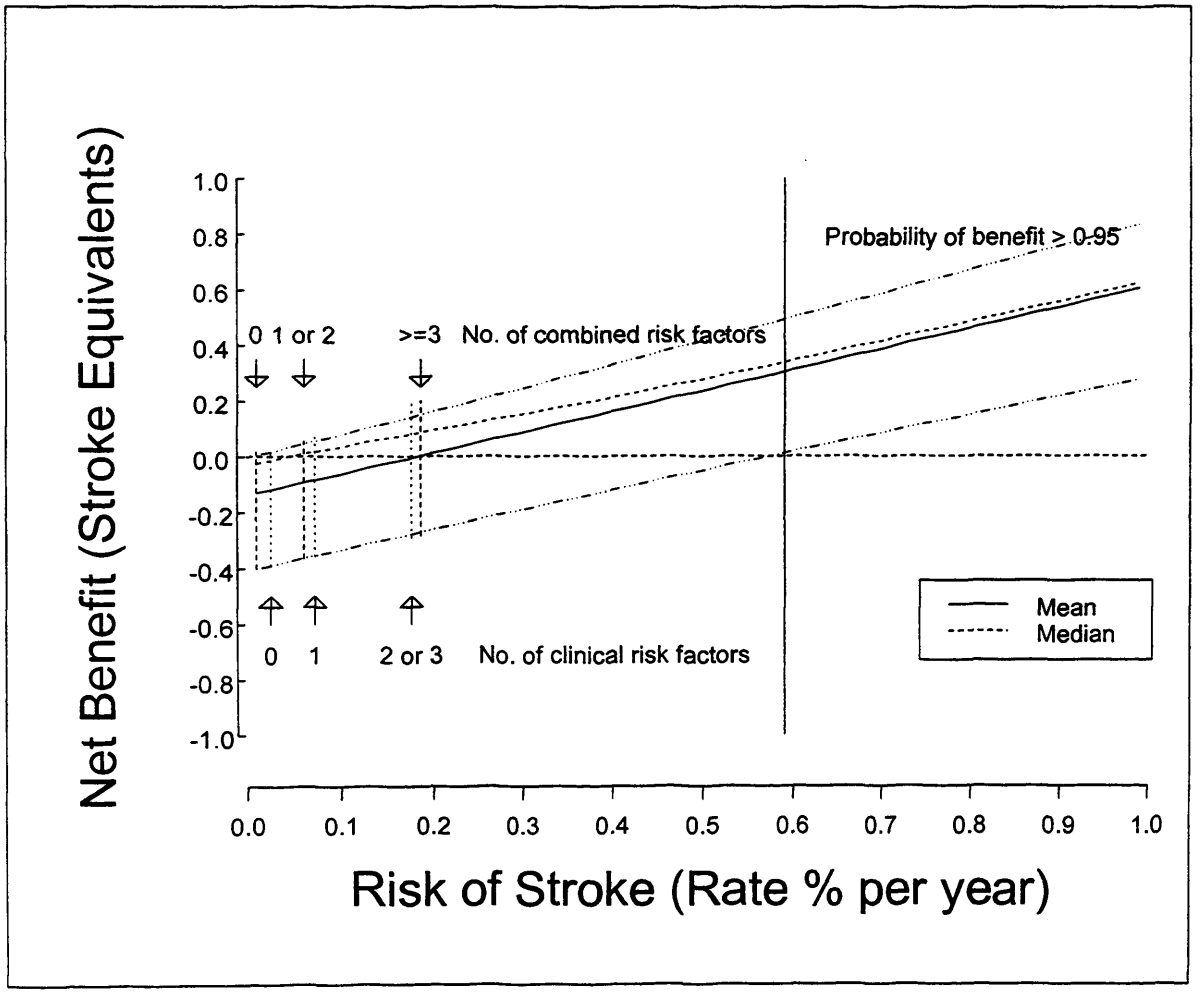
For the clinical risk factors, the surprising result is obtained that even if 2 or 3 factors are present the mean net benefit is negative. However, when at least one clinical risk factor is present, the model estimates that the majority (i.e. >50%) of the population will benefit from warfarin treatment. This can be explained as follows. For the proportion of persons who would have a bad outcome after stroke (i.e. much reduced QoL), the potential benefit of taking warfarin outweighs the risks (leading to the right bell of the distribution), but for the large proportion who would make a nearly full recovery after stroke, taking warfarin, on average, does more harm than good (leading to the left bell of the distribution). Since around 50% of persons have very high QoL after a stroke and 50% have a lower QoL, the overall net benefit is critically dependent on the empirical magnitude in stroke equivalents assigned through the outcome ratio distribution. In this situation, since the aim is to maximise QoL collectively, warfarin should not be given because the mean net-benefit is negative. If data were available that could help identify patients likely to make a good recovery from a stroke, if they were to have one, this would be clinically valuable.

Similar findings are observed for the clinical and echocardiographic risk factors. If zero risk factors are present, the model clearly indicates a negative net benefit. For one or two risk factors, the majority of persons (51.3%) would have a positive benefit, while the mean net-benefit is negative. For three or more risk factors both the majority of patients benefit, and the mean benefit is positive suggesting that for these persons warfarin should be given. However, the 95% credible interval for this and the other

estimates all contain the value zero, so in fact the benefit of warfarin is inconclusive for all risk factors.

In Figure 8.8 the mean and median net benefit line together with accompanying 95% credible intervals are plotted for hypothetical values of risk. As the risk increases the mean and median converge as an increasing majority of patients move into the right-hand bell of the bimodal distribution. The credible intervals for the different risk factor groups are included on this graph; these are fractionally wider than the intervals plotted for the theoretical levels of risk. This is because the absolute levels of risk corresponding to these risk factors are estimated with uncertainty and this uncertainty is reflected in the length of the intervals. The point where the probability of net benefit being positive is greater than 95% is around a risk of stroke of 58% a year, a much higher risk than the cut off for the highest risk groups considered.

Figure 8.8 Net Benefit fully modelling QOL data



8.5 Sensitivity analysis - considering other values for the outcome ratio

Clearly the results of this model are highly dependent on the distribution or value assigned to the outcome ratio. In the original analysis the mean QoL for persons after stroke was reported as being between 0.7 and 0.8 leading to a reduction in QoL of approximately 0.25, and so to an outcome ratio of four. Although it is explained above why such reasoning is flawed, this value maybe used instead of the distribution specified for the outcome ratio to assess its influence on the results.

Figure 8.9 provides the posterior net-benefit distributions for each of the risk factors groups described previously. The first point to note is that since the outcome ratio is specified as a single known number, the posterior distributions are uni-modal (although not symmetric). The results are very different from those calculated previously. For the clinical risk factors, when zero factors are present the mean net benefit is negative, although the credible interval contains zero. When one or more clinical risk factor is present however, net benefit is positive and the credible interval does not include zero, suggesting a clear benefit of warfarin. Similar findings exist for the clinical and echocardiographic risk factors, where the presence of one or more factors leads to a positive net benefit with a 95% credible interval that does not contain zero.

Figure 8.10 plots the mean net benefit and corresponding 95% credible interval over theoretical values of the risk of stroke. When comparing Figures 8.7 and 8.9 the most striking difference is the width of the credibility intervals. Not accounting for the estimation error in the outcome ratio has greatly (and artificially) reduced the uncertainty in the net-benefit estimate. When this is coupled with the incorrectly low value for the mean outcome ratio, very misleading results that are at odds with the analysis above have been produced.

Figure 8.9 Net benefit for different levels of risk of stroke – specifying outcome ratio to be known as four

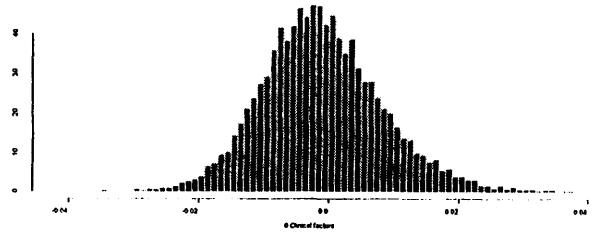
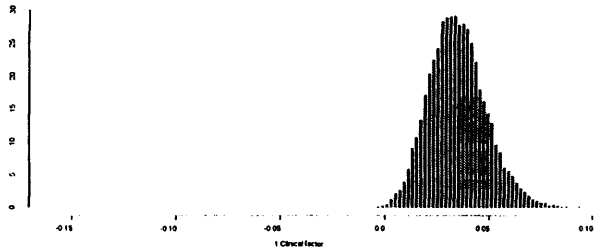
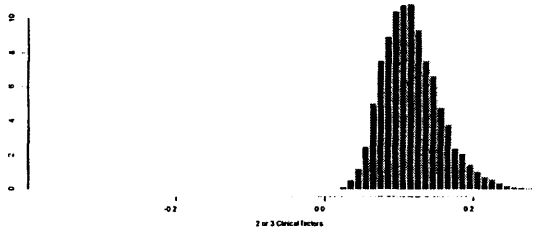
	Mean (s.e.)	Median (95% CrI)	Probability of Benefit > 0	Simulated PDF
No. Clinical risk factors				
0	-0.001 (0.009)	-0.002 (-0.02 to 0.02)	42.7%	
1	0.03 (0.014)	0.03 (0.01 to 0.06)	99.71%	
2 or 3	0.12 (0.04)	0.11 (0.05 to 0.21)	99.96%	

Figure 8.9 Continued

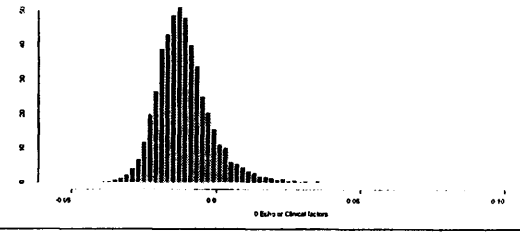
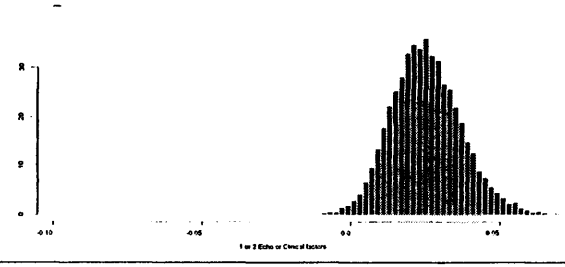
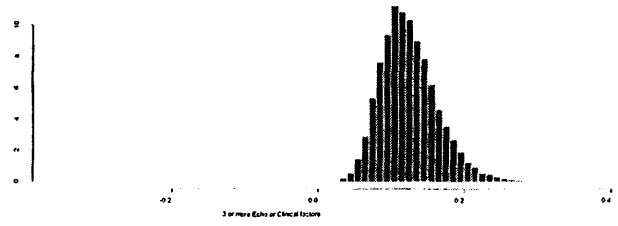
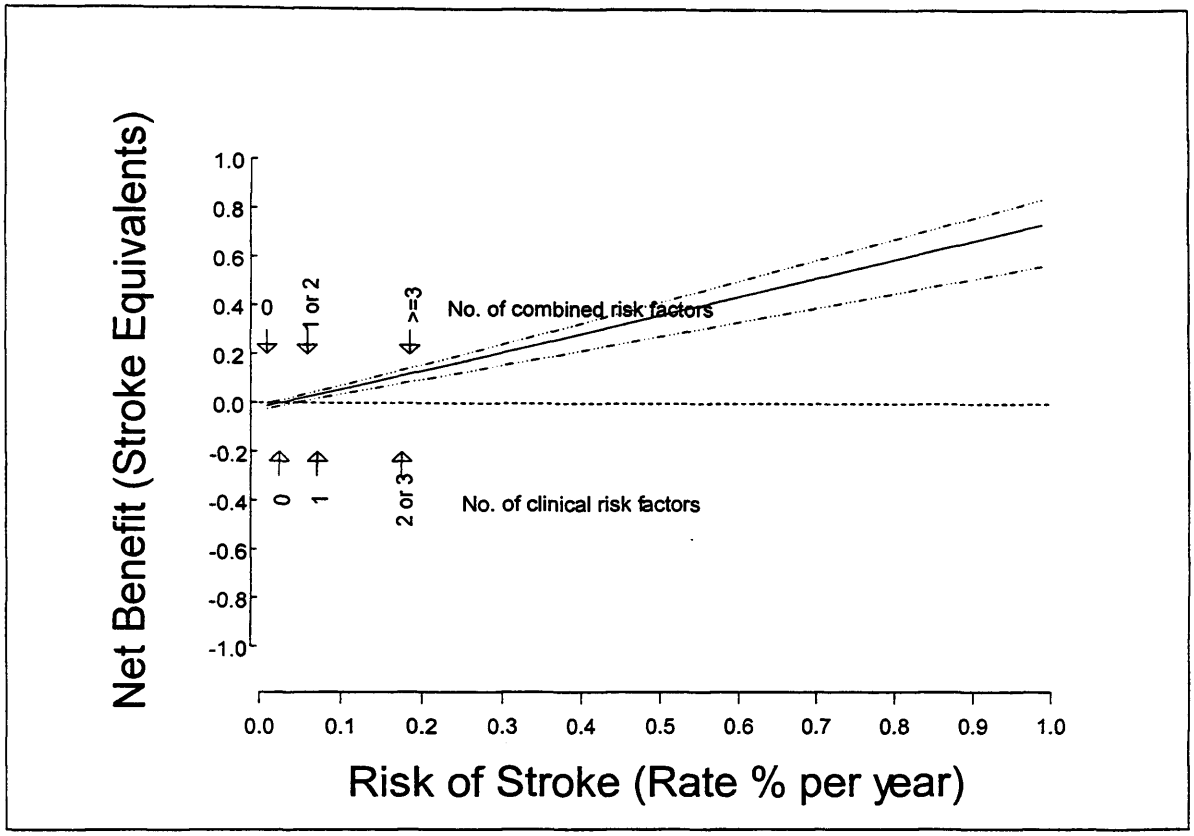
No. Clinical risk factors + echocardiographic features				
0	-0.01 (0.01)	-0.01 (-0.03 to 0.01)	11.6%	
1 or 2	0.03 (0.01)	0.02 (0.004 to 0.05)	98.8%	
≥ 3	0.12 (0.04)	0.12 (0.06 to 0.21)	99.97%	

Figure 8.10 Net Benefit fixing the outcome ratio at four

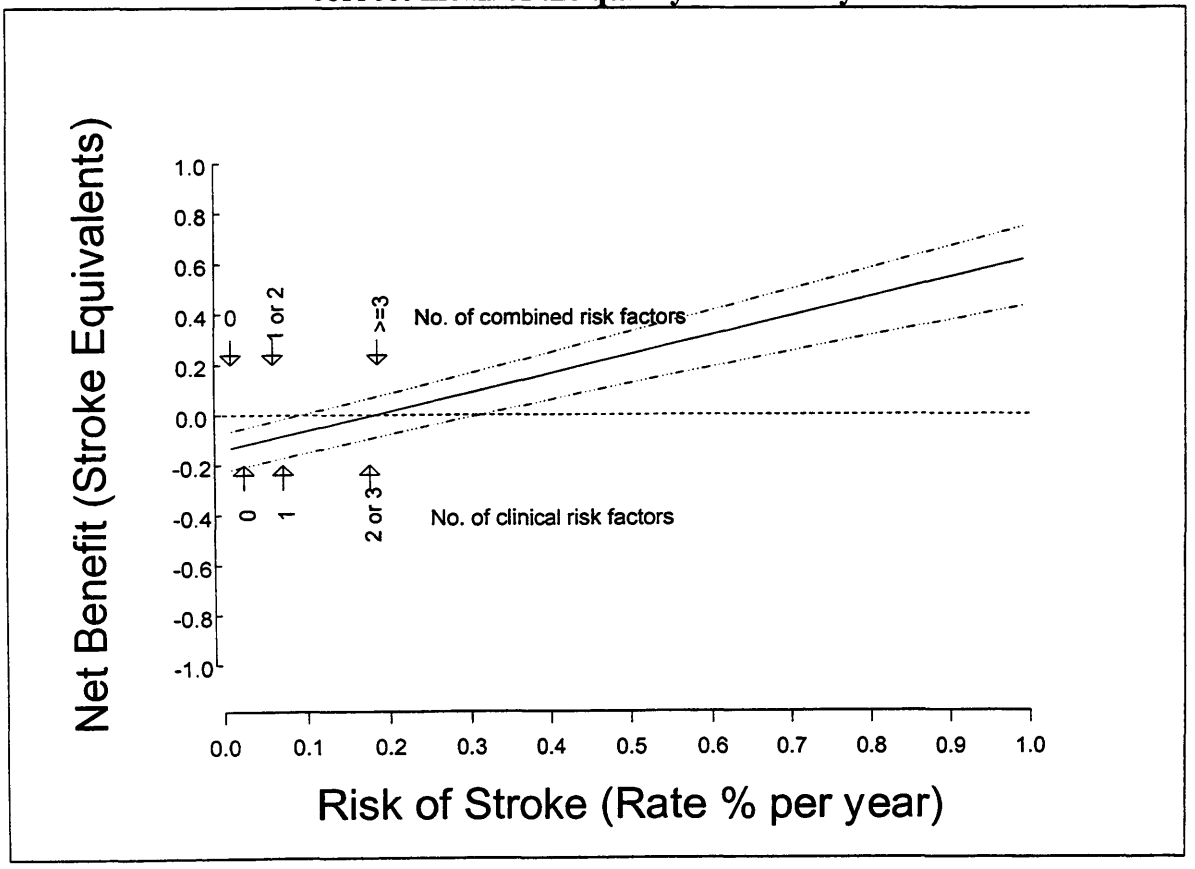


Finally, the single value 26.14 is used for the outcome ratio. This is the mean value of the outcome ratio distribution used in the primary analysis. By including this, the degree by which the results change due to not including the distribution of the outcome ratio can be assessed. The summaries of the posterior distributions for each of the risk factor groups are given in Table 8.4 Interestingly these results are quite different from those of Figure 8.6. Here, only when three or more clinical or echocardiographic risk factors are present is the mean and median net benefit positive, although the credibility interval still includes zero. However unlike any of the previous models, in the two lowest risk groups for clinical and clinical and echocardiographic risk factors, the net benefit is actually statistically significantly harmful (i.e. the 95% Credible interval does not contain zero). Figure 8.11 plots the mean net-benefit and corresponding 95% credible interval over theoretical values of the risk of stroke.

Table 8.4 Net benefit for different levels of risk of stroke – specifying outcome ratio to be known as 26.14

	Mean (s.e.)	*Median (95% CrI)*	*Probability of Benefit > 0*
No. Clinical risk factors			
0	-0.12 (0.04)	-0.12 (-0.21 to –0.05)	0.007%
1	-0.09 (0.04)	-0.08 (-0.17 to –0.01)	0.99%
2 or 3	-0.003 (0.06)	-0.005 (-0.11 to 0.11)	46.7%
No. Clinical risk factors + echocardiographic features			
0	-0.13 (0.04)	-0.13 (-0.22 to -0.06)	0.04%
1 or 2	-0.10 (0.04)	-0.09 (-0.18 to –0.02)	0.41%
≥ 3	0.004 (0.06)	0.003 (-0.10 to 0.12)	52.5%

Figure 8.11 Net benefit fixing outcome ratio at 26.86 – the value derived from the correct mean of the quality of life analysis



These results illustrate that the estimation of net benefit is very sensitive to the QoL values placed on the stroke event, and hence on the distribution or value placed on the outcome ratio. Using a single value rather than the distribution including measurement error changed the conclusions of the analysis, as did using alternative values for the outcome ratio. Thus, if this type of analysis is to be used to inform clinical practice, then great caution is required when specifying the outcome ratio. This issue of the validity of QoL measures such as the one used here, and their critical importance in the net benefit calculation is considered further in Section 8.6.

8.6 Discussion/ further work

In this chapter, a fully stochastic statistical approach to the net benefit model has been outlined in general terms, and applied to the issue of the benefit of giving atrial fibrillation patients warfarin to prevent stroke. In order to make this possible, quantification of the evidence regarding a) the efficacy of warfarin, b) risk of a fatal bleed due to warfarin use, c) the relative reduction in QoL for the competing outcomes, and d) the risk of a stroke in various definable patient subgroups, was necessary. This type of model should not be confused with the net-benefit model recently described in health economics literature (Stinnet and Mullahy, 1998) to evaluate whether an intervention represents good value for money, although the two have similarities, and the possibility of including cost data in the model described here is discussed below.

In applying the model an unjustified assumption in the reasoning regarding the quantification of quality of life decrease after stroke in the original application of this method was identified. (Glasziou and Irwig, 1995) This leads to a serious overestimation in terms of the mean reduction in QoL following a stroke. Sensitivity analyses demonstrated that the results of the model are highly sensitive to the quantification of this parameter. Indeed, the equating of how many strokes are 'equivalent' to a fatal bleed is probably the most contentious part of the model, and open to most criticism. Nonetheless, it can be argued that, whether done through a statistical model, or in a less quantitative or formal way, the potential benefits and harm

need "weighing-up" if a rational decision is to be made. However, it would seem appropriate to recommend that further work be required to establish how different outcomes can most reliably be equated on the same scale (if at all).

It is important to note that only (an estimated) one hundred patients were used to assess the quality of life following a stroke, this is many fewer patients than were used in either the assessment of the efficacy of warfarin at preventing strokes, or the risk of a fatal bleed due to taking warfarin. Probably the most effective way of reducing the uncertainty in the model parameters of interest would be to include more QoL data in the model. The prospect of prospectively planning a net-benefit analysis raises some interesting methodological issues regarding optimal research resource allocation. In this analysis, data from four different types of studies were used – those estimating a) the efficacy of the intervention, b) the rate of side effects, c) the relative reduction in QoL for the competing outcomes, and d) the levels of risk for different subgroups of the population potentially eligible for treatment. Designing and allocating resources optimally to these four study types would be a complex problem, well beyond the issues considered currently when designing a single study, or even a prospective series of similar studies to be combined using meta-analysis. (Margitic et al. 1995) A decision theoretic approach utilising assessments such as the value of perfect information and value of partial information could be very valuable in prospectively planning analyses such as this and informing the most efficient way of allocating future resources. (Claxton, 1999)

Although a complete analysis using the data considered in the original application is presented here, several extensions may be desirable. Firstly a sixth trial of warfarin for atrial fibrillation that was not included in the original meta-analysis has now been published (EAFT (European Atrial Fibrillation Trial) Study Group, 1993) and should be included in the meta-analysis. Secondly, the quality of life assessment reported on those persons following stroke ignores those persons who die due to their stroke. Clearly this means the quality of life results will look too optimistic. Two potential solutions to this problem are i) add complexity to the model to allow for this group of patients; or b) include an estimate of the six month mortality from stroke in the quality of life analysis

by including a proportion (approximately 10%[1]) with a quality of life of 0. Thirdly, the five trials included were all stopped early due to a clear benefit in the treatment arms in all cases, and there is a potential for the treatment effect to be overestimated because of this. (Green et al. 1987; Hughes et al. 1992) Further, adverse event data were only included from three of the six RCTs, the model could easily be extended to include adverse event data from all these trials. The model has been somewhat simplified to consider only the two most serious of a number of outcomes. Serious (but non fatal) and more minor bleeds are relatively frequent outcomes for those taking warfarin, and perhaps would have sufficient impact on quality of life measures to have an impact on net-benefit; hence these further adverse events should be considered. The problems of measuring QoL and the questionable validity of the approach used has already been commented on. There is no one standard way of measuring QoL. For this analysis the results of a time trade-off analysis were used, although the York Health Measurement Questionnaire was also used in the original assessment. (Glasziou et al. 1994) Clearly, different results would be obtained if these results were also included or used instead of the time trade-off data.

Another area where there are potential improvements to be made to the model is the meta-analysis estimating the risk of having a fatal bleed. Included in this analysis were both observational and randomised studies in different patient groups, only a proportion of which had atrial fibrillation. No allowances were made for such heterogeneous study designs and populations beyond the inclusion of a random effect. It would be possible to use a hierarchical model to stratify studies by design and medical condition, using a similar approach to the combination of the Cholesterol studies in Chapter 5. If this were done, perhaps the most realistic estimate of the risk of harm would be gained from the shrunken estimate of the observational atrial fibrillation studies, having borrowed strength from the other subgroups.

Bayesian methods have been used to implement the analysis presented here using MCMC methods. Such an approach offers the flexibility required to fit all the components of this non-standard model, and with the exception of a simplified model being possible using Monte Carlo methods, it is difficult to see how such an analysis

[1] Personal communication with Paul Glasziou

could be possible using Classical methods. All priors were specified as non-informative, although the possibility to incorporate further evidence, especially regarding the quality of life issues, via these distributions should not be ignored. However the elicitation and construction of meaningful priors is a difficult area that requires further research. A further issue regarding the Bayesian methodology used is that the credible intervals were constructed using the 2.5 and 97.5 centiles of the distribution. More appropriate would be to define the region of highest posterior density. Usually there is little difference between these two measures, but with the distinct bi-modality of many of the posteriors considered here differences will exist.

Evaluations of the use of warfarin for atrial fibrillation have recently been carried out by others using distinctly different methodology. (Li et al. 1998; Thompson et al. 2000) Li et al. (Li et al. 1998) describe a method to estimate the absolute estimated benefit for individual patients. This approach uses a database of 35 000 individuals to which a treatment-stratified Cox model is fitted including the principal risk factors and treatment interaction terms. New statistical methodology is proposed to produce a prediction interval for individual patients using a combination of Monte Carlo and Bootstrap methodologies. Their model only considers overall survival rates, and hence does not attempt to equate or model the different outcomes. A further difference between their model and the net benefit one is that the risk of harm (a fatal bleed) for individuals is not assumed to be constant, but rather dependent on patient characteristics. Further, while this method examines a very large database, information from other sources, such as RCTs etc is not considered. Perhaps there is scope for an amalgamation of this and the net benefit methodology allowing a more detailed classification of risk factors for individuals, rather than the classification of individuals into one of the six groups considered above, and allowing the risk of fatal bleeds to also vary between individuals. Further, the implementation of this method in a Bayesian framework would allow the construction of prediction intervals for individuals, and perhaps be simpler to implement than the Monte Carlo and Bootstrap methodologies they propose.

Thompson et al. (Thompson et al. 2000) implemented a full decision analysis model to the same problem, and are similar in spirit to the net benefit model, but going one stage further by incorporating cost data. Very similar information is included in the model, namely a) the RCT data on the efficacy of warfarin to prevent stroke, b) studies

reporting the absolute risk of stroke for different risk factors, c) studies assessing the risk of adverse events in atrial fibrillation patients treated with warfarin, d) QoL data relating to the various possible outcomes, including that from their own study, and e) cost data relating to the various outcomes, including that from their own study. A Markov model is developed and evaluated by simulating a cohort of patients and recording their passage through the model, and hence is different conceptually to the approach taken by the net benefit model. Since their results report more specific patient groups than the net benefit model above, it is difficult to compare the results. However, it would be very interesting to extend the net-benefit model to include cost data, and then use the same data in both models to compare the results obtained from the two methodologies. It is not obvious what the relationship between estimates from the two models would be, but the Markov model is clearly more complex. Such an assessment may help inform if the complexity of such methodology is necessary in clinical decision-making areas such as this, or simpler methods such as net benefit are adequate.

Finally, the application of the model is not restricted to drug treatments. In many public health interventions there are benefits and drawbacks of a new policy. For example, the debate has run for many years concerning the pros and cons of artificially fluoridating drinking water. While there is evidence that doing so prevents tooth decay, adverse side effects such as the increase in prevalence of fluorosis exist. (McDonagh et al. 2000) A modification of the net benefit model to estimate the optimal level of fluoride that should be added to drinking water to maximise benefit and minimise harm would be a valuable exercise. In order to do this the quantification of the relative harm of tooth decay and fluorosis would be required. A further potential example is the enforced use of bicycle helmets. This is another debate which has been running for many years, and while there is evidence of protection in terms of head injuries, negative effects such as strangulation risk and increased prevalence of heart disease due to lower uptake of cycling are evident. (Thompson et al. 1999; Robinson, 1996)

In conclusion, an integrated approach to evaluating net benefit has been described using Bayesian MCMC methodology. This allows distributions representing the uncertainty in all the parameters required for the model to be estimated and utilised. Such a framework allows the synthesis of information from different studies with different designs reporting different but related information, permitting a very broad generalised

quantitative synthesis of evidence. Such a model is particularly appealing since it potentially aids in making the correct treatment decisions for individual patients; perhaps the ultimate goal of any modelling exercise such as this. However, there are a number of unresolved issues which remain and require further research before such a methodology can be recommended. Perhaps the most critical of these is the use of QoL data, as the method has been shown to be highly critical to the data used, and questions regarding the validity of such data, and the most appropriate way to model it, remain.

Chapter 9 Discussion, including lines of further work

9.1 Summary

This thesis comprises an investigation into methods for and the feasibility of combining information from different sources, with emphasis on the assessment of the effectiveness of an intervention in a health context. Bayesian modelling employing MCMC methods is found to be an appealing and powerful approach for achieving this as it allows complex models to be constructed with relative ease compared to the classical alternative. Such models allow more sources of data to be included in an analysis, hopefully modelled in a more appropriate way than a standard meta-analysis allows. However, married with these benefits and added flexibility is the need to think carefully about the specific model structure for a particular application, as often models will need to be custom built. Further, issues related to publication bias and differential quality of the primary studies remain, and, as discussed in Chapter 6 and below, may be an even larger problem in the synthesis of data from multiple study designs than the single design usually considered in a 'traditional' meta-analysis of data from a single type of study. Hence, although there are clear benefits of generalised syntheses of evidence, they are not automatic. The difficulty of sensibly combining sources of information from disparate sources should not be underestimated, and great care is needed when carrying out such analyses.

9.2 Application of Bayesian MCMC methods to the generalised synthesis of evidence

Bayesian MCMC methods have been used extensively in this thesis. This is because their implementation in the WinBUGS program creates an environment which reduces the restrictions on the structures of the models that can be fitted, compared with all classical alternatives.

Often when Bayesian methods are discussed, it is the necessity to place prior distributions on all unknown parameters enabling subjective beliefs it be incorporated

into the analysis that receives most attention. This aspect of the analysis was not the focus of the work in this thesis and "off the shelf" vague priors were used most of the time. However, when there is little data it has been shown (see Sections 5.2 and 7.4 and (Lambert et al. 2001b)) that commonly conceived 'vague' priors may actually be overtly informative. This is particularly true for variance components when the number of units in the analysis was small, or event data was very sparse. The author is currently involved with further research to address this issue.

This is not to say that informative priors do not have a role in the generalised synthesis of evidence. On the contrary, carefully constructing informative priors may expand the breadth of relevant evidence that can be synthesised. Examples of additional information that may exist and may be able to be formulated in terms of priors for a general synthesis application include: a) expert opinion on one or more parameters in the model; b) empirical data of relevance but not included 'directly' in the model (examples include empirical estimates of between study variance parameters derived from previous meta-analyses (Higgins and Whitehead, 1996) or the likely effects of the magnitude of an intervention derived from studies of the intervention in a different disease or patient group); and c) information from studies generating only qualitative evidence. (Dixon-Woods et al. 2001) The generation of such informative priors is far from automatic, and more work is certainly needed in this area generally, (Chaloner and Rhame, 2001; O'Hagan, 1998) as well as specifically in a generalised synthesis context.

The benefits of the flexibility of Bayesian MCMC methods is less well publicised and appreciated by those not working directly in the field than issues relating to prior distributions. The 'LEGO brick' or 'Meccano' type construction offered by WinBUGS allows not only models not available in classical statistical packages to be constructed but also multiple sub-models to be 'bolted' together to form one comprehensive overall 'meta'-model to be constructed in which parameter uncertainty is correctly propagated and correlations between parameters accounted for. This is particularly striking in Chapter 8 where several analyses contribute to constructing a node for the net-benefit expression. This idea has been appreciated by others, possibly the first of whom is Eddy et al., (Eddy et al. 1992) who used similar ideas in the construction of the confidence profile method in the early 1990s.

As mention above, a generalised model will often have to be custom built for a particular application. Perhaps the way to start generating a coherent and cohesive strategy for building a generalised synthesis analysis framework is to identify and document commonly used 'building blocks' in the analysis. For example, a standard random effects meta-analysis on the odds ratio scale (Section 2.3.1), the three level hierarchical model used to combine multiple study designs (Section 3.8.1) and additional code required to include indirect comparisons of treatment effect (Section 5.4.2) could all be thought of as sub-models that could be 'bolted together' as required to build up the complete full model. In a sense Eddy in his book (Eddy et al. 1992) began to do this, but uptake was minimal for reasons discussed in Section 3.6. However with the advent of the WinBUGS software, the climate may now be more amenable. An indicator of this is the success of a recent book by Congdon (Congdon 2001) documenting how common (and not so common!), originally classically derived, statistical models can be fitted in WinBUGS. Perhaps a similar more focused volume, describing 'building blocks' for generalised synthesis of evidence together with examples of how they can be 'bolted' together to form complete analysis would be a good way of raising the profile of generalised synthesis. However, I believe before this could be achieved successfully, more applications, like those described in this thesis, need to be worked through as such methods are still in their infancy.

9.3 Hierarchical models

Hierarchical modelling is now an established method of analysis when observations are clustered into larger units within a dataset. A standard random-effect meta-analysis model is a simple hierarchical model with two levels. The extension of this model to a third level to incorporate study type, originally described by Prevost et al. (Prevost et al. 2000) is used extensively in this thesis (with discussion of its application to the cholesterol lowering data provided in section 5.2.13), with modifications to include adjustment for patients underlying risk and the synthesis of direct and indirect evidence.

This approach to synthesising evidence is appealing because it accounts for between study type heterogeneity as well as between study within study type heterogeneity. However, it would appear this approach has one major setback when the number of

different study types is small. This is that the between study type variance parameter is so poorly estimated, due the lack of data, that its value is heavily influenced by the prior placed on it, and the uncertainty associated with it so high that, even if study types are in broad agreement, uncertainty in the overall pooled estimate will inevitably be very high. This means that including increased numbers of sources of evidence is desirable, either by adding different study designs or perhaps by subdividing existing study designs to create more study level units of analysis is desirable, if a precise estimate is required. Alternatively, an informative prior distribution could deliberately be placed on the between study type variance parameter. Further research is clearly required on these issues.

Finally, only aggregated study level data has been considered in this thesis. Individual patient data (IPD) meta-analyses are carried out, although there is some debate at to whether the extra effort required is justified. Recently it has been shown that IPD regression on patient level covariates is much more powerful than the aggregated meta-regression equivalent. (Lambert et al. 2001a) Methods are becoming available to synthesise IPD and summary level data (Turner et al. 2000; Whitehead et al. 2001; Higgins et al. 2001). Clearly carrying out a generalised synthesis of evidence using IPD is possible, and such extensions of meta-analysis models potentially have a very valuable role to play in a generalised synthesis framework. Research into methods for doing so, and into when the extra effort is worthwhile, are required.

9.4 Sparse data/rare events

When work that has formed this thesis was commenced in 1996, any emphasis in it on rare events, or sparse event data was not perceived. However, as more examples were considered it became clear that quantifying important, but rare, events was an area where generalised synthesis of evidence had a lot to offer. The three data sets considered in Chapters 6 and 7 all include sparse event data. It was not clear what the best method of meta-analysing such data was and Chapter 7 illustrates that results can differ considerably depending on the analysis method used. Although Chapter 7 poses more questions than it answers, it has formed the basis for further work to ascertain the relative merits of competing meta-analysis methods. The first stage of this work is now

complete (Sweeting, 2001) and it would seem MCMC methods offer several advantages over classical methods for combining such data.

There is a lot of scope for further applications of generalised synthesis to rare events, particularly in the analysis and monitoring of adverse events related to drugs.

9.5 Threats to the validity/feasibility of generalised synthesis

The subsections below consider what are perceived as the main validity/feasibility threats to the application of generalised synthesis.

9.5.1 Procedural issues

The procedural methodological aspects of carrying out a generalised synthesis of evidence have not been a focus of this thesis. Nonetheless, issues relating to literature search strategies, data extraction protocols and so on, require as careful consideration in a generalised synthesis framework as they do for a meta-analysis. However, such methods are nowhere near as well developed for generalised synthesis as they are for standard meta-analyses/systematic reviews.

From carrying out the analysis of the cholesterol data described in Chapters 4 and 5 it became clear how difficult it was to define, and locate all the evidence of interest. Indeed, to make the analysis feasible in the time available only a fraction of the total observational evidence was considered. It is well known that carrying out a meta-analysis is a time consuming task, increasing the remit of a review by including the observational evidence could make this job much longer still.

As noted above, currently there is considerable discussion in the meta-analysis literature of when the analysis of individual patient data is desirable and justified in relation to the extra time and cost required. Similarly, in which instances the extra effort in including extra sources of data in a generalised synthesis framework is worthwhile needs clarification. Although a difficult question to answer, further consideration of this issue may help develop guidelines for good generalised synthesis practice.

9.5.2 Publication bias

Chapter 6 highlighted the potential problem of publication bias in a generalised synthesis framework and the fact that the degree of bias may be different for different study types. To further inform how large a problem this is, it would be valuable to carry out assessments of publication bias in subject areas where both RCTs and observational evidence existed to assess whether differential levels of bias existed between study types.

Chapter 6 also considered methods of dealing with publication bias in a generalised synthesis framework. It would appear that methods for dealing with publication bias in meta-analysis of a single type of study can be transferred relatively easily into a generalised synthesis framework. Since different sources of evidence may be susceptible to differing degrees of bias an assessment of publication bias would always seem desirable where possible.

A drawback of the methods used in this thesis to deal with publication bias is that they use classical statistical methods. It would be highly desirable for 'adjustment' methods that could be implemented in a Bayesian framework within WinBUGS to be developed. This would allow all analyses to be carried out in one coherent model. Although Bayesian methods to deal with publication bias have been developed (Givens et al. 1997), no one has yet worked out how to implement them within the WinBUGS program.

9.5.3 Study quality

It has often been said that a meta-analysis is only as good as the quality of data going into it. Exactly the same sentiments apply to the generalised synthesis of evidence. The importance of addressing the variable quality of studies included in a synthesis was flagged as early as the introduction to this thesis. It was also acknowledged there that this important issue would not form a focus of this thesis due to the enormity of the problem.

There is currently still dispute as to how to deal with variable quality of evidence within a single study design (see section 2.7). The problem is further compounded when multiple sources of evidence are being considered. This problem does not have an ideal statistical solution; rather actions can probably best be informed from empirical investigations into the reliability of estimates from different study designs, some of which were reviewed in Chapter 1. However, currently we do not have enough evidence to accurately inform which studies are likely to be reliable and which are not.

Taking a Bayesian approach to synthesis does allow a down-weighting of different sources of evidence through the use of informative prior distributions. Also, a sensitivity analysis plot was constructed in Chapter 7, which allowed the effect of down-weighting different sources of evidence to be shown graphically. An alternative approach to down weighting evidence suspected of being biased, is empirical adjustment for the likely effect of the bias. This was attempted in the Confidence Profile approach. However it has been shown, for RCTs at least, that even the direction, let alone the magnitude, of bias relating to certain study deficiencies may be difficult to predict. (Schulz et al. 1995)

Clearly more work is required in this area, but perhaps not from an essentially statistical viewpoint. Until (if ever) a clearer picture emerges, making assessments of quality and using them as the basis of a sensitivity analysis is perhaps the best approach that can be taken.

9.6 Generalised synthesis and beyond

No comprehensive definition of generalised synthesis has been offered in this thesis, only examples of it. This is deliberate as I am unsure of its boundaries and my perception of these has certainly changed over the duration of writing this thesis. Initially I came from a perspective of thinking about how to extend meta-analysis to incorporate multiple sources of evidence. I think of these as one-parameter primary interest models, as their aim is to estimate some quantity, often effectiveness of interventions, using multiple data sources. An example of this is the cholesterol analysis

in Chapter 5. Then, my focus widened to a broader range of models where data from multiple sources relating to multiple parameters was synthesised as an intermediate step to producing an estimate of the primary parameter of interest. An example of this in this thesis of this is the net-benefit model described in Chapter 8, although economic evaluations and other decision type models often used in healthcare would also come under this category.

Further, work I have been involved with outside this thesis has demonstrated how complex decision models, such as those with a Markov structure, can be constructed using WinBUGS, with meta-analysis sub-models included where evidence is available from more than one study to inform particular parameters required for the model. (Cooper et al. 2001) I find it hard to make distinction between such models and those described in this thesis. Hence, if I now had to define generalised synthesis of evidence, I would first paraphrase Glass (Glass, 1976) and define meta-analysis as the synthesis of information from multiple studies, then go on to define the generalised synthesis of evidence as the synthesis of information from multiple studies, including multiple designs which in some way model the different sources distinctly.

9.7 Conclusions

With the increase in computer power and development of software to fit complex models using Bayesian MCMC methodology, it is now possible to think beyond the models currently used to synthesise medical data. This thesis has presented examples illustrating how information from different sources can be combined together, while taking into account the level of uncertainty associated with the estimates. These methods do not take a radically different approach from or supersede previous methodology for synthesising data; rather they form a natural extension to more standard meta-analysis methods. I hope such efforts will be seen as first steps in a more widely evidence-based future where quantitative models are created routinely to summarise the totality of evidence, and to inform models used to help make decisions for future patients.

Bibliography

Abrams, K.R., Hellmich, M. and Jones, D.R. (1997) Bayesian approach to health care evidence. Technical Report 97-01, Department of Epidemiology and Public Health. University of Leicester.

Abrams, K.R., Lambert, P.C., Sanso, B., Shaw, C. and Marteau, T.M. (2000) Meta-analysis of heterogeneously reported study results - a Bayesian approach. In: Stangl, D.K. and Berry, D.A., (Eds.) *Meta-analysis in medicine and health policy*, Marcel Dekker: New York

Abrams, K.R. and Sanso, B. (1998) Approximate Bayesian inference in random effects meta-analysis. *Stat.Med.* **17**, 201-218.

Adams, D.C., Gurevitch, J. and Rosenberg, M.S. (1997) Resampling tests for meta-analysis of ecological data. *Ecology* **78**, 1277-1283.

Anello, C. and Fleiss, J.L. (1995) Exploratory or analytic meta-analysis: Should we distinguish between them? *J.Clin.Epidemiol.* **48**, 109-116.

Anonymous (1995a) Listening to your baby's heartbeat during labour. Midwives Information and Resource Service (MIDIRS) and the NHS Centre for Reviews and Dissemination. Bristol. *MIDIRS*

Anonymous (1995b) Magnesium, myocardial infarction, meta-analysis and megatrials. *Drug Ther Bull* **33**, 25-27.

Anonymous (1998) Cholesterol and coronary heart disease: screening and treatment. *Effective health care* **4**, 1-16.

Antczakbouckoms, A., Joshipura, K., Burdick, E. and Tulloch, J.F.C. (1993) Metaanalysis of surgical versus nonsurgical methods of treatment for periodontal-disease. *Journal Of Clinical Periodontology* **20**, 259-268.

Antman, E.M., Lau, J., Kupelnick, B., Mosteller, F. and Chalmers, T.C. (1992) A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts: Treatments for myocardial infarction. *J.A.M.A.* **268**, 240-248.

Austin, H., Perkins, L.L. and Martin, D.O. (1997) Estimating a relative risk across sparse case-control and follow-up studies: a method for meta-analysis. *Stat.Med.* **16**, 1005-1015.

Begg, C.B. (1994) Publication bias. In: Cooper, H. and Hedges, L.V., (Eds.) *The handbook of research synthesis*, pp. 399-310. Russell Sage Foundation: New York

Begg, C.B. and Berlin, J.A. (1989) Publication bias and dissemination of clinical research. *J.Natl.Cancer.Inst.* **81**, 107-115.

Begg, C.B. and Mazumdar, M. (1994) Operating characteristics of a rank correlation test for publication bias. *Biometrics* **50**, 1088-1101.

Begg, C.B. and Pilote, L. (1991) A model for incorporating historical controls into a meta-analysis. *Biometrics* **47**, 899-906.

Benson, B.A. and Hartz, A.J. (2000) A comparison of observational studies and rondomized controlled trials. *New England Journal of Medicine* **342**, 1878-1886.

Berard, A. and Bravo, G. (1998) Combining studies using effect sizes and quality scores: application to bone loss in postmenopausal women. *J.Clin.Epidemiol.* **51**, 801-807.

Berkey, C.S., Anderson, J.J. and Hoaglin, D.C. (1996) Multiple-outcome meta-analysis of clinical trials. *Stat.Med.* **15**, 537-557.

Berkey, C.S., Antczak-Bouckoms, A., Hoaglin, D.C., Mosteller, F. and Pihlstrom, B.L. (1995) Multiple-outcomes meta-analysis of treatments for periodontal disease. *J Dent Res* **74**, 1030-1039.

Berkey, C.S., Hoaglin, D.C., Antczak-Bouckoms, A., Mosteller, F. and Colditz, G.A. (1998) Meta-analysis of multiple outcomes by regression with random effects. *Stat.Med.* **17**, 2537-2550.

Berkey, C.S., Hoaglin, D.C., Mosteller, F. and Colditz, G.A. (1995) A random-effects regression model for meta-analysis. *Stat.Med.* **14**, 395-411.

Berlin, J.A., Laird, N.M., Sacks, H.S. and Chalmers, T.C. (1989) A comparison of statistical methods for combining event rates from clinical trials. *Stat.Med.* **8**, 141-151.

Biggerstaff, B.J. (1997) Confidence intervals in the one-way random effects model for meta-analytic applications. Technical Report. Department of Statristics, Colorado State University (1997).

Biggerstaff, B.J. and Tweedie, R.L. (1997) Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Stat.Med.* **16**, 753-768.

Birge, R.T. (1932) The calculation of errors by the method of least squares. *Phys.Rev.* **16**, 1-32.

Black, N. (1996) Why we need observational studies to evaluate the effectiveness of health care. *Br.Med.J.* **312**, 1215-1218.

Brand, R. and Kragt, H. (1992) Importance of trends in the interpretation of an overall odds ratio in the meta-analysis of clinical trials. *Stat.Med.* **11**, 2077-2082.

Briton, A., McKee, M., Black, N., McPherson, K., Sanderson, C. and Bain, C. (1998) Choosing between randomised and non-randomised studies: a systematic review. *Health Technology Assessment* **2**(13),

Brooks, S.P. and Gelman, A. (1998) Alternative methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* **7**, 434-455.

Bucher, H.C., Guyatt, G.H., Griffith, L.E. and Walter, S.D. (1997) The results of direct and indirect treatment comparisons in meta-analysis of randomised controlled trials. *J.Clin.Epidemiol.* **50**, 683-691.

Byar, D.P. (1980) Why data bases should not replace randomized clinical trials. *Biometrics* **36**, 337-342.

Byar, D.P. (1991) Problems with using observational databases to compare treatments. *Stat.Med.* **10**, 663-666.

Calvert, G.D. (1994) A review of observational studies on the relationship between cholesterol and coronary heart disease. *Aust NZ J Med* **24**, 89-91.

Carlin, J.B. (2000) Tutorial in biostatistics. Meta-analysis: formulating, evaluating, combining, and reporting (letter). *Stat.Med.* **19**, 753-761.

Carroll, R. J., Simpson, D. G., Zhou, H., and Guth, D. Stratified ordinal regression: A tool for combining information from disparate toxicological studies. #26. 94. Research Triangle Park, NC, National Institute of Statistical Sciences.

Chalmers, I., Sackett, D. and Silagy, C. (1997) The Cochrane Collaboration. In: Maynard, A. and Chalmers, I., (Eds.) *Non-random reflections on health services research*, pp. 231-249. BMJ Publishing Group: London

Chalmers, T.C., Smith H Jr, Blackburn, B., Silverman, B., Schroeder, B., Reitman, D. and Ambroz, A. (1981) A method for assessing the quality of a randomized control trial. *Controlled.Clin.Trials.* **2**, 31-49.

Chaloner, K. and Rhame, F.S. (2001) Quantifying and documenting prior beliefs in clinical trials. *Statistics in Medicine* **20**, 581-600.

Chelimsky, E., Silberman, G. and Droitcour, J. (1993) Cross design synthesis. *Lancet* **341**, 498

Chene, G. and Thompson, S.G. (1996) Methods for summarizing the risk associations of quantitative variables in epidemiologic studies in a consistent form. *Am.J.Epidemiol.* **144**, 610-621.

Claxton, K. (1999) Bayesian approaches to the value of information: implications for the regulation of new pharmaceuticals. *Health Economics* **8**, 269-274.

Cochran, W.G. (1937) Problems arising in the analysis of a series of similar experiments. *J.Roy.Statist.Soc.* **(Supplement), 4**, 102-118.

Cochran, W.G. (1954) The combination of estimates from different experiments. *Biometrics* **10**, 101-129.

Cochrane, A.L. (1979) 1931-1971: a critical review, with particular reference to the

medical profession. In: *Medicines for the year 2000*, Office of Health Economics: London

Colditz, G.A., Hankinson, S.E. and Hunter, D.J. (1995) The use of estrogens and progestins and the risk of breast cancer in postmenopausal women. *The New England Journal of Medicine* **332**, 1589-1593.

Collins, R., Yusuf, S. and Peto, R. (1985) Overview of randomised trials of diuretics in pregnancy. *BMJ* **290**, 17-23.

Concato, J., Shah, N. and Horwitz, R.I. (2000) Randomized, controlled trials, observational studies, and the hierarchy of research designs. *JAMA* **342**, 1887-1892.

Congdon, P. (2001) *Bayesian statistical modelling*. Chichester: John Wiley & Sons Ltd.

Cook, R.J. and Walter, S.D. (1997) A logistic model for trend in 2 x 2 x kappa tables with applications to meta-analyses. *Biometrics*. **53**, 352-357.

Cook, T.D. and Campbell, D.T. (1979) *Quasi-experimentation: Design & analysis issues for field settings.*, Houghton Mifflin: Boston.

Cooper, N.J., Abrams, K.R., Sutton, A.J., Lambert, P. and Turner, D. A Bayesian approach to Markov modeling in cost-effectiveness analyses: application to taxane use in advanced breast cancer. (Submitted 2001).

Copas, J. (1999) What works?: selectivity models and meta-analysis. *Journal of the Royal Statistical Society, Series A* **161**, 95-105.

Copas, J.B. and Li, H.G. (1997) Inference for non-randon samples. *Journal of the Royal Statistical Society, Series B* **59**, 55-95.

Copas, J.B. and Shi, J.Q. (2000) Reanalysis of epidemiological evidence on lung cancer and passive smoking. *Br.Med.J.* **320**, 417-418.

Cowles, M. K., Best, N. G., and Vines, K. CODA - Convergence diagnostics and output analysis software for Gibbs samples produced by the BUGS Language. Version 0.30. Technical Report. 94. MRC Biostatistics Unit, University of Cambridge, England.

Cox, L. H. and Piegorsch, W. W. Combining environmental information: Environmetric research in ecological monitoring, epidemiology, toxicology, and environmental data reporting. #12. 94. Research Triangle Park, NC, National Institute of Statistical Sciences.

Cummings, P. and Psaty, B.M. (1994) The association between cholesterol and death from injury. *Ann Intern Med* **120**, 848-855.

Davey Smith, G. and Pekkenen, J. (1992) Should there be a moratorium on the use of cholesterol lowering drugs? *Br.Med.J.* **304**, 431-434.

Deeks, J., Bradburn, M., Localio, R. and Berlin, J. (1999) Much ado about nothing: statistical methods for meta-analysis with rare events. Presented at Systematic Reviews: Beyond the basics, January Oxford (Abstract available at http://www.his.ox.ac.uk/csm/talks.html#p23.)

Deeks, J., Glanville, J., and Sheldon, T. Undertaking systematic reviews of research on effectiveness: CRD guidelines for those carrying out or commissioning reviews. #4. 96. Centre for Reviews and Dissemination, York, York Publishing Services Ltd.

Deeks, J.J. and Altman, D.G. (2001) Effect measures for meta-analysis of trials with binary outcomes. In: Egger, M., Davey Smith, G. and Altman, D.G., (Eds.) *Systematic reviews in health care. Meta-analysis in context*, 2nd edn. London: BMJ Publishing Group

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc B* **39**, 1-38.

DerSimonian, R. and Laird, N. (1986) Meta-analysis in clinical trials. *Controlled.Clin.Trials.* **7**, 177-188.

Detsky, A.S., Naylor, C.D., ORourke, K., McGeer, A.J., LAbbe, K.A., O'Rourke, K. and L'Abbe, K.A. (1992) Incorporating variations in the quality of individual randomized trials into meta-analysis. *J.Clin.Epidemiol.* **45**, 255-265.

Dickersin, K. and Berlin, J.A. (1992) Meta-analysis: state-of-the-science. *Epidemiol.Rev.* **14**, 154-176.

Dixon-Woods, M., Fitzpatrick, R. and Roberts, K. (2001) Including qualitative research in systematic reviews: opportunities and problems. *Journal of Evaluation in Clinical Practice* **7**, 125-133.

Dominici, F., Parmigiani, G., Wolpert, R.L. and Hasselblad, V. (1999) Meta-analysis of migrane headache treatments: combining information from heterogeneous designs. *American Statistical Association* **94**, 16-28.

Downs, S.H. and Black, N. (1998) The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *Journal of Epidemiology & Community Health* **52**, 377-384.

Droitcour, J., Silberman, G. and Chelimsky, E. (1993) Cross-design synthesis: A new form of meta-analysis for combining results from randomized clinical trials and medical- practice databases. *International Journal of Technology Assessment in Health Care* **9**, 440-449.

Drummond, M.F., O'Brien, B., Stoddart, G.L. and Torrance, G.W. (1997) *Methods for the economic evaluation of health care programmes*, 2nd edition. Oxford: Oxford University Press.

Duffy, S.W., Rohan, T.E. and Altman, D.G. (1989) A method for combining matched

and unmatched binary data. Application to randomized, controlled trials of photocoagulation in the treatment of diabetic retinopathy. *Am.J.Epidemiol.* **130**, 371-378.

DuMouchel, W. (1989) Bayesian Metaanalysis. In: Berry, D.A., (Ed.) *Statistical Methodology in the Pharmaceutical Sciences*, pp. 509-529. Marcel Dekker: New York

DuMouchel, W. Hierarchical Bayes linear models for meta-analysis. 27. (1994a.) Technical Report #27, National Institute of Statistical Sciences, PO Box 14162, Research Triangle Park, NC 27709

DuMouchel, W. (1994b) Predictive cross-validation in Bayesian meta-analysis. *Draft Form Proceedings of Fith Valencia International Meeting on Bayesian Statistics*, Dernardo, J. et al., eds. Valencia, Spain.

DuMouchel, W. (1998) Repeated measures meta-analysis. *Bulletin of the International Statistical Institute* **Tome LVII, Book 1**, 285-288.

DuMouchel, W. and Normand, S.L. (2000) Computer-modeling and graphical strategies for meta-analysis. In: Stangl, D.K. and Berry, D.A., (Eds.) *Meta-analysis in medicine and health policy*, New York: Marcel Dekker

DuMouchel, W.H. and Harris, J.E. (1983) Bayes methods for combining the results of cancer studies in humans and other species (with comment). *J.Am.Statist.Assoc.* **78**, 293-308.

Duval, S. and Tweedie, R. (2000a) R. A non-parametric "trim and fill" method of assessing publication bias in meta-analysis. *Biometrics,* **56**:455-463.

Duval S, Tweedie R. (2000b) Practical estimates of the effect of publication bias in meta-analysis. *Australasian Epidemiologist* 5: 14-17.

Duval, S. and Tweedie, R. (2000c) Trim and fill: A simple funnel plot based method of testing and adjusting for publication bias in meta-analysis. *The Journal of the American Statistical Association,***95**,89-98

EAFT (European Atrial Fibrillation Trial) Study Group (1993) Secondary prevention in non-rheumatic atrial fibrillation after transient ischaemc attack or minor stroke. *Lancet* **342**, 1255-1262.

Easterbrook, P.J., Berlin, J.A., Gopalan, R. and Matthews, D.R. (1991) Publication bias in clinical research. *Lancet* **337**, 867-872.

Eddy, D.M. (1989) The confidence profile method: A Bayesian method for assessing health technologies. *Operations Research* **37**, 210-228.

Eddy, D. M. and Hasselblad, V. FastPro: Software for MetaAnalysis by the Confidence Profile Method. 92. San Diego, California, Academic Press Inc.

Eddy, D.M., Hasselblad, V., McGivney, W. and Hendee, W. (1988) The value of

mammography screening in women under the age of 50 years. *JAMA* **259**, 1512-1519.

Eddy, D.M., Hasselblad, V. and Shachter, R. (1990a) A Bayesian method for synthesizing evidence: The confidence profile method. *International Journal of Technology Assessment in Health Care* **6**, 31-55.

Eddy, D.M., Hasselblad, V. and Shachter, R. (1990b) An introduction to a Bayesian method for meta-analysis: The confidence profile method. *Medical Decision Making* **10**, 15-23.

Eddy, D.M., Hasselblad, V. and Shachter, R. (1992) *Meta-analysis by the Confidence Profile Method*, Academic Press: San Diego.

Efron, B. and Tibshirani, R.J. (1993) *An introduction to the Bootstrap*, 1st edn. Chapman & Hall: New York.

Egger, M., Davey Smith, G. and Altman.D.G. (2000) *Systematic revews in health care: Meta-analysis in context*, London: BMJ Books.

Egger, M., Smith, G.D., Schneider, M. and Minder, C. (1997) Bias in meta-analysis detected by a simple, graphical test. *Br.Med.J.* **315**, 629-634.

Egger, M., Sterne, J.A.C. and Davey Smith, G. (1998) Meta-analysis software. *Br.Med.J.* **316**:(website only: http://bmj.com/archive/7126/7126ed9.htm)

Emerson, J.D. (1994) Combining estimates of the odds ratio: the state of the art. *Stat.Methods Med.Res.* **3**, 157-178.

Enderby, P. and Emerson, J. (1995) *Does speech and language therapy work?*, Whurr Publishers Ltd: London.

Engles.E.A., Schmid, C.H., Terrin, N., Olkin, I. and Lau, J. (2001) Heterogeneity ans statistical significance in meta-analysis: an empirical study of 125 meta-analyses. *Statistics in Medicine* **19**, 1707-1728.

Farmer, R. and Miller, D. (1996) *Lecture notes on Epidemiology and public health medicine*, 4th edn. Blackwell Scientific Publications.

Fleiss, J.L. (1981) *Statistical methods for rates and proportions*, 2nd edn. Wiley: New York.

Fleiss, J.L. (1993) The statistical basis of meta-analysis. *Stat.Methods Med.Res.* **2**, 121-145.

Fleiss, J.L. (1994) Measures of effect size for categorical data. In: Cooper, H. and Hedges, L.V., (Eds.) *The handbook of research synthesis*, pp. 245-260. Russell Sage Foundation: New York

Fortin, P.R., Lew, R.A., Liang, M.H., Wright, E.A., Beckett, L.A., Chalmers, T.C. and Sperling, R.I. (1995) Validation of a meta-analysis: the effects of fish oil in rheumatoid arthritis. *J Clin Epidemiol* **48**, 1379-1390.

Freemantle, N. and Mason, J. (1997) Publication bias in clinical trials and economic analyses. *Pharmacoeconomics* **12**, 10-16.

Galbraith, R.F. (1988) A note on graphical presentation of estimated odds ratios from several clinical trials. *Stat.Med.* **7**, 889-894.

Gelber, R.D. and Goldhirsch, A. (1987) Interpretation of results from subset analyses within overviews of randomized clinical trials. *Stat.Med.* **6**, 371-378.

General Accounting Office (1992). Cross design synthesis: a new strategy for medical effectiveness research. Washington, DC, GAO.

General Accounting Office (1994). Breast conservation versus mastectomy. Patient survival in day-to-day medical practice and in randomized studies. Report GAO/PEMID-95-9. Washington, DC, GAO.

Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1996) *Markov Chain Monte Carlo in practice*, Chapman and Hall: London.

Givens, G.H., Smith, D.D. and Tweedie, R.L. (1997) Publication bias in meta-analysis: a Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate. With discussion by Begg,C.B. and others. *Statistical Science* **12**, 221-250.

Glass, G.V. (1976) Primary, secondary and meta-analysis of research. *Educ.Res.* **5**, 3-8.

Glasziou, P. and Irwig, L. (1995) Generalizing randomized trial results using additional epidemiologic information. *Am.J.Epidemiol.* **141**, S 47

Glasziou, P.P., Bromwich, S. and Simes, R.J. (1994) Quality of life six months after myocardial infarction treated with thrombolytic therapy. *The Medical Journal of Australia* **161**, 532-536.

Glasziou, P.P. and Irwig, L.M. (1995) An evidence based approach to individualizing treatment. *Br.Med.J.* **311**, 1356-1359.

Glasziou, P.P., Simes, R.J. and Gelber, R.D. (1990) Quality adjusted survival analysis. *Statistics in Medicine* **9**, 1259-1276.

Gleser, L.J. and Olkin, I. (1994) Stochastically dependent effect sizes. In: Cooper, H. and Hedges, L.V., (Eds.) *The handbook of research synthesis*, pp. 339-356. Russell Sage Foundation: New York

Goldstein, H., Yang, M., Omar, R.Z., Turner, R.M. and Thompson, S.G. (2000) Meta-analysis using multilevel models with an application to the study of class size effects. *Applied Statistics* **49**, 399-412.

Green, S.J., Fleming, T.R. and Emerson, S. (1987) Effects on overviews of early stopping rules for clinical-trials. *Stat.Med.* **6**, 361

Greenland, S. (1987) Quantitative methods in the review of epidemiological literature.

Epidemiol.Rev. **9**, 1-30.

Greenland, S. (1994) Invited commentary: a critical look at some popular meta-analytic methods. *Am J Epidemiol* **140**, 290-296.

Greenland, S. and Salvan, A. (1990) Bias in the one-step method for pooling study results. *Stat.Med.* **9**, 247-252.

Grodstein, F., Stampfer, M.J. and Colditz, G.A. (1997) Postmenopausal hormone therapy and mortality. *The New England Journal of Medicine* **336**, 1769-1775.

Hardy, R.J. and Thompson, S.G. (1996) A likelihood approach to meta-analysis with random effects. *Stat.Med.* **15**, 619-629.

Hasselblad, V. and Hedges, L.V. (1995) Meta-analysis of screening and diagnostic tests. *Psychol.Bull.* **117**, 167-178.

Hasselblad, V.I.C. and Mccrory, D.C. (1995) Meta-analytic tools for medical decision making: A practical guide. *Med Decis Making* **15**, 81-96.

Hauck, W.W. (1984) A comparative study of conditional maximum likelihood estimation of a common odds ratio. *Biometrics* **40**, 1117-1123.

Hedges, L.V. (1992) Modeling publication selection effects in meta-analysis. *Statistical Science* **7**, 246-255.

Hedges, L.V. and Olkin, I. (1985) *Statistical Methods for Meta-Analysis*, Academic Press: London.

Hedges, L.V. and Vevea, J.L. (1996) Estimating effects size under publication bias: small sample properties and robustness of a random effects selection model. *J Educ Behav Stat* **21**, 299-333.

Hemminki, E. and McPherson, K. (1997) Impact of postmenopausal hormone therapy on cardiovascular events and cancer: pooled data from clinical trials. *Br.Med.J.* **315**, 149-153.

Higgins, J. P. T. Exploiting information in random effects meta-analysis. 97. Department of Applied Statistics, The University of Reading.

Higgins, J.P.T. and Whitehead, A. (1996) Borrowing strength from external trials in a meta-analysis. *Stat.Med.* **15**, 2733-2749.

Higgins, J.P.T., Whitehead, A., Turner, R.M., Omar, R.Z. and Thompson, S.G. (2001) Meta-analysis of continuous outcome data from individual patients. *Statistics in Medicine* **20**, 2219-2241.

Hlatky, M.A. (1991) Using databases to evaluate therapy. *Stat.Med.* **10**, 647-652.

Hornbuckle, J., Vail, A., Abrams, K.R. and Thornton, J.G. (2000) Bayesian interpretation of trials: the example of intrapartum electronic fetal heart rate monitoring. *Br J Obstet Gynaecol* **107**, 3-10.

Hughes, M.D., Freedman, L.S. and Pocock, S.J. (1992) The impact of stopping rules on heterogeneity of results in overviews of clinical trials. *Biometrics* **48**, 41-53.

Huicho, L., Campos, M., Rivera, J. and Guerrant, R.L. (1996) Fecal screening tests in the approach to acute infectious diarrhea: A scientific overview. *Pediatric Infectious Disease Journal* **15**, 486-494.

Huque, M.F. and Dubey, S.D. (1994) A metaanalysis methodology for utilizing study-level covariate- information from clinical-trials. *Communications In Statistics-Theory And Methods* **23**, 377-394.

Hutton, J.L. and Williamson, P.R. (2000) Bias in meta-analysis due to outcome varible selection within studies. *Applied Statristics* **49**, 359-370.

Ioannidis, J.P.A., Haidich, A.-B., Pappa, M., Pantazis, N., Kokori, S.I., Tektonidou, M.G., Contopoulos-Ioannidis, D.G. and Lau, J. (2001) Comparison of evidence of treatment effects in randomised and nonrandomised studies. *JAMA* **286**, 821-830.

Iyengar, S. and Greenhouse, J.B. (1988) Selection models and the file drawer problem. *Statistical Science* **3**, 109-135.

Jefferson, T., Mugford, M., Gray, A. and DeMicheli, V. (1996) An exercise in the feasibility of carrying out secondary economic analysis. *Health Economics* **5**, 155-165.

Juni, P., Witschi, A., Bloch, R. and Egger, M. (1999) The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA* **282**, 1054-1060.

Katerndahl, D.A. and Lawler, W.R. (1999) Variability in meta-analytic results concerning the value of cholesterol reduction in coronary heart disease: a meta-meta-analysis. *Am.J.Epidemiol.* **149**, 429-441.

Kunz, R. and Oxman, A.D. (1998) The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *BMJ* **317**, 1185-1190.

L'Abbe, K.A., Detsky, A.S. and O'Rourke, K. (1987) Meta-analysis in clinical research. *Annals of Internal Medicine* **107**, 224-233.

Lambert, P., Sutton, A.J., Abrams, K.R. and Jones, D.R. (2001a) A comparison of summary patient level covariates in meta-regression with individual patient data meta-analyses. *Journal of Clinical Epidemiology*

Lambert, P., Sutton, A.J., Jones, D.R. and Abrams, K.R. (2001b) How vague is vague? (*In preparation*)

Lambert, P.C. and Abrams, K.R. (1996) Meta-analysis using multilevel models. *Multilevel Modelling Newsletter* **7**, 17-19.

Lancet (1992) Cross design synthesis: a new strategy for studying medical outcomes? *Lancet* **340**, 944-946.

Larose, D.T. and Dey, D.K. (1997a) Grouped random effects models for Bayesian meta-analysis. *Stat.Med.* **16**, 1817-1829.

Larose, D. T. and Dey, D. K. (1997b) Modeling dependent covariate subclass effects in Bayesian meta-analysis. Technical Report #96-22, University of Connecticut.

Lau, J., Ioannidis, J.P. and Schmid, C.H. (1998) Summing up evidence: one answer is not always enough. *The Lancet* **351**, 123-127.

Law, M.R., Wald, N.J. and Thompson, S.G. (1994a) By how much and how quickly does reduction in serum cholesterol concentration lower risk of ischaemic heart disease? *Br.Med.J.* **308**, 367-373.

Law, M.R., Wald, N.J., Wu, T., Hackshaw, A. and Bailey, A. (1994b) Systematic underestimation of association between serum cholesterol concentration and ischaemic heart disease in observational studies: data from the BUPA study. *Br.Med.J.* **308**, 363-366.

Lee, P.M. (1989) *Bayesian statistics: an introduction*, Edward Arnold: London.

Li, W., Girard, P., Boissel, J.P. and Gueyffier, F. (1998) The calculation of a confidence interval on the absolute estimated benefit for an individual patient. *Computer and Biomedical Research* **31**, 244-256.

Li, Y.Z., Shi, L. and Roth, H.D. (1994) The bias of the commonly-used estimate of variance in metaanalysis. *Communications In Statistics-Theory And Methods* **23**, 1063-1085.

LI, Z.H. and Begg, C.B. (1994) Random effects models for combining results from controlled and uncontrolled studies in a metaanalysis. *J.Am.Statist.Assoc.* **89**, 1523-1527.

Liao, J.G. (1999) A hierarchical Bayesian model for combining multiple 2 x 2 tables using conditional likelihoods. *Biometrics* **55**, 268-272.

Light, R.J. (1987) Accumulating evidence from independent studies - what we can win and what we can lose. *Stat.Med.* **6**, 221-231.

Light, R.J. and Pillemar, D.B. (1984) *Summing Up: The science of Reviewing Research.*, Harvard University Press: Cambridge, Mass.

Macaskill, P., Walter, S.D. and Irwig, L. (2001) A comparison of methods to detect publication bias in meta-analysis. *Statistics in Medicine* **20**, 641-654.

Macmahon, S. (1994) Cholesterol reduction and death from noncoronary causes - evidence from randomized controlled trials. *Australian And New Zealand Journal Of Medicine* **24**, 120-123.

Mantel, N. and Haenszel, W. (1959) Statistical aspects of the analysis of data from retrospective studies of disease. *J.Natl.Cancer.Inst.* **22**, 719-748.

Margitic, S.E., Morgan, T.M., Sager, M.A. and Furberg, C.D. (1995) Lessons learned

from a prospective meta-analysis. *J.Am.Geriatr.Soc.* **43**, 435-439.

Marshall, E.C. and Spiegelhalter, D.J. (1998) Reliability of league tables of in vitro fertilisation clinics:retrospective analysis of live birth rates. *BMJ* **316**, 1701-1705.

McDonagh, M.S., Whiting, P.F., Wilson, P.M., Sutton, A.J., Chestnutt, I., Cooper, J., Misso, K., Bradley, M., Treasure, E. and Kleijnen, J. (2000) Systematic review of water fluoridation. *BMJ* **321**, 855-859.

McIntosh, M.W. (1996) The population risk as an explanatory variable in research synthesis of clinical trials. *Stat.Med.* **15**, 1713-1728.

Mehta, C.R., Patel, N.R. and Grey, R. (1997) Computing an exact confidence interval for the common odds ratio in several 2 x 2 contingency tables. *J.Am.Statist.Assoc.* **80**, 969-973.

Mengersen, K.L., Tweedie, R.L. and Biggerstaff, B.J. (1995) The impact of method choice in meta-analysis. *Aust.J.Stats* **37**, 19-44.

Midgette, A.S., Wong, J.B., Beshansky, J.R., Porath, A., Fleming, C. and Pauker, S.G. (1994) Cost-effectiveness of streptokinase for acute myocardial- infarction - a combined metaanalysis and decision-analysis of the effects of infarct location and of likelihood of infarction. *Medical Decision Making* **14**, 108-117.

Moher, D., Jadad, A.R., Nichol, G., Penman, M., Tugwell, P. and Walsh, S. (1995) Assessing the quality of randomized controlled trials - an annotated-bibliography of scales and checklists. *Controlled.Clin.Trials.* **12**, 62-73.

Moher, D., Jadad, A.R. and Tugwell, P. (1996) Assessing the quality of randomised controlled trials: current issues and future directions. *International Journal of Technology Assessment in Health Care* **12**, 195-208.

Moher, D., Klassen, T.P., Jadad, A.R., Tugwell, P., Moher, M. and Jones, A.L. (1998a) Assessing the quality of randomised controlled trials: implications for the conduct of meta-analyses. *Health Technology Assessment* **2**(12),

Moher, D., Pham, B., Jones, A., Cook, D.J., Jadad, A.R., Moher, M., Tugwell, P. and Klassen, T.P. (1998b) Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analysis? *Lancet* **352**, 609-613.

Moreno, V., Martin, M.L., Bosch, F.X., De Sanjose, S., Torres, F., Munoz, N. and Desanjose, S. (1996) Combined analysis of matched and unmatched case-control studies: Comparison of risk estimates from different studies. *Am.J.Epidemiol.* **143**, 293-300.

MRC Health Services and Public Health Research Board (2001) A framework for development and evaluation of RCTs for complex interventions to improve health (discussion document). (http://www.mrc.ac.uk)

Muldoon, M.F., Manuck, S.B. and Matthews, K.A. (1990) Lowering cholesterol

concentrations and mortality: a quantitative review of primary prevention trials. *Br.Med.J.* **301**, 309-314.

Muller, P., Parmigiani, G., Schildkraut, J. and Tardella, L. (1999) A Bayesian hierarchical approach for combining case-control and prospective studies. *Biometrics* **55**, 858-866.

Naylor, C.D. (1988) Two cheers for meta-analysis: problems and opportunities in aggregating results of clinical trials. *Can.Med.Assoc.J.* **138**, 891-895.

Naylor, C.D. (1989) Meta-analysis of controlled clinical trials. *Journal of Rheumatology* **16**, 424-426.

Neal, R. Suppressing random walks in Markov chain Monte Carlo using ordered over-relaxation. In: Jordan, M.I., (Ed.) *Learnining in Graphical Models*, pp. 205-230. Dorderecht: Kluwer Academic Publishers.

O'Hagan, A. (1988) *Probability: methods and measurement*, Chapman and Hall: London.

O'Hagan, A. (1994) *Bayesian Inference*, Edward Arnold: London.

O'Hagan, A. (1998) Eliciting expert beliefs in substantial practical applications. *Statistician* **47**, 21-35.

O'Hagan, A. and Stevens, J.W. (2001) A framework for cost-effectiveness analysis from clinical trial data. *Health Economics* **10**, 302-315.

Oglesby, P. and Hennekens, C. (1992) Long-term mortality after primary prevention for cardiovascular disease. *JAMA* **267**, 2185-2186.

Oxman, A.D. (2001) The Cochrane Collaboration in the 21st century: ten challenges and one reason why they must be met. In: Egger, M., Davey Smith, G. and Altman, D.G., (Eds.) *Systematic reviews in health care: meta-analysis in context*, 2nd edn. London: BMJ Publishing Group

Oxman, A.D.Ed. (1996) *The Cochrane Collaboration handbook: preparing and maintaining systematic reviews*, Second edn. Cochrane Collaboration: Oxford.

Perkins, L.L., Clark, B.D., Klein, P.J. and Cook, R.R. (1995) A meta-analysis of breast implanats and connective tissue disease. *Annals of Plastic Surgery* **35**, 561-570.

Peto, R. (1987) Why do we need systematic overviews of randomised trials? *Stat.Med.* **6**, 233-240.

Peto, R., Pike, M.C., Armitage, P., Breslow, N.E., Cox, D.R., Howard, S.V., Mantel, N., McPherson, K., Peto, J. and Smith, P.G. (1977) Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II: Analysis and examples. *Br.J.Cancer.* **35**, 1-39.

Petticrew, M., Gilbody, S. and Sheldon, T.A. (1999) Relation between hostility and coronary heart disease. Evidence does not support link. *Br.Med.J.* **319**, 917

Phillips, A. and Holland, P.W. (1987) Estimators of the variance of the Mantel-Haenszel log-odds-ratio estimate. *Biometrics* **43**, 425-431.

Prevost, T.C., Abrams, K.R. and Jones, D.R. (2000) Hierarchical models in generalised synthesis of evidence: an example based on studies of breast cancer screening. *Statistics in Medicine* **19**, 3359-3376.

Psaty, B., Koepsell, T., Lin, D., Weiss, N., Siscovick, D., Rosendall, F., Pahor, M. and Furberg, C.D. (1999) Assessment and control for confounding by indication in observational studies. *The Journal of the American Geriatrics Society* **47**, 749-754.

Raudenbush, S.W. (1994) Random effects models. In: Cooper, H. and Hedges, L.V., (Eds.) *The handbook of research synthesis*, pp. 301-322. Russell Sage Foundation: New York

Raudenbush, S.W., Becker, B.J. and Kalaian, H. (1988) Modeling multivariate effect sizes. *Psychol.Bull.* **103**, 111-120.

Reeves, B.C., MacLehose, R.R., Harvey, I.M., Sheldon, T.A., Russel, I.T. and Black, A.M.S. (1998) Black, N., Brazier, J., Fitzpatrick, R. and Reeves, B., (Eds.) *Comparisons of effect size estimates derived from randomised and non-randomised studies*, 1st edn. London: BMJ Publishing Group.

Robins, J., Breslow, N. and Greenland, S. (1986a) Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics* **42**, 311-323.

Robins, J., Greenland, S. and Breslow, N.E. (1986b) A general estimator for the variance of the Mantel-Haenszel odds ratio. *Am.J.Epidemiol.* **124**, 719-723.

Robinson, D. (1996) Head injuries and bicycle helmet laws. *Accident Analysis and Prevention* **28**, 463-475.

Rogatko, A. (1992) Bayesian-approach for metaanalysis of controlled clinical-trials. *Communications In Statistics-Theory And Methods* **21**, 1441-1462.

Rosenthal, R. (1978) Combining the results to independent studies. *Profess.Psychol.* **17**, 136-137.

Rosenthal, R. (1979) The file drawer problem and tolerance for null results. *Psychol.Bull.* **86**, 638-641.

Rubin, D. (1992) A new perspective. In: Wachter, K.W. and Straf, M.L., (Eds.) *The future of meta-analysis*, pp. 155-165. Russell Sage Foundation: New York

Rubin, D.B. (1981) The Bayesian Bootstrap. *The Annals of Statistics* **9**, 130-134.

Rubins, H.B. (1995) Cholesterol in patients with coronary heart disease: How low should we go? *J Gen Intern Med* **10**, 464-471.

Sackett, D.L., Rosenberg, W.M., Gray, J.A., Haynes, R.B. and Richardson, W.S. (1996)

Evidence-based medicine: what it is and what it isn't. *Br.Med.J.* **312**, 71-72.

Sacks, H.S., Berrier, J., Reitman, D., Ancona-Berk, V.A. and Chalmers, T.C. (1987) Meta-analysis of randomized controlled trials. *New.Engl.J.Med.* **316**, 450-455.

Sankey, S.S., Weissfeld, L.A., Fine, M.J. and Kapoor, W. (1996) An assessment of the use of the continuity correction for sparse data in metaanalysis. *Communications In Statistics-Simulation And Computation* **25**, 1031-1056.

SAS Institute Inc. (1992) SAS Technical Report P-229. SAS/STAT Software: Changes and Enhancements. Release 6.07.

Sato, T. (1990) Confidence limits for the common odds ratio based on the asymptotic distribution of the Mantel-Haenszel estimator. *Biometrics* **46**, 71-80.

Schmid, C.H., Lau, J., McIntosh, M.W. and Cappelleri, J.C. (1998) An empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials. *Stat.Med.* **17**, 1923-1942.

Schneider, B. (1989) Analysis of Clinical Trial Outcomes: Alternative Approaches to Subgroup Analysis. *Controlled.Clin.Trials.* **10**, 176S-186S.

Schulz, K.F., Chalmers, I., Hayes, R.J. and Altman, D.G. (1995) Empirical evidence of bias: Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *J.A.M.A.* **273**, 408-412.

Seltzer, M. (1991) The use of data augmentation in fitting hierarchical models to education data. Unpublished doctorial dissertation, University of Chicargo.

Senn, S. (1994) Importance of Trends in the Interpretation of an Overall Odds Ratio in the Meta-Analysis of Clinical Trials. *Stat Med* **13**, 293-296.

Senn, S., Sharp, S., Thompson, S. and Altman, D. (1996) Relation between treatment benefit and underlying risk in meta- analysis. *Br.Med.J.* **313**, 1550-1551.

Shadish, W.R. and Haddock, C.K. (1994) Combining estimates of effect size. In: Cooper, H. and Hedges, L.V., (Eds.) *The handbook of research synthesis*, pp. 261-284. Russell Sage Foundation: New York.

Sharp, S. (1998) Meta-analysis regression. *Stata Technical Bulletin* **42**:16-22.

Sharp, S.J., Thompson, S.G. and Altman, D.G. (1996) The relation between treatment benefit and underlying risk in metaanalysis. *Br.Med.J.* **313**, 735-738.

Shea, S., DuMouchel, W. and Bahamonde, L. (1996) A meta-analysis of 16 randomised controlled trials to evaluate computer-based clinical reminder systems for preventive care in the ambulatory setting. *JASA* **3**, 399-409.

Simes, R.J. (1986) Publication bias: the case for an international registry of clinical trials. *J.Clin.Oncol.* **4**, 1529-1541.

Simpson, E.H. (1951) The interpretation of interaction in contingency tables. *Journal*

Of The Royal Statistical Society Series B **13**, 238-241.

Sinclair, J.C. and Bracken, M.B. (1994) Clinically useful measures of effect in binary analyses of randomized trials. *J.Clin.Epidemiol.* **47**, 881-889.

Skene, A.M. and Wakefield, J.C. (1990) Hierarchical models for multicentre binary response studies. *Stat.Med.* **9**, 919-929.

Smith, G.D., Song, F., Sheldon, T.A. and Song, F.J. (1993) Cholesterol lowering and mortality: The importance of considering initial level of risk. *Br.Med.J.* **306**, 1367-1373.

Smith, R. (1999) What is publication? A continuum. *BMJ* **318**, 142

Smith, S.J., Caudill, S.P., Steinberg, K.K. and Thacker, S.B. (1995) On combining dose-response data from epidemiological studies by meta-analysis. *Stat Med* **14**, 531-544.

Smith, T.C., Spiegelhalter, D. and Parmar, M.K.B. (1995a) Bayesian meta-analysis of randomized triles using graphical models and BUGS. In: *Bayesian Biostatistics*, pp. 411-427.

Smith, T.C., Spiegelhalter, D.J. and Thomas, A. (1995b) Bayesian approaches to random-effects meta-analysis: A comparative study. *Stat Med* **14**, 2685-2699.

Song, F., Glenny, A. and Altman, D.G. (2000a) Indirect comparison in evaluation relative efficacy illustrated by antimicrobial prophylaxis in colorectal surgery. *Controlled Clinical Trials* **21**, 488-497.

Song, F., Easterwood, A., Gilbody, S., Duley, L. and Sutton, A.J. (2000b) Publication and other selection biases in systematic reviews. *Health Technology Assessment* **4**(10).

Song, F., Kahn, K.S., Dinnes, J. and Sutton, A. (2002) Asymmetric funnel plots and the probelm of publication bias in meta-analyses of diagnostic accuracy. *International Journal of Epidemiology* (*to appear*)

Spector, T.D. and Thompson, S.G. (1991) Research methods in epidemiology .5. the potential and limitations of metaanalysis. *J.Epidemiol.Comm.Hlth.* **45**, 89-92.

Spiegelhalter, D., Thomas, A. and Gilks, W. (1994) *BUGS Examples 0.30.1*, MRC Biostatistics Unit: Cambridge.

Spiegelhalter, D.J., Thomas, A. and Best, N.G. (2000a) *WinBUGS Version 1.2. User manual*, MRC Biostatistics Unit: Cambridge.

Spiegelhalter, D.J., Myles, J.P., Jones, D.R. and Abrams, K.R. (2000b) Bayesian methods in Health Technology Assessment. *Health Technology Assessment* **4**(38).

Spitzer, W.O. (1991) Meta-meta-analysis: unanswered questions about aggregating data. *J.Clin.Epidemiol.* **44**, 103-107.

Steichen, T.J. (1998) Tests for publication bias in meta-analysis. *Stata Technical Bulletin* **41:sbe20**, 9-15.

Steichen, T.J., Egger, M. and Sterne, J. (1998) Modification of the metabias program. *Stata Technical Bulletin* **STB 44: sbe19.1**, 3-4.

Steinberg, K.K., Smith, S.J., Thacker, S.B. and Stroup, D.F. (1994) Breast cancer risk and duration of estrogen use: The role of study design in meta-analysis. *Epidemiology* **5**, 415-421.

Sterne, J.A.C., Egger, M. and Sutton, A.J. (2000a) Meta-analysis software. In: Egger, M., Davey Smith, G. and Altman, D.G., (Eds.) *Systematic Reviews in Health Care: Meta-analysis in context.*, 2nd edn. BMJ Books: London

Sterne, J.A.C., Gavaghan, D. and Egger, M. (2000b) Publication and related bias in meta-analysis: Power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology* **53**, 1119-1129.

Stevenson, J. (1998) Meta-analysis of trials and studies of adjuvant chemotherapy in the treatment of childhood medulloblastoma. MSc Thesis, Leicester University.

Stinnett, A.A. and Mulahy, J. (1998) Net health benefits: A new framework for the analysis of uncertainty in cost-effectiveness analysis. *Medical Decision Making* **18**, S68-S80

Stram, D.O. (1996) Meta-analysis of published data using a linear mixed-effects model. *Biometrics* **52**, 536-544.

Stroup, D.F., Berlin, J.A., Morton, S.C., Olkin, I., Williamson, G.D., Rennie, D., Moher, D., Becker, B.J., Sipe, T.A., Thacker, S.B. and for the Meta-analysis Of Observational Studies in Epidemiology (MOOSE) Group (2000) Meta-analysis of observational studies in epidemiology: a proposal for reporting. *JAMA* **283**, 2008-2012.

Strube, M.J. (1985) Combining and comparing significance levels from non-independent hypothesis tests. *Psychol.Bull.* **97**, 334-341.

Sutton, A.J. and Abrams, K.R. (1998) Questions and Answers: If RCTs are so good for evidence based medicine, why are purchasers beginning to say that trials are artificial and contrived and that the want 'real world' evidence of effectiveness? *Health Services Research and Policy* **3**:255-256.

Sutton, A.J. and Abrams, K.R. (2001) Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research* **10**, 277-303.

Sutton, A. J., Abrams, K. R., Jones, D. R., Sheldon, T. A., and Song, F. (1998) Systematic reviews of trials and other studies. Health Technology Assessment **2**(19).

Sutton, A.J., Lambert, P.C., Hellmich, M., Abrams, K.R. and Jones, D.R. (1998) Meta-analysis in practice: a critical review of available software. In: Berry, D.A. and Stangl, D.K., (Eds.) *Meta-analysis in medicine and health policy*, Marcel

Dekker:

Sutton, A.J., Abrams, K.R., Jones, D.R., Sheldon, T.A. and Song, F. (2000a) *Methods for meta-analysis in medical research*, John Wiley: London.

Sutton, A.J., Duval, S.J., Tweedie, R.L., Abrams, K.R. and Jones, D.R. (2000b) Empirical assessment of effect of publication bias on meta-analyses. *BMJ* **320**, 1574-1577.

Sutton, A.J., Song, F., Gilbody, S.M., Abrams, K.R. (2000c) Modelling publication bias in meta-analysis: a review. *Statistical Methods in Medical Research* **9**:421-445.

Sweeting, M. (2001) Meta-analysis of rare events. MSc dissertation, University of Leicester.

Tang, J.L. (2000) Weighting bias in meta-analysis of binary outcomes. *Journal of Clinical Epidemiology* **53**, 1130-1136.

Tang, J.L., Armitage, J.M., Lancaster, T., Silagy, C.A., Fowler, G.H. and Neil, H.A.W. (1998) Systematic review of dietary intervention trials to lower blood total cholesterol in free-living subjects. *Br.Med.J.* **316**, 1213-1220.

Tang, J.L. and Liu, J.L. (2000) Misleading funnel plot for detection of bias in meta-analysis. *Journal of Clinical Epidemiology* **53**, 477-484.

Thacker, S.B. (1988) Meta-analysis. A quantitative approach to research integration. *J.A.M.A.* **259**, 1685-1689.

Thacker, S.B., Stroup, D.F. and Peterson, H.B. (1997) Continuous electronic fetal heart rate monitoring during labour. In: Neilson JP, Crowther CA, Hodnett ED, Hofmeyer GJ, Keirse MJNC, editors. Cochrane Database of Systematic Reviews Pregnancy and Childbirth Module. Cochrane Updates on Disk: Issue 2; Oxford: Update Sorfware.

Thompson, D.C., Rivara, F.P. and Thompson, R.S. Helmets for preventing head and facial injuries in bicycling. *Cochrane Libary. Issue 4. Oxford: Update Software, 1999*

Thompson, R., Parkin, D., Eccles, M., Sudlow, M. and Robinson, A. (2000) Decision analysis and guidelines for anticoagulant therapy to prevent stroke in patients with atrial fibrillation. *The Lancet* **355**, 956-962.

Thompson, S.G. (1993) Controversies in meta-analysis: the case of the trials of serum cholesterol reduction. *Stat.Methods Med.Res.* **2**, 173-192.

Thompson, S.G. and Barber, J.A. (2000) How should cost data in pragmatic randomised trials be analysed? *BMJ* **320**, 1197-1200.

Thompson, S.G. and Sharp, S.J. (1999) Explaining heterogeneity in meta-analysis: a comparison of methods. *Stat.Med.* **18**, 2693-2708.

Thompson, S.G., Smith, T.C. and Sharp, S.J. (1997) Investigation underlying risk as a

source of heterogeneity in meta-analysis. *Stat Med* **16**, 2741-2758.

Tobias, A. (1999) Assessing the influence of a single study in the meta-analysis estimate. *Stata Technical Bulletin* **47:sbe26**, 15-17.

Tori, V., Simon, R., Russek-Cohen, E., Midthune, D. and Friedman, M. (1992) Statistical model to determine the relationship of response and survival in patients with advanced ovarian cancer treated with chemotherapy. *J.Natl.Cancer.Inst.* **84**, 407-414.

Tritchler, D. (1999) Modelling study quality in meta-analysis. *Stat.Med.* **18**, 2135-2145.

Turner, R.M., Omar, R.Z., Yang, M., Goldstein, H. and Thompson, S.G. (2000) A multilevel model framework for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine* **19**, 3417-3432.

Tweedie, R.L. and Mengersen, K.L. (1992) Lung cancer and passive smoking: Reconciling the biochemical and epidemiological approaches. *Br.J.Cancer.* **66**, 700-705.

Van Houwelingen, H.C., Zwinderman, K.H. and Stijnen, T. (1993) A bivariate approach to meta-analysis. *Stat.Med.* **12**, 2273-2284.

Vanhonacker, W.R. (1996) Meta-analysis and response surface extrapolation: a least squares approach. *American Statistician* **50**, 294-299.

Verdinelli, I., Andrews, K., Detre, K., and Peduzzi, P. (1995) The Bayesian approach to meta-analysis: a case study. Departmet of Statistics, Carnegie Mellon University.

Various authors (1997) Impact of postmenopausal hormone therapy on cardiovascular events and cancer - letters. *BMJ* **315**, 676-678.

Waclawiw, M.A. and Liang, K.Y. (1994) Empirical bayes estimation and inference for the random effects model with binary response. *Stat.Med.* **13**, 541-551.

Wald, N.J., Nanchahal, K., Thompson, S.G. and Cuckle, H.S. (1986) Does breathing other people's tobacco smoke cause lung cancer? *Br.Med.J.* **293**, 1217-1222.

Walter, S.D. (2000) Choice of effect measure for epidemiological data. *Journal of Clinical Epidemiology* **53**, 931-939.

Waternaux, C. and DuMouchel, W. (1993) Combining information across sites with hierarchical Bayesian linear models. San Francisco. Proceedings of the Section on Bayesian Statistics.

Whitehead, A., Bailey, A.J. and Elbourne, D. (1999) Combining summaries of binary outcomes with those of continuous outcomes in a meta-analysis. *Journal of Biopharmaceutical Statistics*. **9**(1):1-16.

Whitehead, A., Omar, R.Z., Higgins, J.P.T., Savaluny, E., Turner, R.M. and Thompson,

S.G. (2001) Meta-analysis of ordinal outcomes using individual patient data. *Statistics in Medicine* **20**, 2243-2260.

Whitehead, A. and Whitehead, J. (1991) A general parametric approach to the meta-analysis of randomised clinical trials. *Stat.Med.* **10**, 1665-1677.

Wortman, P.M. (1983) Evaluation research: A methodological perspective. *Annual Review of Psychology* **34**, 223-260.

Wortman, P.M. (1994) Judging research quality. In: Cooper, H. and Hedges, L.V., (Eds.) *The handbook of research synthesis*, pp. 97-110. Russell Sage Foundation: New York

Yusuf, S., Peto, R., Lewis, J., Collins, R., Sleight, P. and et al. (1985) Beta blockade during and after myocardial infarction: an overview of the randomised trials. *Progress in Cardiovascular Diseases* **27**, 335-371.

Yusuf, S., Wittes, J., Probstfield, J. and Tyroler, H.A. (1991) Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA* **266**, 93-98.

Appendix A

Supplement to Chapters 4 and 5 on the cholesterol synthesis

A.I. Previous meta-analyses of the cholesterol lowering RCTs (in chronological order)

M1. Mann JI, Marr JW. Miller WE, Lewis B, editors.Lippoproteins, atherosclerosis and coronary heart disease. Amsterdam: Elsevier/North Holland Biomedical Press; 1981;Coronary heart disease prevention trials of diets to control hyperlipidemia. p. 197-210.

M2. Peto R, Yusuf S, Collins R. Cholesterol-lowering trial results in their epidemiologic context. Circulation 1985;72:451

M3. Yusuf S, Furberg CD. Olsson AG, editors.Atherosclerosis: Biology and Clinical Science. Edinburugh: Churchill Livingstone; 1987;Single factor trials: control through life - style changes. p. 389-92.

M4. Yusuf S, Culter J. Olsson AG, editors.Atherosclerosis: Biology and Clinical Science. Edinburugh: Churchill Livingstone; 1987;Single factor trials: drug studies. p. 393-8.

M5. Tyroler HA. Steinberg D, Olefsky JM, editors.Hypercholesterolemia and atherosclerosis: pathogenesis and prevention. New York: Churchill Livingstone; 1987;Lowering plasma cholesterol levels decreases risk of coronary heart disease: an overview of clinical trials. p. 99-117.

M6. Yusuf S, Wittes J, Friedman L. Overview of results of randomized clinical trials in heard disease. II. Unstable angina, heart failure, primary prevention with aspirin, and risk factor modification. JAMA 1988;260:2259-63.

M7. Oliver MF. Reducing cholesterol does not reduce mortality. J Am Coll Cardiol 1988;12:814-7.

M8. Tyroler HA. Overview of clinical trials of cholesterol lowering in relationship to epidemiologic studies. Am J Med 1989;84(suppl 4A):14-9.

M9. Holme I. An analysis of randomized trials evaluating the effect of cholesterol reduction on total mortality and coronary heart disease incidence. Circulation 1990;82:1916-24.

M10. Rossouw JE, Lewis B, Rifkind BM. The value of lowering cholesterol after myocardial infarction. N Engl J Med 1990;323:1112-9.

M11. Muldoon MF, Manuck SB, Matthews KA. Lowering cholesterol concentrations and mortality: a quantitative review of primary prevention trials. Br Med J 1990;301:309-14.

M12. Silberberg JS, Henry DA. The benefits of reducing cholesterol levels - the need to distinguish primary from secondary prevention: 2. Implications for heart disease prevention in Australia. Medical Journal of Australia 1991;155:670-4.

M13. Rossouw JE, Canner PL, Hulley SB. Deaths from injury, violence, and suicide in secondary prevention trials of cholesterol lowering [letter]. N Engl J Med 1991;325:1813

M14. Rossouw JE, Lewis B, Rifkind BM. Mortality experience in cholesterol-reduction trials. N Engl J Med 1991;324:923

M15. Muldoon MF, Manuck SB, Matthews KA. Mortality experience in cholesterol-reduction trials. N Engl J Med 1991;324:922-3.

M16. Silberberg JS, Henry DA. The benefits of reducing cholesterol levels: the need to distinguish primary from secondary prevention. 1. A meta-analysis of cholesterol-lowering trials. Med J Aust 1991;155:665-70.

M17. Davey Smith G, Pekkenen J. Should there be a moratorium on the use of cholesterol lowering drugs? Br Med J 1992;304:431-4.

M18. Ravnskov U. Cholesterol lowering trials in coronary heart-disease -frequency of citation and outcome. Br Med J 1992;305:15-9.

M19. Holme I. Meta-analysis of cholesterol reduction trials: coronary disease and mortality. Primary Cardiol 1992;18:63-70.

M20. MacManon S. Lowering cholesterol: effects on trauma death, cancer death and total mortality. Aust N Z J Med 1992;22:580-2.

M21. Lau J, Antman EM, Jimenez-Silva J, Kupelink B, Mosteller SF, Chalmers TC, JimenezSilva J, Kupelnick B, Mosteller F. Cumulative meta-analysis of therapeutic trials for myocardial infarction. New Engl J Med 1992;327:248-54.

M22. Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts: Treatments for myocardial infarction. J A M A 1992;268:240-8.

M23. Holme I. Relation of coronary heart disease incidence and total mortality to plasma cholesterol reduction in randomised trials; use of meta-analysis. British Heart Journal 1993;69 (supplement):542-7.

M24. Cucherat M, Boissel JP. Meta-analysis of results from clinical trials on prevention of coronary heart disease by lipid lowering interventions. Clin Trials Meta-Anal 1993;28:109-29.

M25. Smith GD, Song F, Sheldon TA, Song FJ. Cholesterol lowering and mortality: The importance of considering initial level of risk. Br Med J 1993;306:1367-73.

M26. Atkins D, Psaty BM, Koepsell TD, Longstreth WT, Larson EB. Cholesterol reduction and the risk for stroke in men: a meta-analysis of randomized, controlled trials. Ann Intern Med 1993;119:136-45.

M27. Macmahon S. Cholesterol reduction and death from noncoronary causes -evidence from randomized controlled trials. Australian And New Zealand Journal Of Medicine 1994;24:120-3.

M28. Law MR, Wald NJ, Thompson SG. By how much and how quickly does reduction in serum cholesterol concentration lower risk of ischaemic heart disease? Br Med J 1994;308:367-73.

M29. Law MR, Thompson SG, Wald NJ. Assessing possible hazards of reducing serum cholesterol. Br Med J 1994;308:373-9.

M30. Superko HR, Krauss RM. Coronary artery disease regression - convincing evidence for the benefit of aggressive lipoprotein management. Circulation 1994;90:1056-69.

M31. Truswell AS. Review of dietary intervention studies: effect on coronary events and on total mortality. Aust N Z J Med 1994;24:98-106.

M32. Howes LG, Simons LA. Efficacy of drug intervention for lipids in the prevention of coronary artery disease. Aust N Z J Med 1994;24:107-12.

M33. Hong MK, Mintz GS, Popma JJ. Limitations of angiography for analyzing coronary atherosclerosis progression or regression. Ann Intern Med 1994;121:348-54.

M34. Paterson RW, Paat JJ, Steele GH, Hathaway SC, Wong JC. Impact of intensive lipid modulation on angiographically defined coronary disease: clinical implications. South Med J 1994;87:236-42.

M35. Marchioli R, Prieto JC, Tognoni G. Surrogate end-points: the case of trials on coronary atherosclerotic plaque regression. Clin Trials Meta-Anal 1994;29:139-76.

M36. Gould AL, Rossouw JE, Santanello NC, Heyse JF, Furberg CD. Cholesterol reduction yields clinical benefit: A new look at old data. Circulation 1995;91:2274-82.

M37. Hebert PR, Gaziano JM, Hennekens CH. An overview of trials on cholesterol lowering and risk of stroke. Arch Intern Med 1995;155:50-5.

M38. Denke MA. Cholesterol-lowering diets: a review of the evidence. Arch Intern Med 1995;155:17-26.

M39. Rossouw JE. Lipid-lowering intervention in angiographic trials. The American Journal of Cardiology 1995;76:86c-92c.

M40. Holme I. Relationship between total mortality and cholesterol reduction as found by meta-regression analysis of randomized cholesterol-lowering trials. Controlled Clin Trials 1996;17:13-22.

M41. Rembold CM. Number-needed-to-treat analysis of the prevention of myocardial infarction and death by antidyslipidemic therapy. Journal of Family Practice 1996;42:577-86.

M42. Marchioli R, Marfisi RM, Carinci F, Tognoni G. Metaanalysis, clinical-trials, and transferability of research results into practice - the case of cholesterol-lowering interventions in the secondary prevention of coronary heart-disease. Archives of Internal Medicine 1996;156:1158-72.

M43. Hebert PR, Gaziano JM, Chan KS, Hennekens CH. Cholesterol lowering with statin drugs, risk of stroke, and total mortality. An overview of randomised trials. JAMA 1997;278:313-21.

A.II. RCTs that examine clinical endpoints

R1a. Singh RB, Rastogi SS, Verma R, Bolaki L, Singh R. An Indian experiment with nutritional modulation in acute myocardial infarction. Am J Cardiol 1992;69:879-85.

R1b. Singh RB, Rastogi SS, Verma R, Luxmi B, Reema BL, Singh R. Randomised controlled trial of cardioprotective diet in patients with recent acute myocardial infarction: results of one year follow up. Br Med J 1992;304:1015-9.

R2a. Marmorston J, Moore FJ, Hopkins CE, Kuzma OT, Weiner J. Clinical studies of long-term estrogen therapy in men with myocardial infarction. Proc Soc Exp Biol Med 1962;110:4000-8.

R3a. Stamler J, Pick R, Pick A, Kaplan BM, Berkson DM. Effectiveness of estrogens for therapy of myocardial infarction in middle-age men. JAMA 1963;183:632-8.

R4a. McCaughan D. The long-term effects of probucol on serum lipid levels. Arch Intern Med 1981;141:1428-32.

R5a. Harrold BP, Marmion VJ, Gough KR. A double blind controlled trial of clofibrate in the treatment of diabetic retinopathy. Diabetes 1969;18:285-91.

R6a. Carlson LA, Rosenhamer G. Reduction of mortality in the Stockholm ischaemic heart disease secondary prevention study by combined treatment with clofibrate and nicotinic acid. Acta Med Scand 1988;223:405-18.

R6b. Carlson LA, Danielson M, Ekberg I, Klintemar B, Rosenhamer G. Reduction of myocardial reinfarction by the combined treatment with clofibrate and nicotinic acid. Atherosclerosis 1977;28:81-6.

R6c. Rosenhamer G, Carlson LA. Effect of combined clofibrate-nicotinic acid treatment in ischaemic heart disease. Atheroscierosis 1980;37:129-38.

R7a. Leren P. The effect of plasma cholesterol lowering diet in male survivors of myocardial infarction. A controlled clinical trial. Acta Med Scand 1966;(suppl 466):1-92.

R8a. Research Committee. Low fat diet in myocardial infarction. A controlled trial. Lancet 1965;ii:501-4.

R9a. Burr ML, Gilbert JF, Holliday RM, Sweetnam PM, Elwood PC, Deadman NM. Effects of changes in fat, fish, and fibre intakes on death and myocardial reinfarction: diet and reinfarction trial (DART). Lancet 1989;ii:757-61.

R10a. Schock HK. The US veterans administration cardiology drug-lipid study: an interim report. Adv Exp Med Biol 1968;4:405-20.
Reprint : Not in File,

R10b. Detre KM, Shaw L. Long-term changes of serum cholesterol with cholesterol-altering drugs in patients with coronary heart disease. Veterans Administration drug-lipid cooperative study. Circulation 1974;50:998-1005.

R11a. Group of Physicians of the Newcastle upon Tyne Region. Trial of clofibrate in the treatment of ischaemic heart disease. Five year study. Br Med J 1971;iv:767-75.

R12a. Oliver MF, Boyd GS. Influence of reduction of serum lipids on prognosis of coronary heart-disease. A five year study using oestrogen. Lancet 1961;ii:499-505.

R13a. Acheson J, Huchinson EC. Controlled trial of clofibrate in cerebral vascular disease. Atherosclerosis 1972;15:177-83.

R14a. Watts GF, Lewis B, Brunt JNH, Lewis ES, Coltart DJ, Smith LDR. Effects of coronary artery disease of lipid-lowering diet, or diet plus cholestyramine, in the St Thomas atherosclerosis regression study (STARS). Lancet 1992;339:563-9.

R15a. Canner PL, Berge KG, Wenger NK, Stamler J, Friedman L, Prineas RJ. Fifteen year mortality in coronary drug project patients: long-term benefit with niacine. J Am Coll Cardiol 1986;8:1245-55.

R15b. Coronary Drug Project Research Group. Clofibrate and niacine in coronary heart disease. JAMA 1975;231:360-80.

R15c. Coronary Drug Project Research Group. The Coronary Drug Project: initial findings leading to modification of its research protocol. JAMA 1970;214:1303-13.

R15d. Coronary Drug Project Research Group. Natural history of myocardial infarction in the coronary drug project: long-term prognostic importance of serum lipid levels. Am J Cardiol 1978;42:489-98.

R16a. Dayton S, Pearce ML, Hashimoto S, Dixon WJ, Tomiyasu U. A controlled clinical trial of a diet high in unsaturated fat in preventing complications of atherosclerosis. Circulation 1969;40*suppl II):1-63.

R16b. Ederer F, Leren P, Turpeinen O, Frantz ID. Cancer among men on cholesterol-lowering diets. Lancet 1971;ii:203-5.

R16c. Pearce ML, Dayton S. Incidence of cancer in men on a diet high in polyunsaturated fat. Lancet 1971;ii:464-7.

R17a. Research Committee. Controlled trial of soya-bean oil in myocardial infarction. Lancet 1968;ii:693-700.

R18a. Research Committee of the Scottish Society of Physicians. Ischaemic heart disease: a secondary prevention trial using clofbrate. Br Med J 1971;iv:775-84.

R18b. Dewar HA, Oliver MF. Trial of clofibrate. Br Med J 1972;i:506

R19a. Sahni RS, Maniet AR, Voci C, Banka VS. Prevention of re-stenosis by lovastatin after successful coronary angioplasty. Am Heart J 1991;121:1600-8.

R19b. Fail PS, Sahni RS, Maniet AR, Voci C, Banka VS. The long-term clinical efficacy of lovastatin therapy following successful coronary angioplasty. Clin Res 1992;40:400A

R19c. Pitt B, Furberg C, McGovern M. Reduction in cardiovascular events during treatment with pravastatin: pooled analysis from coronary and carotid atherosclerosis intervention trials [abstract]. Eur Heart J 1995;15:S487

R19d. Pitt B, Ellis SG, Mancini J, Rosman HS, McGovern ME, for the PLAC I Investigators. Design and recruitment in the United States of a multicenter quantitative angiographic trial of Pravastatin to Limit Atherosclerosis in the Coronary Arteries (PLAC 1). Am J Cardiol 1993;72:31-5.

R20a. Dorr AE, Gundesen K, Schneider JC, Spencer TW, Martin WB. Colestipol hydrochloride in hypercholesterolemic patients - effect on serum cholesterol and mortality. J Chron Dis 1978;31:5-14.

R21a. Dumont JM. Effect of cholesterol reduction by simvastatin on progression of coronary atherosclerosis: design, baseline characteristics, and progress of the Multicenter Anti-Atheroma Study (MAAS). Controlled Clin Trials 1993;14:209-28.

R21b. Woodhill JM, Palmer AJ, Leelarthaepin B, et al. Kritchevsky D, Paoletti R, Holmes WL, editors.Drugs, Lipid Metabolism, and Atherosclerosis. New York, NY: Plenum Press; 1978;Low fat, low cholesterol diet in secondary prevention of coronary heart disease.

R22a. Rose GA, Thompson WB, Williams RT. Corn oil in treatment of ischaemic heart disease. Br Med J 1965;i:1531-3.

R23a. Brensike JF, Levy RI, Kelsey SF, Passamani ER, Richardson JM, Loh IK. Effects of therapy with cholestyramine on progression of coronary atheriosclerosis: results of the NHIBI type II coronary intervention study. Circulation 1984;69:313-24.

R24a. Frantz ID, Dawson EA, Ashman PL, Gatewood LC, Bartsch GE, Kuba K. Test of effect of lipid lowering by diet on cardiovascular risk. The Minnesota coronary survey. Atherosclerosis 1989;9:129-35.

R25a. Buchwald H, Varco RL, Matts JP, Long JM, Fitch LL, Campbell GS. Effect of partial bypass surgery on mortality and morbidity from coronary heart disease in patients with hypercholesterolemia. N Engl J Med 1990;323:946-55.

R25b. Buchwald H, Matts JP, Fitch LL. Program on the Surgical Control of the Hyperlipidemias (POSCH): design and methodology. J Clin Epidemiol 1989;42:1111-27.

R25c. Buchwald H, Matts JP, Hansen BJ, Long JM, Fitch LL, POSCH Group. Program on Surgical Control of the Hyperlipidemias (POSCH): recruitment experience. Controlled Clin Trials 1987;8:94s-104s.

R25d. Buchwald H, Matts JP, Fitch LL. Changes in sequential coronary arteriograms and subsequentcoronary events. JAMA 1992;268:1429-33.

R25e. Matts JP, Buchwald H, Fitch LL, Campos CT, Varco RL, Campbell GS, Pearce MB, Yellin AE, Smink RD, Jr., Sawin HS, Jr. Subgroup analyses of the major clinical endpoints in the Program on the Surgical Control of the Hyperlipidemias (POSCH): overall mortality, atherosclerotic coronary heart disease (ACHD) mortality, and ACHD mortality or myocardial infarction. J Clin Epidemiol 1995;48(3):389-405.

R26a. Blankenhorn DH, Nessim SA, Johnson RL, Sanmarco ME, Azen SP, Cashin-Hemphill L. Beneficial effects of combined colestipol-niacin therapy on coronary atherosclerosis and coronary venous bypass grafts. JAMA 1987;257:3233-40.

R26b. Cashin-Hemphill L, Mack WJ, Pogoda JM, Sanmarco ME, Azen SP, Blankenhorn DH. Beneficial effects of colestipol-niacin on coronary athersclerosis. JAMA 1990;264:3013-7.

R27a. Frick MH, Heininen OP, Huttunen JK, Koskinen P, Manttari M, Manninen V. Efficacy of gemfibrozil in dyslipidaemic subjects with suspected heart disease. An ancillary study in the Helsinki heart study frame population. Ann Med 1993;25:41-5.

R28a. Lipid Research Clinics Program. The lipid research clinics coronary primary prevention trial results. I. Reduction in incidence of coronary heart disease. JAMA 1984;251:351-64.

R29a. Frick MH, Elo O, Happa K, Heinonen AP, Heinsalmi P, Helo P. Helsinki heart study: primary prevention trial with gemfibrozil in middle-aged men with dyslipidemia. N Engl J Med 1987;317:1237-45.

R30a. Bradford RH, Shear CL, Chremos AN, Dujorne C, Franklin FA, Hesney M. Expanded clinical evaluation of lovastatin (EXCEL) study results. I. Efficiacy in modifying plasma lipoproteins and adverse event profile in 8245 patients with moderate hypercholesterolemia. Arch Intern Med 1991;151:43-9.

R30b. Bradford RH, Shear CL, Athanassios N. Expanded clinical evaluation of lovastatin (EXCEL) study: design and patient characteristics of a double-blind, placebo-controlled study in patients with moderate hypercholesterolemia. Am J Cardiol 1990;66:44-55B.

R30c. Tolbert JA. The cholesterol controversy. Br Med J 1992;304:713

R30d. Shear CL, Franklin FA, Stinnett S. Expanded Clinical Evaluation of Lovastatin (EXCEL) study results: effect of patient characteristics on lovastatin-induced changes in plasma concentrations of lipids and lipoproteins. Circulation 1992;85:1293-303.

R31a. Committee of Principal Investigation. A co-operative trial in the primary prevention of ischaemic heart disease using clofibrate. Br Heart J 1978;40:1069-118.

R31b. Heady JA, Morris JN, Oliver MF. WHO clofibrate/cholesterol trial: clarifications. Lancet 1992;ii:1405-6.

R31c. Report from the Committee of Principal Investigators. WHO cooperative trial on primary prevention of ischaemic heart disease with clofibrate to lower serum cholesterol: final mortality follow-up. Lancet 1984;ii:600-4.

R32a. Ornish D, Brown SE, Scherwitz LW, Billings JH, Armstrong WT, Ports TA. Can lifestyle changes reverse coronary heart disease? The lifestyle heart trial. Lancet 1990;336:129-33.

R33a. Kane JP, Malloy MJ, Ports TA, Phillips NR, Diehl JC, Havel RJ. Regression of coronary athersclerosis during treatment of familial hypercholesterolemia with combined drug regimens. JAMA 1990;264:3007-12.

R34a. Brown G, Albers JJ, Fisher LD, Schaefer SM, Lin JT, Kaplan C. Regression of coronary artery disease as a result of intensive lipid-lowering therapy in men with high levels of apoliprotein B. N Engl J Med 1990;323:1289-98.

R35a. Gross L, Figueredo R. Long-term cholesterol-lowering effect of colestipol resin in humans. J Am Geriatr Soc 1973;21:552-6.

R36a. Begg TB, Rifkind BM. Valutazione deila terapia con clofibrate nelle arteriopatie periferiche. Minerva Med 1971;62:3469-75.

R37a. Multiple Risk Factor Intervention Trial Research Group. Multiple risk factor intervention trial. Risk factor changes and mortality results. JMA 1982;248:1465-77.

R38a. Miettinen TA, Huttunen JK, Naukkarinen V. Multifactorial primary prevention of cardiovascular diseases in middle aged men. Risk factor changes, incidence, and mortality. JAMA 1985;254:2097-102.

R39a. World Health Organization European Collaborative Group. European collaborative trial of multifactorial prevention of coronary heart disease: final report of the 6-year results. Lancet 1986;i:869-72.

R39b. Kornitzer M, DeBacker G, Dramaix M, Kittel F, Thilly C, Graffar M. Belgian heart disease prevention project: incidence and mortality results. Lancet 1983;i:1066-70.

R39c. Kornitzer M, Rose G. WHO European collaborative trial of multifactorial prevention of coronary heart-disease. Preventive Medicine 1985;14:272-8.

R40a. Wilheimsen L, Berglund G, Elmfeldt D, Tibblin G, Wedel H, Pannert K, Vedin A, Wilhelmsson C, Werko L. The multifactor primary prevention trial in Goteborg, Sweeden. Eur Heart J 1986;7:279-88.

R41a. Holme I, Hjermann I, Helgeland A, Leren P. The Oslo study: diet and anti-smoking advice. Prev Med 1985;14:279-92.

R41b. Hjermann I, Byre KV, Holme I, Leren P. Effect of diet and smoking intervention on the incidence of coronary heart disease. Report from the Oslo Study Group of a Randomised Trial in Healthy Men. Lancet 1981;iii:1303-10.

R41c. Leren P. The Oslo diet-heart study. Eleven-year report. Circulation 1970;935-42.

R41d. Lehen P. The Oslo diet-heart study. Circulation 1970;42:935-42.

R42a. Waters D, Higginson L, Gladstone P, Kimball B, Le May M, Boccuzzi SJ, Lesperance J. CCAIT (Canadian Coronary Atherosclerosis Intervention Trial) study group. Effects of monotherapy with an HMG-CoA reductase inhibitor on the progression of coronary atherosclerosis as assessed by serial quantitative arteriography (CCAIT). Circulation 1994;89:959-68.

R42b. Waters D, Higginson L, Gladstone P. Design features of a controlled clinical trial to assess the effect of an HMG Co A reductase inhibitor on the progression of coronary artery disease. Controlled Clin Trials 1993;14:45-74.

R43a. MASS (Multicenter Anti-Atheroma Study) Investigators. Effects of simvastatin on coronary atheroma: the multicenter anti-atheroma study. Lancet 1994;344:633-8.

R43b. Woodhill JM, Palmer AJ, Leelarthaepin B, McGilchrist C, Blacket RB. Low fat, low cholesterol diet in secondary prevention of coronary heart disease. Adv Exp Med Biol 1978;109:317-30.

R44a. Pravastatin Multinational Study Group for Cardiac Risk Patients. Effects of pravastatin in patients with serum total cholesterol levels from 5.2 to 7.8 mmol/liter (200 to 300 mg/dl) plus two additional atherosclerotic risk factors. Am J Cardiol 1993;72:1031-7.

R45a. Shepherd J, Cobbe SM, Ford I, Isles CG, Lorimer AR, MacFarlane PW. Prevention of coronary heart disease with pravastatin in men with hypercholesterolemia. N Engl J Med 1995;333:1301-7.

R45b. The West of Scotland Coronary Prevention Study Group. A coronary primary prevention study of Scottish men aged 45-64 years: trial design. J Clin Epidemiol 1992;45:849-60.

R46a. Sacks FM, Pasternak RC, Gibson CM, Rosner B, Stone PH. HARP (Harvard Atherosclerosis Reversibility Trial group. Effect on coronary atherosclerosis of decrease in plasma cholesterol concentration in normocholesterolemic parients. Lancet 1994;344:1182-6.

R47a. Haskell WL, Alderman EL, Fair JL, Maron DJ, Sackey SF, Superko HR, Williams PT. Effects of intensive multiple risk factor reduction on coronary atherosclerosis and clinical cardiac events in men and women with coronary artery disease. Circulation 1994;89:975-90.

R48a. Pedersen TR, Kjekshus J, Berg K, Haghfelt T, Faergeman O, Thorgeirsson G, Pyorala K, Miettinen T, Wilhelmsen L, Olsson AG, et al. Randomized trial of cholesterol-lowering in 4444 patients with coronary-heart-disease - the scandinavian simvastatin survival study (4s). Lancet 1994;344:1383-9.

R48b. The Scandinavian Simvastatin Survival Study Group. Design and baseline results of the Scandinavian Simvastatin Survival Study of patients with stable angina and/or previous myocardial infarction. Am J Cardiol 1993;71:393-400.

R48c. Moccetti T, Malacrida R, Pasotti E, Sessa F, Genoni M, Barlera S, Turazza F, Maggioni AP. Epidemiologic variables and outcome of 1972 young patients with acute myocardial infarction. Data from the GISSI-2 database. Investigators of the Gruppo Italiano per lo Studio della Sopravvivenza nell'Infarto Miocardico (GISSI-2). Arch Intern Med 1997;157(8):865-9.

R48d. Scandinavian Simvastatin Survival Study Group. Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: the Scandinavian Simvastatin Survival Study (4S). Lancet 1995;345:1274-5.

R49a. Furberg CD, Adams HP, Applegate WB, Byington RP, Espeland MA, Hartwell T, Hunninghake DB, Lefkowitz DS, Probstfield J, Riley WA, et al. Effect of Lovastatin on early carotid atherosclerosis and cardiovascular events. Circulation 1994;90:1679-87.

R50a. Jukema JW, Bruschke AVG, van Boven AJ, Reiber JHC, Bal ET, Zwinderman AH, Jansen H. Effects of lipid lowering by pravastatin on progression and regression of coronary artery disease in symptomatic men with normal to moderately elivated serum cholesterol levels (Regression Growth Evaluation Statin Study [REGRESS]). Circulation 1995;91:2528-40.

R51a. Schuler G, Hambrecht R, Schlierf G, Niebauer J, Hauer K, Neumann J, Hoberg E. Regular physical exercise and low fat diet - effects of progression of coronary artery disease. Circulation 1992;86:1-11.

R52a. Blankenhorn DH, Azen SP, Kramsch SM, Mack WJ, Cashin-Hemphill L, Hodis HN, MARS Research Group. Coronary angiographic changes with lovastatin therapy (MARS). Ann Intern Med 1993;119:969-76.

R52b. Hodia HN, Mack WJ, Azen SP. Triglyceride- and cholesterol-rich lipoprotiens have a different effect on mild/moderate and severe lesion progression as assessed by quantitative coronary angiography in a controlled trial of lovastatin. Circulation 1994;90:42-9.

R53a. de Lorgeril M, Renaud S, Mamelle N. Mediterranean alpha-linoleic acid-rich diet in secondary prevention of coronary heart disease. Lancet 1994;343:1454-9.

R54a. Weintraub WS, Boccuzzi SJ, Klein L. Lack of effect of lovastatin on restenosis after coronary angioplasty. N Engl J Med 1994;331:1331-7.

R54b. Furberg CD, Byington RP, Crouse JR, Espeland MA. Pravastatin, lipids, and major coronary events. Am J Cardiol 1994;73:1133-4.

R54c. Pitt B, Mancini J, Ellis SG, Rosman HS, McGovern ME, for the PLAC I Investigators. Pravastatin Limitation of Atherosclerosis in the Coronary Arteries (PLAC I). J Am Coll Cardiol 1994;A793-42.

R54d. Furberg CD, Pitt B, Byington RP, Park J, McGovern ME, for the PLAC I and II Investigators. Reduction in coronary events during treatment with pravastatin. Am J Cardiol 1995;76:60C-3C.

R55a. Crouse JR, Byington RP, Bond MG. Pravastatin, lipids, and atherosclerosis in the carotid arteries: design features of a clinical trial with carotid atherosclerosis outcome. Control Clin Trials 1992;13:495-506.

R55b. Furberg CD, Crouse JR, Byington RP, Bond MG, Espeland MA. PLAC-2: effects of pravastatin on progression of carotid atherosclerosis and clinical events. J Am Coll Cardiol 1993;21:71 [Abs]

R55c. Crouse JRI, Byington RP, Bond MG. Pravastatin, lipids, and atherosclerosis in the carotid arteries (PLAC-II). Am J Cardiol 1995;75:455-9.

R55d. Byington RP, Jukema JW, Salonen JT. Reduction in cardiovascular events during pravastatin therapy: pooled analysis of clinical events of the Pravastatin Atherosclerosis Intervention Program. Circulation 1995;92:2419-25.

R55e. Byington, R.P., Furberg, C.D., Crouse, J.R., Espeland, M.A., and Bond, M.G. Pravastatin, lipids, and atherosclerosis in the carotid arteries (PLAC-II). *The American Journal of Cardiology* 76:54C-59C, 1995.

R56a. Andserson TJ, Meredith IT, Yeung AC, Frei B, Selwyn AP, Ganz P. The effect of cholesterol-lowering and antioxidant therapy on endothelium-dependent coronary vasomotion. N Engl J Med 1995;332:448-93.

R57a. Jones PH, Farmer JA, Cressman MD. Once-daily pravastatin in patients with primary hypercholesterolemia: a dose-response study. Clin Cardiol 1991;14:146-51.

R58a. Salonen R, Nyyssonen K, Porkkala E. Kuopio Atherosclerosis Prevention Study (KAPS): a population-based primary prevention trial of the effect of LDL lowering on atherosclerotic progression in carotid and femoral arteries. Circulation 1995;92:1758-64.

R58b. Salonen R, Nyyssonen K, Porkkala-Sarataho E, Salonen JT. The kuopio atherosclerosis prevention study (KAPS): Effect of pravastatin treatment on lipids, oxidation resistance of lipoproteins, and atherosclerotic progression. The American Journal of Cardiology 1995;76:34C-9C.

R59a. Sacks FM, Pfeffer MA, Moye LA, Rouleau JL, Rutherford JD, Cole TG, Brown L, Warnica JW, Arnold JM, Wun CC, et al. The effect of pravastatin on coronary events after

myocardial infarction in patients with average cholesterol levels. Cholesterol and Recurrent Events Trial investigators [see comments]. N Engl J Med 1996;335:1001-9.

R60a. Kallio V, Hamalainen H, Hakkila J, Lauurila OJ. Reduction in sudden deaths by a multifactorial intervention programme after acute myocardial infarction. Lancet 1979;ii:1091-4.

R61a. Bestehorn HP, Rensing UF, Roskamn H et al. The effect of Simvastatin on progression of coronary heart disease. The multicenter coronary intervention study (CIS). European Heart Journal. 1997;18:226-234.

R62a. Anonymous. Prevention of cardiovascular events and death with pravastatin in patients with coronary heart disease and a broad range of initial cholesterol levels. The Long-Term Intervention with Pravastatin in Ischaemic Disease (LIPID) Study Group. New England Journal of Medicine. 1998;339:1349-57.

R63a. Downs JR, Clearfield M, Weis S et al. Primary prevention of acute coronary events with lovastatin in men and women with average cholesterol levels: Results of AFCAPS. JAMA 1998;279:1615-1622.

R64a. Riegger G, Abletshauser C, Ludwig M et al. The effect of Fluvastatin on cardiac events in patients with symptomatic coronary artery disease during one year of treatment. Atherosclerosis 1999;144:263-270.

Cluster dynamic cohort crossover studies

X1a. Miettinen M, Turpeinen O, Karvonen MJ, Elosuo R, Paavilainen E. Effect of cholesterol-lowering diet on mortality from coronary heart disease and other causes - a twelve year clinical trial in men and women. Lancet 1972;ii:835-8.

X1b. Turpeinen O, Karvonen MJ, Pekkarinen M, Miettinen M, Elosuo R, Paavilainen E. Dietary prevention of coronary heart disease: the Finnish mental hospital study. Int J Epidemiol 1979;8:99-118.

A.III. The ten largest related Cohort studies investigating cholesterol levels and adverse events

C1a. Law, M.R., Wald, N.J., Wu, T., Hackshaw, A., and Bailey, A. A systematic underestimation of association between serum cholesterol concentration and ischaemic heart disease in observational studies: data from the BUPA study. *Br.Med.J.* 308:363-366, 1994.

C2a. Pocock, S.J., Shaper, A.G., and Phillips, A.N. Concentrations of high density lipoprotein cholesterol, triglycerides, and total cholesterol in ischaemic heart disease. *BMJ.* 298:998-1002, 1989.

C2b. Wannamethee G, Shaper AG, Whincup PH, Walker M. Low serum total cholesterol concentrations and mortality in middle aged British men. *BMJ.* 1995;**311**:409-13.

C3a. Rosengren, A., Welin, L., Tsipogianni, A., and Wilhelmsen, L. Impact of cardiovascular risk factors on coronary heart disease and mortality among middle aged diabetic men: a general population study. *BMJ.* 299:1127-1131, 1989.

C4a. Neaton, J.D. and Wentworth, D. Serum cholesterol, blood pressure, cigarette smoking, and death from coronary heart disease. *Arch Intern Med* 152:56-64, 1992.

C4b. Neaton, J.D., Blackburn, H., and Jacobs, D. Multiple Risk Intervention Trial Research Group. Serum cholesterol level and mortality findings for men screened in the Multiple Risk Factor Intervention Trial. *Arch Intern Med* 152:1490-1500, 1992.

C5a. Isles, C.G., Hole, D.J., Hawthorne, V.M., and Lever, A.F. Relation between coronary risk and coronary mortality in women of the Renfrew and Paisley survey: comparison with men. *Lancet* 339:702-706, 1992.

C6a. Shipley, M.J., Pocock, S.J., and Marmot, M.G. Does plasma cholesterol concentration predict mortality from coronary heart disease in elderly people? 18 year follow up in Whitehall study. *Br.Med.J.* 303:89-92, 1991.

C7a. Stemmermann, G.N., Chyou, P., Kagan, A., Nomura, A.M.Y., and Yano, K. Serum cholesterol and mortality among Japanese-American males: the Honolulu (Hawaii) heart program. *Arch Intern Med* 151:969-972, 1991.

C7b. Iribarren, C., Reed, D.M., Burchfiel, C.M., and Dwyer, J.H. Serum total cholesterol and mortality. Confounding factors and risk modification in Japanese-American men. *JAMA* 273(24):1926-1932, 1995.

C8a. Tornberg, S.A., Holm, L., Cartensen, J.M., and Eklund, G.A. Cancer incidence and cancer mortality in relation to serum cholesterol. *J Natl Cancer Inst* 81:1917-1921, 1989.

C9a. Goldbourt, U. and Yaari, S. Cholesterol and coronary heart disease mortality: a 23 year follow-up study of 9,902 men in Israel. *Arteriosclerosis* 10:512-519, 1990.

C10a. The Pooling Project Research Group Relationship of blood pressure, serum cholesterol, smoking habit, relative weight and ECG abnormalities to incidence of major coronary events: final report of the Pooling Project. *Journal of Chronic Diseases* 31:201-306, 1978.

SPECIAL NOTE

THIS ITEM IS BOUND IN SUCH A MANNER AND WHILE EVERY EFFORT HAS BEEN MADE TO REPRODUCE THE CENTRES, FORCE WOULD RESULT IN DAMAGE

A.V. Dataset for the 60 Cholesterol RCTs

id	year	br(chd)	br(grp)	pt-grp	intv_gl	intvg_2	f_up	nt	nc	totm_t	totm_c	chd_t	chd_c	%chol	col_abs	col_bse	sex	female%	multi_t	id
R1	1992	127.5	1	2	2	4	2.0	204	202	28	51	25	45	7.2	0.4	5.9	2	10	0	R1
R2	1962	110.4	1	2	1	2	2.5	285	147	70	38	62	35	99.9	99.9	99.9	1	0	0	R2
R3	1963	78.8	1	2	1	2	5.0	156	119	37	40	34	39	3.0	0.2	6.4	1	0	0	R3
R4	1981	72.7	1	2	1	6	1.0	88	30	2	3	2	2	8.0	0.7	7.9	1	0	0	R4
R5	1969	63.5	1	3	1	1	1.0	30	33	0	3	0	2	99.9	99.9	99.9	2	48	0	R5
R6	1988	62.1	1	2	1	1	5.0	279	276	61	82	47	73	13.0	0.8	6.4	2	20	1	R6
R7	1970	56.0	1	2	2	4	5.0	206	206	41	55	37	50	14.4	1.1	7.7	1	0	0	R7
R8	1965	50.9	1	2	2	4	3.0	123	129	20	24	17	20	6.6	0.6	6.8	1	0	0	R8
R9	1989	50.5	1	2	2	4	2.0	1018	1015	111	113	97	97	3.5	0.3	6.5	1	0	0	R9
R10	1968	50.3	1	2	1	6	3.2	427	143	81	27	71	23	6.3	0.6	6.2	1	0	0	R10
R11	1971	48.9	2	2	1	1	5.0	244	253	31	51	25	44	9.8	0.6	6.5	2	20	0	R11
R12	1961	43.7	2	2	1	2	5.0	50	50	17	12	13	10	12.4	0.8	6.1	1	0	0	R12
R13	1972	39.5	2	2	1	1	6.0	47	48	23	20	13	5	8.5	0.6	7.5	2	32	0	R13
R14	1992	36.4	2	2	3	5	3.0	30	60	0	4	0	4	17.5	1.1	7.2	1	0	1	R14
R15	1975	35.6	2	2	1	1	8.0	5552	2789	1025	723	826	632	9.5	0.6	6.5	1	0	1	R15
R16	1969	32.4	2	1	2	4	4.0	424	422	174	178	41	50	12.7	0.9	6.1	1	0	1	R16
R17	1968	29.1	2	2	2	4	2.0	199	194	28	31	25	25	14.3	1.0	7.0	1	0	0	R17
R18	1971	27.3	2	2	1	1	6.0	350	367	42	48	34	35	16.0	0.6	7.0	2	21	0	R18
R19	1991	26.5	2	2	1	3	2.0	79	78	4	5	2	4	12.8	0.7	5.4	2	15	0	R19
R20	1978	21.7	2	1	1	6	1.9	1149	1129	37	48	19	31	9.4	0.5	7.9	2	52	0	R20
R21	1978	21.5	2	2	2	4	5.0	221	237	39	28	35	26	4.0	0.3	7.3	2	12	0	R21
R22	1965	20.8	2	2	2	4	2.0	54	26	8	1	8	1	3.6	0.6	6.7	2	99.9	0	R22
R23	1984	17.5	2	2	1	6	5.0	71	72	5	7	5	6	16.3	0.9	8.4	2	19	1	R23
R24	1989	11.5	2	1	2	4	1.0	4541	4516	269	248	61	54	13.8	0.7	5.4	2	51	0	R24
R25	1990	10.9	2	2	3	5	9.7	421	417	49	62	32	44	22.5	1.5	6.5	2	9	0	R25
R26	1987	5.7	3	2	1	4	2.0	94	94	0	1	0	1	22.0	1.3	6.3	1	0	1	R26
R27	1993	5.1	3	2	1	6	5.0	311	317	19	12	17	8	8.5	0.8	6.9	1	0	0	R27
R28	1984	3.2	3	1	1	6	7.4	1906	1900	68	71	32	44	8.5	0.7	7.2	1	0	0	R28
R29	1987	1.9	3	1	1	6	5.0	2051	2030	44	43	14	19	10.1	0.7	6.9	1	0	0	R29
R30	1991	1.3	3	1	1	3	0.9	6582	1663	33	3	28	3	24.0	1.1	6.7	2	41	0	R30
R31	1978	1.2	3	1	1	1	5.3	5331	5296	236	181	91	77	9.0	0.6	6.8	1	0	0	R31
R33	1990	0.0	3	3	1	3	2.0	48	49	0	1	0	0	24.5	2.2	9.6	2	57	1	R33
R34	1990	0.0	3	2	1	3	2.5	94	52	1	0	1	0	24.9	1.5	7.0	1	0	1	R34
R35	1973	0.0	3	2	1	6	1.0	23	29	1	2	1	0	9.6	0.6	8.0	2	71	0	R35
R36	1971	24.2	2	2	1	1	5.0	76	79	4	10	4	9	10.8	0.7	6.5	99.9	99.9	0	R36
R37	1982	2.8	3	1	2	4	7.0	6428	6438	265	260	115	124	2.0	0.6	6.2	1	0	1	R37
R38	1985	0.3	3	1	1	1	5.0	612	610	10	5	4	1	6.3	0.4	7.1	1	0	1	R38
R39	1986	2.5	3	1	2	4	6.0	30489	26971	1325	1186	428	398	99.9	99.9	5.9	1	0	1	R39
R40	1986	4.8	3	1	1	1	10.0	10004	10011	1293	1304	462	453	6.5	0.0	6.5	1	0	1	R40

id	year	br(chd)	br(grp)	pt-grp	intv_g1	intvg_2	f_up	nt	nc	totm_t	totm_c	chd_t	chd_c	%chol	col_abs	col_bse	sex	female%	multi_f	id
R41	1970	4.5	3	1	2	4	5.0	604	628	16	24	6	14	12	1.1	7.7	1	0	1	R41
R42	1994	3.0	3	2	1	3	2.0	165	166	2	2	2	1	20	1.3	6.5	2	18	1	R42
R43	1994	8.2	3	2	1	3	4.0	193	188	4	11	4	6	22.7	1.4	6.4	2	12	1	R43
R44	1993	11.3	2	1	1	3	0.5	530	532	0	3	0	3	18.5	1.3	6.8	2	24	1	R44
R45	1995	4.6	3	1	1	3	4.9	3305	3293	106	135	50	73	20	1.4	7.0	1	0	1	R45
R46	1994	9.1	3	2	1	3	2.5	40	39	99.9	99.9	1	1	25.7	1.5	5.5	2	11	1	R46
R47	1994	4.9	3	2	1	3	4.0	145	155	3	3	2	3	16.4	0.9	6.0	2	14	1	R47
R48	1993	18.3	2	2	1	3	5.4	2221	2223	182	256	136	207	25	1.6	6.8	2	19	1	R48
R49	1994	4.4	3	2	1	3	3.0	460	459	1	8	0	6	27	1.5	6.1	2	48	1	R49
R50	1995	5.8	3	2	1	3	2.0	450	434	5	7	3	5	19	1.2	6.0	1	0	1	R50
R51	1992	0.0	3	2	2	4	1.0	56	57	2	1	2	0	10.1	0.3	6.1	1	0	1	R51
R52	1993	4.0	3	2	1	3	2.0	123	124	2	1	1	1	32.2	1.8	6.0	2	9	1	R52
R53	1994	24.3	2	2	2	4	2.3	302	303	8	20	3	16	0.03	0.0	6.5	2	9	0	R53
R54	1994	99.9	99.9	2	1	3	0.5	203	201	3	1	99.9	99.9	99.9	99.9	5.3	2	28	0	R54
R55	1995	5.0	3	2	1	3	3.0	206	202	4	6	3	3	21.6	1.3	5.9	2	22.5	1	R55
R57	1991	0.0	3	2	1	3	0.2	83	42	1	0	1	0	17.5	1.3	7.7	2	50	0	R57
R58	1995	4.5	3	1	1	3	3.0	224	223	3	4	2	3	21.0	1.5	6.0	1	0	0	R58
R59	1996	13.1	2	2	1	3	5.0	2081	2078	180	196	112	130	20.0	1.1	5.4	2	14	0	R59
R60	1979	115.3	1	2	2	4	3.0	188	187	41	56	35	55	+8.3	0	6.0	2	20	0	R60
R61	1997	7.0	3	2	1	3	2.3	129	125	1	4	1	2	0.8	0.05	6.3	1	0	1	R61
R62	1998	14.7	2	2	1	3	6.1	4512	4520	498	633	287	373	18.0	1.0	5.6	2	17	0	R62
R63	1998	0.88	3	1	1	3	5.2	3304	3301	80	77	11	15	18.4	1.15	5.71	2	15	1	R63
R64	1999	22.72	2	2	1	3	1.0	187	178	2	4	2	4	12.5	1.0	7.45	2	38.4	0	R64

Variable descriptions and details

id - Identification number for each RCT. This number corresponds to the reference given in the bibliography. i.e. 1 above refers to R1 in the bibliography

year - The date toe firs trial report was published

br(chd) - The baseline risk of CHD mortality in the control group - see details in body of report for how this was calculated.

br(grp) - This is a categorical variable based on br(chd) above. Baseline risk is categorised into high = 1, medium =2, and low = 3 risk. High risk corresponds to > 50, medium risk between 50 and 10, and low risk < 10 CHD deaths per 1000 person years.

pt-grp - Predominant patient group: 1 = primary prevention (no pre-existing CHD), 2 = secondary prevention (pre-existing CHD), and 3 = diabetic.

intv_g1 - Intervention type variable 1. Categorises the trials by the most aggressive treatment given to the treatment group: 1 = drug, 2 = diet and 3 = surgery. Note: on if drugs were used when diet produced an un-satisfactory response, it is coded as a drug trial.

intvg_2 - Intervention type variable 2. More detailed categorisation than above: 1 = fibrate drugs, 2 = hormones, 3 = statins, 4 = diet, 5 = surgery, 6 = other drugs. Again, trial was defined by the most aggressive treatment given, if more than one was given

f_up - Follow-up - the duration of the trial in years

nt - the number of patients in the treatment arm of the trial

nc - the number of patients in the control arm of the trial

totm_t - the number of deaths (total mortality) in the treatment arm of the trial

totm_c - the number of deaths (total mortality) in the control arm of the trial

chd_t - number of deaths from CHD in the treatment arm of the trial

chd_c - number of deaths from CHD in the control arm of the trial

%chol - The percentage reduction in serum cholesterol between the treatment and control arms of the trial

col_abs - the absolute reduction in serum cholesterol between the treatment and control arms of the trial (in mmol/l)

col_bse - baseline level of cholesterol in patients (average between the 2 arms in mmol/l)

sex - indicates the sex of patients in the trial: 1 = all male; 2 = mixed (there are no all female trials)

female% - The percentage patient composition which is female (0 corresponds to all male trials)

multi_t - Indicator of number of treatments administered. 0 = single intervention, 1= multiple treatment regime

Missing values - 9.9 or 99.9 corresponds to missing values in any variable.

A.VI. Cohort study dataset

ID	OR	SE_OR	ln_OR	se_ln_or	Followup
C2	0.80	0.038	-0.223	0.048	14.8
C4	0.63	0.008	-0.462	0.013	7.1
C5	0.77	0.044	-0.261	0.057	15
C7	0.71	0.039	-0.342	0.055	19
C9	0.86	0.025	-0.151	0.029	23

Variable descriptions and details

ID - study id - note C5 is date from the males in the study only

OR - Odds ratio derived

SE_OR - standard error of the odds ratio

ln_OR - log of the OR

se_ln_or - standard error of the log odds ratio

Followup - Length of study

A.VII. WinBUGS code for fitting 3-level hierarchical model to the cholesterol data (no covariates)

```
model
{
# Randomised controlled trials
# ========================
 for (i in 1:R) {

     chdt[i]       ~ dbin(pt[i],nt[i])
     chdc[i]       ~ dbin(pc[i],nc[i])
     logit(pc[i]) <- mu[i]
     logit(pt[i]) <- mu[i] + delta[i]

    mu[i]       ~ dnorm(0.0,1.0E-5)
  delta[i] ~ dnorm(theta[1],tau.theta[1])
}


# Cohort studies model
# ==================
for (i in 1:C) {
     coh.prec[i]  <- 1/(coh.se.lnor[i]*coh.se.lnor[i])

     coh.lnor[i]  ~ dnorm(coh.psi[i],coh.prec[i])

     coh.psi[i]  ~ dnorm(theta[2], tau.theta[2])
}


# Combining both sources
# ====================
for(i in 1:T) {
    theta[i] ~dnorm(mean,tau.mean)
   tau.theta[i] ~ dgamma(0.001,0.001)
    var.theta[i] <- 1/tau.theta[i]
or[i] <- exp(theta[i])
}

mean  ~ dnorm(0,1.0E-6)
or.overall <- exp(mean)
tau.mean ~ dgamma(0.001,0.001)
var.mean<-1/tau.mean
}


Data

list(coh.lnor=c(-0.223,-0.462,-0.261,-0.342,-0.151),
   coh.se.lnor=c(0.048,0.013,0.057,0.055,0.029),
          R=60,C=5,T=2)

nt[]      nc[]      chdt[]    chdc[]
204       202       25        45
285       147       62        35
156       119       34        39
88        30        2         2
30        33        0         2
279       276       47        73
206       206       37        50
```

123	129	17	20
1018	1015	97	97
427	143	71	23
244	253	25	44
50	50	13	10
47	48	13	5
30	60	0	4
5552	2789	826	632
424	422	41	50
199	194	25	25
350	367	34	35
79	78	2	4
1149	1129	19	31
221	237	35	26
54	26	8	1
71	72	5	6
4541	4516	61	54
421	417	32	44
94	94	0	1
311	317	17	8
1906	1900	32	44
2051	2030	14	19
6582	1663	28	3
5331	5296	91	77
94	52	1	0
23	29	1	0
76	79	4	9
6428	6438	115	124
612	610	4	1
30489	26971	428	398
10004	10011	462	453
604	628	6	14
165	166	2	1
193	188	4	6
530	532	0	3
3305	3293	50	73
40	39	1	1
145	155	2	3
2221	2223	136	207
460	459	0	6
450	434	3	5
56	57	2	0
123	124	1	1
302	303	3	16
206	202	3	3
83	42	1	0
224	223	2	3
2081	2078	112	130
188	187	35	55
129	125	1	2
4512	4520	287	373
3304	3301	11	15
187	178	2	4

Initial values

```
list(tau.theta=c(0.5,0.5), theta=c(0,0), mean=0, tau.mean=1,
delta=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0),
mu=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)
)
```

A.VIII. WinBUGS code to fit general model of DuMouchel for application to the computer-based reminder systems meta-analysis

```
model

{

for(i in 1:lines){                                                    # loops round all lines of data

y[i] ~ dnorm(eq[study[i], treatment[i], outcome[i], group[i], time[i],repeat[i]], w[i])

eq[study[i], treatment[i], outcome[i], group[i],time[i], repeat[i]] <-

        # fixed effects
        out[outcome[i]] + trt[treatment[i]] +

        # random effects
        z.trt.out[treatment[i], outcome[i]]   +   z.study[study[i]] +
        z.study.trt[study[i], treatment[i]] +  z.study.out[study[i], outcome[i]] +
        z.study.group[study[i], group[i]] +  z.study.time[study[i],time[i]]

                                                                                                    }

for(x in 1:treatments){
for(m in 1:outcomes) {
z.trt.out[x, m] ~ dnorm(0,tau[1]);
}
}

for(m in 1:studies) {
 z.study[m] ~ dnorm(0,tau[2]);
}

for(m in 1:studies) {
for(x in 1:treatments) {
z.study.trt[m, x] ~ dnorm(0,tau[3]);
}
}

for(m in 1:studies) {
for(x in 1:outcomes) {
z.study.out[m, x] ~ dnorm(0,tau[4]);
}
}

for(m in 1:studies) {
for(x in 1:groups) {
z.study.group[m, x] ~ dnorm(0,tau[5]);
}
}

for(m in 1:studies) {
for(x in 1:times) {
z.study.time[m, x] ~ dnorm(0,tau[6]);
}
}


for(x in 1:6) {
tau[x] ~ dgamma(0.001,0.001)
sigma[x] <- 1.0 / sqrt(tau[x])
}


for(m in 1:outcomes) {
```

```
out[m] ~ dnorm(0,0.001);
}


trt[treatments] <-0      # This is required for a unique solution -  corner centring
for(m in 1:treatments-1) {
trt[m] ~ dnorm(0,0.001);
}


###############################
# Computing parameters of interest  #
###############################

#one outcome - compare each treatment to placebo for each outcome
for (m in 1:outcomes)  {  # loops for each  outcome
for (k in 1:treatments-1) {
lnor[k,m] <- trt[k] - trt[treatments] + z.trt.out[k,m] - z.trt.out[treatments,m]
or[k,m]<- exp(lnor[k,m])
}
}


}
```

Data

```
list(lines=330, outcomes=6, treatments=4, groups= 5, times=3, studies=19,

study=c(1,1,2,2,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,5,5,5,5,5,5,5,5,5,5,5,5,
5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,7,7,7,7,7,7,7,7,7,7,7,7,8,8,8,8,8,8,9,9,9
,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,10,10,10,10,10,10,10,10,10,10,10,10,10,10,10,10,10,10,10,10,11,11,11,11,11,11,11,11,11,11,11,
11,11,11,11,11,11,11,11,11,11,11,11,11,11,12,12,12,12,13,13,14,14,14,14,14,14,14,14,14,14,14,14,14,14,14,14,14,14,14,14,14,1
4,14,15,15,15,15,15,15,15,15,15,15,15,15,15,15,15,15,15,15,15,15,15,15,15,15,15,15,15,15,15,15,15,15,15,15,15,15,15,15,15,16
,16,16,16,16,16,16,16,16,16,16,16,16,16,16,16,16,16,16,16,16,16,16,16,16,16,16,16,16,16,16,16,16,16,16,16,16,16,16,16,16,16,
16,17,17,17,17,17,17,17,17,17,17,18,18,18,19,19,19),

repeat=c(1,1,1,1,1,1,1,2,1,3,4,1,2,5,11,1,2,1,3,4,1,2,5,1,1,2,2,1,1,1,1,2,2,1,1,2,2,1,1,1,1,2,2,1,1,2,2,1,1,1,1,2,2,1,1,2,1,1,2,1,1,1,1,2,1
,1,2,1,1,1,1,2,1,1,2,1,1,1,1,2,1,1,2,1,1,1,1,2,1,1,2,1,1,1,2,3,1,2,1,2,1,2,1,2,1,2,3,1,2,1,2,1,2,1,2,1,1,1,2,1,2,1,1,1,2,1,2,1,1,1,1,1,1,1,1,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,3,1,2,1,2,1,2,3,1,2,1,2,1,2,3,1,2,1,2,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,2,1,1,1,1,1,1,1
,2,1,1,2,3,1,1,1,2,1,1,2,3,1,1,1,2,1,1,2,3,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,
2,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,1,1,1,1,1,1,1,1,2,2,2,2,3,3,3,3,3,3,3,3,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1),

group=c(1,2,1,2,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,2,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,2,3,3,3,3,3,3,3,3,3,3,1,1,1,1,1,1,1,1,2,2,2,2
,2,2,2,2,3,3,3,3,3,3,3,3,4,4,4,4,4,4,4,4,5,5,5,5,5,5,5,5,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,2,2,2,2,1,1,1,1,1,1,2,2,2,2,2,2,2,1,1,1,2,2,2,1,2,
3,4,1,2,3,4,1,2,3,4,4,3,2,1,4,3,2,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,3,3,3,3,1,2,3,4,5,1,2,3,4,5,1,2,3,4,5,1,2,3,4,5,1,2,3,4,5,1,2,3,4,1,2,1
,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,3,3,3,3,1,1,2,2,3,3,4,4,1,1,2,2,3,3,4,4,1,1,2,2,3,3,4,4,1,1,2,2,3,3,4,4,1,1,2,2,3,3,4,4,1,1,2,2,1,1,2,
2,1,1,2,2,1,1,2,2,1,1,2,2,1,1,2,2,1,1,2,2,1,1,2,2,1,1,2,2,1,1,2,2,1,1,2,2,1,1,1,1,1,1,1,1,1,1,1,1,2,3,1,2,3
),

time=c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2
,2,2,2,2,2,2,2,2,2,2,2,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1
,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,
1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,2,1,1,1,1,1,1),

outcome=c(1,1,3,3,4,5,2,2,6,2,2,3,6,2,4,5,2,2,6,2,2,3,6,2,3,3,3,3,5,5,4,4,4,4,3,3,3,3,5,5,4,4,4,4,3,3,3,3,5,5,4,4,4,4,6,2,6,5,3,3,4,1,6,2,
6,5,3,3,4,1,6,2,6,5,3,3,4,1,6,2,6,5,3,3,4,1,6,2,6,5,3,3,4,1,4,4,4,5,5,3,3,1,1,2,2,4,4,4,5,5,3,3,1,1,2,2,4,3,5,4,6,6,4,3,5,4,6,6,4,3,5,4,3,5,4
,4,4,4,6,6,6,6,2,2,2,2,5,5,5,5,3,3,3,3,4,4,4,5,5,3,3,4,4,4,5,5,3,3,4,4,4,5,5,3,3,1,1,1,1,1,5,5,5,5,5,6,6,6,6,6,6,6,6,6,1,1,1,1,1,3,3,3,3,3,
3,2,2,4,6,6,6,3,5,2,2,4,6,6,6,3,5,2,2,4,6,6,6,3,5,1,1,1,1,1,1,1,1,4,4,4,4,4,4,4,4,3,3,3,3,3,3,3,3,5,5,5,5,5,5,5,5,6,6,6,6,6,6,6,6,2,2,2,2,2,2
,2,2,3,3,3,3,6,6,6,6,4,4,4,4,3,3,3,3,5,5,5,5,1,1,1,1,1,1,1,1,2,2,2,2,1,1,1,1,1,1,6,6,4,4,3,3,5,5,6,6,6,6,6,6),

treatment=c(4,2,4,2,4,4,4,4,4,4,4,4,4,4,2,2,2,2,2,2,2,2,2,2,4,2,4,2,4,2,4,2,4,2,4,3,4,3,4,3,4,3,4,3,4,1,4,1,4,1,4,1,4,1,4,4,4,4,4,4,4,4,4,4,
4,4,4,4,4,4,3,3,3,3,3,3,3,4,4,4,4,4,4,4,4,4,4,2,2,2,2,2,2,2,2,4,4,4,4,4,4,4,4,4,4,4,4,4,1,1,1,1,1,1,1,1,1,1,1,4,4,4,4,4,4,2,2,2,2,2,2,2,2,2,2,1,1,1,4
,3,2,1,4,3,2,1,4,3,2,1,4,3,2,1,4,3,2,1,4,4,4,4,4,4,4,4,3,3,3,3,3,3,3,3,2,2,2,2,2,2,2,2,4,2,3,3,4,4,2,3,3,4,4,2,3,3,4,4,2,3,3,4,4,2,3,3,4,4,3,2,1,4,
1,4,4,4,4,4,4,4,4,2,2,2,2,2,2,2,2,1,1,1,1,1,1,1,1,4,4,4,2,4,3,4,1,4,4,4,2,4,3,4,1,4,4,4,2,4,3,4,1,4,4,4,2,4,3,4,1,4,4,4,2,4,3,4,1,4,3,4,2,4,3
,4,2,4,3,4,2,4,3,4,2,4,3,4,2,4,3,4,2,4,3,4,2,4,3,4,2,4,3,4,2,4,3,4,2,4,3,4,2,4,1,4,1,4,1,4,1,4,1,4,3,3,4,3,3
),

y=c("vector of 330 responses omitted for confidentiality reasons"),

w=c("vector of 330 weights omitted for confidentiality reasons")
)
```

Initial values

```
list(out=c(-1,-1,-1,-1,-1,-1), trt=c(0.5,0.5,0.5,NA), tau=c(5,5,5,5,5,5))
```

Appendix B

Annotated WinBUGS code for fitting the net-benefit model in Chapter 8

```
model
        {#

        ## Meta-analysis for OR of reduction in stroke by using warfrin

          for( i in 1 : Num.benefit ) {#
                rc[i] ~ dbin(pc[i], nc[i])
                rt[i] ~ dbin(pt[i], nt[i])
                logit(pc[i]) <- mu[i]
                logit(pt[i]) <- mu[i] + delta[i]
                mu[i] ~ dnorm(0.0,1.0E-5)
                delta[i] ~ dnorm(d, tau)
          }
          d ~ dnorm(0.0,1.0E-6)
          tau ~ dgamma(0.001,0.001)
          var <- 1 / tau
                or<- exp(d)
                rrr <- 1 - or

### Meta-analysis for the estimate of harm

### Non rcts ###

for(j in 1 : Num.harm) {#

 bleeds[j] ~ dpois(expected[j])
log(expected[j]) <- beta[j] + log(fup[j])
 beta[j]~ dnorm(h,tau.h)
}

### RCTs ###

 for(j in 1 :  rct.harm) {#

 bleeds.r[j] ~ dpois(expected.r[j])
 log(expected.r[j]) <- beta.r[j] + log(nt[j]*ave.fup[j])
  beta.r[j]~ dnorm(h,tau.h)
 }

h ~ dnorm(0,1.0E-6)
harm <- exp(h)
tau.h ~ dgamma(0.001,0.001)
var.h <- 1 / tau.h


#### Quality of life following stroke ####

        r.qolcat[1:16] ~ dmulti(p[1:16], N.qol)

        N.qol <- sum(r.qolcat[1:16])
        for (i in 2:16) {#
        p[i] ~ dbeta(1,1)
}
        p[1] <- 1- sum(p[2:16])
```

```
qol.cat ~ dcat(p[])
qol.score <-score[qol.cat]
qol.reduction <- 1- qol.score
outcome.ratio <- 1/(qol.reduction)


#### Specifying risks estimated by clinical and echocardiographic predictors ####

clin.risk[1] ~ dnorm(-3.68888,8.468168)
clin.risk[2] ~ dnorm(-2.63109,23.36712)
clin.risk[3] ~ dnorm(-1.73727,14.03161)
clin.echo.risk[1] ~ dnorm(-4.60517,1.712246)
clin.echo.risk[2] ~ dnorm(-2.81341, 26.34225)
clin.echo.risk[3] ~ dnorm(-1.68201, 16.90100)


#### Estimation of net-benefit for levels of risks due to predictors #########

for (i in 1:3) {#
        nb.clin.risk[i] <-(exp(clin.risk[i])*rrr)-(harm*outcome.ratio)

        nb.clin.echo.risk[i] <-(exp(clin.echo.risk[i])*rrr)-(harm*outcome.ratio)
}


### Estimation of net-benefit for range of theoretical absolute levels of risk ###

for (i in 1:50) {#
nb.risk[i] <-(risk[i]*rrr)-(harm*outcome.ratio)
}


}
```

Data

```
        list(nt=c(212,210,187,335,260), nc=c(208,211,191,336,265), Num.benefit=5,
rt=c(2,6,5,4,4),rc=c(13,18,11,21,19),

Num.harm=22, bleeds=c(4,3,2,1,0,2,1,4,0,0,3,0,0,0,0,0,0,3,5,2,10),
fup=c(45.3,96.8,93.4,121.8,52.5,857.0,362.5,418.3,168.0,109.8,1872.0,8.3,139.0,6.0,8.3,113.3,
24.0,49.0,506.0,726.2,365.7,904.0),
rct.harm=3, ave.fup = c(2.094339623,1.238095238,1.26684492), bleeds.r=c(1,1,2),
r.qolcat=c(1, 54,0,0,10,13,0,0,0,0,10,0,0,0,12,0),

score=c(0.99,0.98,0.866666667,0.8,0.733333333,0.666666667,0.6,0.533333333,0.466666667,
0.4,0.333333333,0.266666667,0.2,0.133333333,0.066666667,0),

risk=c(0.01,0.03,0.05,0.07,0.09,0.11,0.13,0.15,0.17,0.19,0.21,0.23,0.25,0.27,0.29,0.31,0.33,0.35,0.37,0.3
9,0.41,0.43,0.45,0.47,0.49,0.51,0.53,0.55,0.57,0.59,0.61,0.63,0.65,0.67,0.69,0.71,0.73,0.75,0.77,0.79,0.81
,0.83,0.85,0.87,0.89,0.91,0.93,0.95,0.97,0.99))
```

Inits

```
list(tau=0.5, d=0, mu=c(0,0,0,0,0), h=0.1, tau.h = 0.1,
beta=c(0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1),
beta.r=c(0.1,0.1,0.1),
        p= c(NA,0.05,0.05,0.05,0.05,0.05,0.05,0.05,0.05,0.05,0.05,0.05,0.05,0.05,0.05,0.05,0.05,0.05))
```