An Exploration of Evidence Synthesis Methods for Adverse Events

Fiona Claire Warren MBBS; MSc

Submitted for the degree of Doctor of Philosophy

at the University of Leicester

April 2010

An Exploration of Evidence Synthesis Methods for Adverse Events

Fiona Claire Warren MBBS; MSc

Abstract

Adverse events following the use of medical interventions are a major source of concern for patients, healthcare professionals and pharmaceutical companies. Therefore, evidence synthesis of potential adverse events are very important in determining whether an association exists, and the strength of such an association. It is also desirable to be able to quantitatively balance potential harms against the benefits of the intervention. However, standard statistical techniques for meta-analysis are often unsuitable when applied to datasets where the primary intervention is an adverse event.

A review of standard meta-analysis methods, including Bayesian methods, is conducted. The specific challenges of meta-analysis in relation to adverse events datasets are described, with some of the main areas of contention being sparsity of events, subgroup analysis, class effects with regard to drug interventions, and issues related to time factors within the individual studies. Methods used in existing meta-analyses where the primary outcome is an adverse event have also been reviewed; this demonstrates the methods already used in this field and highlights some of their limitations, and where the methods could be extended.

In the light of the reviews of methods and previous meta-analyses, four case-studies are performed. The first uses data from GlaxoSmithKline to investigate a potential relationship between paroxetine and suicidality, using many of the standard methods for comparison purposes. The second uses individual patient data for a time-to-event analysis of anti-TNF drugs used for rheumatoid arthritis. This clinical example is extended by the use of Mixed Treatment Comparisons for within-class comparisons, and to asses the effect of dose. Finally, a harm–benefits model is used to asses the interplay of risk of endometrial cancer against breast cancer recurrence for tamoxifen users. These models present novel ways of analysing adverse events data and demonstrate some of the difficulties in their use.

Acknowledgements

I have had the great good fortune to be guided and supported throughout the duration of this research by two very talented and committed supervisors, Professors Keith Abrams and Alex Sutton. I would like to thank them both in equal measure for their very different contributions. I am grateful to Alex for his very different way of seeing the world and encouraging me to approach issues in ways that would not come naturally to me. Keith has been invaluable in providing his vision in developing the methodological concepts and his expertise in programming that has enabled me to put into practice many of these ideas. It has been a privilege and a pleasure to work with them both.

I am grateful to Su Golder of the Centre for Reviews and Dissemination, University of York, for her assistance in identifying the previous meta-analyses of adverse outcomes used in the systematic review in this thesis. I would also like to thank Dr Tim Bongartz and Dr Eric Matteson of the Mayo Clinic College of Medicine, Minnesota, USA, and Dr Daniel Mines of Wyeth Research, Pennsylvania, USA, for their collaboration in the meta-analyses regarding etanercept treatment for rheumatoid arthritis, and for their permission to include this work in this thesis.

Finally, I would like to thank my mother, Jennifer Warren, for all her support (and proof-reading!) throughout the past few years.

Contents

1

Introduction

1.1 Aims and objectives of this thesis

The underlying source of concern of this thesis is as follows:

Drugs and other clinical interventions often have unintended outcomes, either beneficial or harmful. Increased understanding of such effects would enable improvements in clinical decision-making, both in general guidelines and in making decisions regarding individual patients. Diverse sources of information regarding such unintended outcomes exist, and it is necessary to synthesise data from all available sources to optimise the knowledge on which to base decisions. There are, however, many difficulties in bringing together disparate information types, and these must be explored and overcome as rigorously as possible to facilitate the development of the most accurate knowledge base possible.

This hypothesis leads to the following broad aims:

1. to identify specific methodological issues in relation to evidence synthesis and decision-modelling in the field of unintended effects of clinical therapies;
2. to identify examples of drugs, and other forms of intervention, with potential for adverse effects that can be analysed as case studies;

3. to address the identified issues in a coherent way, using the case studies as a means of exploring how these issues can be tackled; and
4. to develop methods for evidence synthesis for adverse outcomes and decision-modelling for use when balancing treatment efficacy against the potential for harmful effects.

These aims can be further clarified into more specific objectives:

1. to review the literature regarding evidence synthesis and decision-modelling methods;
2. to identify specific challenges in this area that are directly relevant to adverse outcomes;
3. to review previous meta-analyses where an adverse event has been the primary outcome;
4. to develop techniques to address methodological issues identified;
5. to investigate at least three case studies, incorporating the methods and techniques developed, thus facilitating refinement and improvement in the methodologies;
6. the development of methods for evidence synthesis that can be adapted to meet the requirements of each scenario;

7. to develop methods that will enable the useful incorporation of a broad array of data from a variety of sources into a decision-modelling and evidence synthesis model; and
8. further assessment and critique of these methodologies by additional case studies, ideally including surgical or public health examples as well as pharmacological examples.

These aims and objectives will be revisited at appropriate intervals during this project in order to evaluate progress and the extent to which they are being fulfilled.

1.2 Methodological background

In the field of statistics applied specifically to biomedical outcomes, in many clinical scenarios it is reasonable to apply pre-existing established methods to the analysis of data derived from that scenario. Some clinical scenarios may, however, present a unique set of challenges, some of which may be similar to those presented by other scenarios and some of which may not. Alternatively, a particular set of statistical issues may occur in conjunction in relation to a particular clinical field.

Evidence synthesis for adverse events presents several challenges, many of which occur in other clinical areas, but nevertheless have a particular association with data related to adverse events. In this situation it is reasonable to use the analytical problems to drive the methodological development, but it is also valuable to consider methods applied to evidence synthesis in other contexts and then to apply these to adverse events datasets.

The development of evidence synthesis methods with adverse events specifically in mind has been considered previously (Sutton *et al*. 2002). This area is an example of how problems of analysis generated by certain types of clinical data can create an agenda for development of statistical techniques. These techniques may then be applied to other clinical areas that may have similar issues in relation to evidence synthesis. This thesis extends and broadens methodologies previously used, by applying them to datasets that are centred on an adverse outcome.

The Cochrane Collaboration (Loke *et al*. 2008), in its remit to undertake systematic reviews, has considered adverse events as an area worthy of consideration as a defined clinical area that requires specific methodology, but the emphasis is on the non-statistical elements, such as choice of outcomes, study types and search strategies. Evidence synthesis methods with the specific aim of application to adverse events have not been brought together previously, and despite the fact that this area is very wide, it is the intent of this thesis to address this issue.

1.3 Clinical background

Drug treatments and other interventions that have the ability to provide therapeutic benefits also have, in many cases, the power to harm. To quote Paracelsus (1493–1541):

> **Alle Ding' sind Gift und nichts ohn' Gift; allein die Dosis macht, dass ein Ding kein Gift ist.**

In English, this translates as: all things are poison and nothing is without poison, only the dose permits something not to be poisonous. However, there is often wide individual variation in the dose of a substance that will bring about adverse effects, and these may be related to a variety of individual characteristics. These may include age, sex, environmental factors or genetic factors. The medications used clinically with the intention to cure disease, relieve symptoms or both, have as much potential to cause damage as other substances used purely to cause harm, with no known therapeutic value.

However, in some cases there may be additional positive benefits associated with the use of a particular drug that are unforeseen by the prescriber or manufacturer of the drug. Such benefits may only come to light through anecdotal evidence or through observational studies. Unintended outcomes of an intervention may therefore be positive or negative.

Thinking more broadly, other medical interventions, which may include surgical procedures or public health programmes, may also have unintended outcomes, either beneficial or harmful.

In prescribing interventions for a particular patient, the intent of the clinician is to maximise the positive effects whilst minimising, and preferably eliminating entirely, any damaging effects. Hence, a thorough understanding of the efficacy of that intervention is essential, but no less important is an understanding of any harmful effects. In the same way that individual patients may require different doses of a drug to achieve a therapeutic effect, there are variations between individuals in the doses required to cause a deleterious outcome.

With appropriate knowledge, the clinician can, in discussion with the patient when there are issues of personal preference to be taken into consideration, make decisions that will promote the intended therapeutic effects whilst reducing the potential for harm. In effect, each patient must 'play the odds' between benefits and disbenefits, armed with knowledge of the intervention's potential for both, and with knowledge of their own characteristics that influence the actions of an intervention. In many cases this may be a delicate balancing act; in cases where the potential disbenefits are of a minor or self-limiting nature there is a clear incentive to use an intervention with a possible beneficial effect, but if there is a risk, however minimal, of serious or long-term sequelae, then it is less clear-cut to determine whether an intervention should be used.

Clinical decision-making is facilitated by clear-cut guidelines for each intervention, which make full use of all available data and are applicable to all patients taking account of his/her relevant characteristics. Such guidelines can of course be modified to take into account individual patient preferences, for example how well they tolerate certain symptoms or adverse outcomes from interventions. As discussed in the next section, the overarching aim of this work is to develop appropriate and sound statistical methodologies for evidence synthesis, in cases where the primary outcome of interest is an adverse event. Such methodologies would be of use at various levels of clinical application, for example:

1. detecting 'signals' from the data, in terms of any indication of concern regarding an intervention;
2. investigating further when there are sufficient data to raise concern about a possible adverse event;
3. combining data from multiple sources to increase power of statistical analyses and promote validity of conclusions; and

4. methods for quantifying harms against benefits, which could be used to inform clinical decision-making at the level of general guidelines and individual patient management.

1.4 Outline of the thesis

1.4.1 Methodological aspects

As a first step it is possible to identify some specific technical issues that require addressing, usually as a result of the nature of the data available in this area. Such issues include:

1. incorporation of sparse data (e.g. from randomised controlled trials (RCTs) and observational studies);
2. combination of heterogeneous data sources (RCTs, observational studies (pharmacoepidemiology), anecdotal reporting by the 'yellow card' system (see Section 2.2), and case reports);
3. assessment of harms and benefits by evidence synthesis and decision-modelling;
4. subgroup analysis (e.g. pharmacogenetics, age, sex);
5. evidence synthesis for a specific intervention with multiple clinical indications;
6. combining individual patient data with summary statistics;
7. addressing reporting bias (suppression of adverse events or not proactively looking for adverse events);
8. estimation of dose effects;
9. evidence synthesis across multiple outcomes (i.e. multiple adverse effects);
10. comparison across different drugs from the same class; and
11. consideration of adverse events over different time-periods (time-course aspects).

These aspects of evidence synthesis will be addressed in more depth in Chapter 5, with the aim of conducting case-studies using specific examples of adverse events datasets to illustrate selected methodological issues.

1.4.2 Potential case studies

From initial scrutiny of recent literature, there are many examples of a therapeutic intervention that has been linked with some unforeseen effect, either beneficial or harmful. Some of these include:

1. hormone replacement therapy and breast cancer (Million Women Study Collaborators 2003);
2. statins and cancer (Bonovas *et al*. 2005; 2008);
3. non-steroidal anti-inflammatory drugs and severe gastro-intestinal bleeding (Hernández-Díaz & García Rodríguez 2000);
4. selective serotonin re-uptake inhibitors and suicidality (Gunnell *et al*. 2005);
5. tamoxifen and endometrial cancer (Braithwaite *et al*. 2003);
6. warfarin (for atrial fibrillation) and risk of haemorrhage and cerebrovascular accident (Cooper *et al*. 2006);
7. rheumatoid arthritis drugs (anti-tumour necrosis factor drugs) and cancer (Bongartz *et al*. 2006); and
8. antiretroviral drugs and abnormalities of lipid metabolism (Miller *et al*. 2000).

Selected examples from the above list are chosen to serve as case-studies, illustrating statistical methods that are relevant to each example. The issue of suicidality and selective serotonin reuptake inhibitors (SSRIs) is discussed in Chapter 7. Anti-tumour necrosis factor (anti-TNF) drugs used in rheumatoid arthritis and the associated risk of cancer is addressed in Chapters 8–10. Tamoxifen use in breast cancer patients, with the aim of reducing recurrence, and the associated risk of endometrial cancer is investigated using a harm–benefit model in Chapter 11.

1.4.3 Structure of the thesis

A discussion of the clinical problems of adverse events is provided in Chapter 2, in order to demonstrate the nature and extent of the issues involved. Established meta-analysis methods are considered in Chapter 3, with Bayesian meta-analysis methods presented in Chapter 4. These three chapters comprise a framework to bring together the clinical issue of adverse events with meta-analysis methodology, and the specific challenges that may arise in relation to synthesis of evidence where the primary outcome is an adverse or unintended event; this comprises Chapter 5.

A systematic review of previous meta-analyses of datasets where the main outcome is an adverse or unintended event is presented in Chapter 6. This is followed by a case-study where multiple methods of meta-analysis are applied to the same dataset, with the aim of making a comparison across methods (Chapter 7). The use of individual patient data (IPD) in a time-to-event meta-analysis is the main focus of Chapter 8, and this approach is contrasted with meta-analysis methods with a binary outcome (and no time-to-event element). The clinical issue of interest in Chapter 8 is extended in Chapters 9 and 10 by looking more deeply at the different forms of treatment and how they may be compared; this closer scrutiny of the clinical issues required a commensurate extension of statistical methods, to include mixed treatment comparison (MTC) methods and Bayesian hierarchical models.

Modelling benefits of a clinical intervention in parallel to potential deleterious effects is the subject of Chapter 11, using a Bayesian net-benefit model. Finally, Chapter 12 aims to set the developments of this thesis in a wider context, to discuss how this thesis has developed existing methodology, and to propound ways in which evidence synthesis methods for adverse events may be extended.

1.5 Summary

Inherent in many clinical interventions is the potential for significant adverse outcomes (and in some cases unforeseen benefits). Evidence synthesis methods to provide as much insight as possible into adverse outcomes are essential to inform clinical practice, at the level of general guidelines, and individual patient

management. The development of suitable statistical methods for evidence synthesis of adverse outcomes, taking into account the specific difficulties presented by such data, is the basis for this work, and underpins the more specific aims and objectives.

2

Background to adverse events

2.1 Overview of adverse drug reactions

2.1.1 Definitions and classifications

Ideally, a statistical overview of methods appropriate to adverse outcomes would include a variety of clinical interventions, including pharmacological, surgical and public health programmes. However, in this work, the main focus is on drug therapies and their unintended, usually harmful, effects. The concept of an 'adverse drug reaction' (ADR) therefore requires some thought and definition.

Firstly, however, it is helpful to define what is meant by a drug. A definition is cited by a World Health Organisation (WHO) publication of 1969, taken from the earlier work of Borda *et al*. (1968). A drug is defined as:

> **'any substance or product that is used or intended to be used to modify or explore physiological systems or pathological states for the benefit of the recipient'.**

The WHO defines an adverse drug reaction as being (WHO 1969):

> **'one which is noxious, unintended, and which occurs at doses normally used in man for prophylaxis, diagnosis or therapy.'**

A similar concept is the adverse event, which has been described by Michel *et al*. (2004) as an unintended injury caused by a disease process and which resulted

in death, life-threatening illness, disability at time of discharge, admission to hospital or prolongation of hospital stay. This definition appears to have been derived from consulting earlier work of several authors.

Similar definitions are used by Leape *et al*. (1998):

> **[an adverse drug event is] an injury, large or small, caused by the use (including non-use) of a drug. There are two types of adverse drug events (ADEs); those caused by errors and those that occur despite proper usage. If an adverse drug event is caused by error it is by definition preventable. Non-preventable adverse drug events (injury but no error) are called adverse drug reactions (ADRs).**

Hence, ADEs encompass both preventable and non-preventable occurrences.

Further terminology has included the possible ADR, which refers to an ADR that follows a reasonable temporal sequence and for which the ADR is a known response to the drug, although the response may also be explained by the patient's clinical state (Lazarou *et al*. 1998, citing Karch & Lasagna 1975). A serious ADR is one that requires hospitalisation, prolongs hospitalisation, is permanently disabling or results in death (Lazarou *et al*. 1998). This definition is similar to that of the adverse event as defined by Michel (2004) but referring only to ADRs rather than ADEs.

As the aim of this study is to investigate unintended effects of drugs that have been correctly prescribed and used, the ADEs are not of direct relevance. Hence, based on the above terminology, this work will include the events denoted as ADRs, that is, non-preventable events. This definition excludes all errors of prescribing (such as incorrect dosages or other clinical errors such as inaccurate diagnosis leading to the wrong drug being prescribed), administration (by health professionals), compliance (where the medication is being self-administered), and overdose (intentional or unintentional). Also excluded are cases where a particular drug is ineffective for a certain patient for any reason.

However, it is possible that some studies reporting on adverse reactions may include both ADRs and ADEs combined, hence leading to unwanted events being taken into account and thus to heterogeneity in the data sources.

Some further subclassifications of ADRs are also useful. A useful classification system has been set out by Edwards & Aronson (2000). Six classes of adverse reactions are set out, the first being dose-related effects (predictable from the pharmacological action of a drug). As known overdoses are not being considered here, this category would encompass adverse reactions that occur in certain individuals at doses that are thought to be safe. This may be due to the specific metabolism of that individual which again could be related to a pharmacogenetic effect. The second category is non-dose related, which includes uncommon effects that are not related to the pharmacological action of a drug. Such effects include immunological actions and idiosyncratic reactions such as malignant hyperthermia. The third category includes dose- and time-related effects, the chronic effects that occur after long usage that are related to cumulative dose of the drug. The fourth category is time-related effects that are usually related to the dose effect and usually become apparent some time after the drug has been taken. This includes teratogenesis and carcinogenic effects. The fifth category is withdrawal effects, usually occurring quite soon after cessation of the drug. The sixth category includes unexpected failures of therapy.

For the purposes of this study only the first four categories are of relevance as withdrawal effects and therapeutic failures are not the focus here. Whilst it is interesting to consider the possible causes of an ADR, the main emphasis of this study is to analyse existing methods and further develop new methods to identify possible patterns of ADRs, using all available data. Where feasible, some consideration will be given to potential mechanisms for a specific drug being associated with a specific ADR, especially if this mechanism can be plausibly tied in with the identified patterns of occurrence.

2.1.2 The extent of adverse drug reactions

A review and meta-analysis of studies investigating the incidence of adverse drug reactions in hospital in-patients was conducted by Lazarou *et al*. (1998). This study was restricted to primary data from the USA, and excluded errors of prescribing and administration/compliance and overdose or drug abuse. Only serious (requiring hospitalisation or leading to permanent disability) or fatal reactions were included. The patient base included hospital in-patients who suffered an adverse reaction whilst in hospital, and those in the community

who experienced an adverse reaction requiring hospitalisation. (This effectively excludes those ADRs that occur in the community and result directly in death, without hospital admission.) Only prospective studies were included, with a total of 39 primary studies being used. Random effects models were employed throughout. The authors found an incidence for serious ADRs of 6.7%, with a 95% confidence interval (CI) of 5.2%; 8.2%, and for fatal ADRs of 0.32% (95% CI 0.23%; 0.41%). It was also estimated that for the year 1994, there were 106 000 (95% CI 76 000; 137 000) fatal ADRs, with 1 547 000 (95% CI 1 033 000; 2 060 000) hospital admissions due to ADRs. Based on these figures, the authors concluded that ADRs were between the fourth and sixth commonest cause of death in the USA for that year.

The meta-analysis by Lazarou *et al*. (1998) has been severely criticised by Kvasz *et al*. (2000). This critique discovered many problematic issues in the original meta-analysis. These problems were related to both the study design and statistical analysis. Among these was the problem of heterogeneity between studies. There were many sources of heterogeneity including differences in definitions of adverse reactions, and in preventability of events. There was also heterogeneity in the types of patient included (for example, adults or children) and in the types of hospital or ward involved in the study. There was some variability between the studies surrounding the question of what actually constituted an ADR. In some studies, only 'probable/definite' ADRs were included, while in some others, 'probable' ADRs were also included.

From a statistical analysis (as opposed to study design) perspective, a major problem was the derivation of numerators and denominators. Some studies did not directly report the number of patients with adverse events, and these were therefore estimated for the purpose of the meta-analysis, and hence may over- or under-estimate the true numerators, thus adding an extra source of variability and lack of precision. Also, the denominator used in the meta-analysis was considered to be questionable. The chosen denominator was the total number of patients admitted to hospital, whereas a more appropriate choice would have been the total number of patients receiving prescription drugs. A major statistical issue, related to the problem of studies where no events occur (which will be considered further in Chapters 3, 5, 7, 8, 9 and 10) is the fact that a study of fatal adverse events only will exclude many studies with zero events, as was the case in the study by Lazarou *et al*. (1998). In this way, the risk

of fatal adverse events will be severely overestimated. Several other statistical areas of contention in the Lazarou *et al*. (1998) paper are highlighted by Kvasz *et al*. (2000). These include ascertainment and publication bias, and extrapolation with small numbers, leading to errors and invalid confidence intervals.

In the UK, the issue of number of hospital admissions due to ADRs in the UK was addressed by Pirmohamed *et al*. (2004). This study was conducted over a 6-month period between 2001 to 2002, in two hospitals in the Merseyside area. Again, deliberate overdoses and episodes of non-compliance were excluded. Admissions thought to be due to an ADR accounted for 6.5% of admissions (95% CI 6.2%; 6.9%). It was also found that the median age for admissions due to an ADR (76 years, interquartile range 65–83) was greater than that for other types of admission (66 years, interquartile range 46–79). The proportion of women was also significantly higher in the ADR group (59%) than in the non-ADR group (52%), p-value < 0.0001. Of all the ADRs, 80% (95% CI 78%; 92%) were deemed to have been the direct cause of the admission, whereas for the remaining 20% (95% CI 18%; 22%) of admissions, the ADR was identified through screening and although not the direct cause of the admission, may have been a contributory factor. In total, 2.3% of the ADR-related admissions died as a direct result of the ADR. The most common cause of death was gastrointestinal bleeding, due to a variety of drug therapies including aspirin (alone or in combination with other drugs), paroxetine and warfarin. Therefore, ADR-related deaths represented 0.15% (95% CI 0.1%; 2%) of all patients admitted during the study period.

In the study discussed above (Pirmohamed *et al*. 2004), the issue of avoidability was also raised, with some very interesting discoveries. It was found that only 28% (95% CI 25%; 30%) of ADRs resulting in admission were classed as 'unavoidable' using the classification system of Hallas *et al*. (1990). Of the ADR admissions, 9% (95% CI 7%; 10%) were classified as 'definitely avoidable', while 63% (95% CI 60%; 66%) were 'possibly avoidable'. In the classification of Hallas *et al*. (1990) an ADR was classed as definitely avoidable if it would have been prevented by the application of current good medical practice; 'possibly avoidable' indicated that an ADR could have been prevented by efforts exceeding current good medical practice.

Other studies also highlight the extent of the problem of injury due to ADRs. Weingart *et al*. (2000) discuss a benchmark study carried out in the USA by

Brennan *et al*. (1984), who reviewed the medical charts of over 30 000 patients in New York. They found that adverse events due to medical treatment resulted in injury that prolonged hospital stay or produced disability at time of death occurred in 3.7% of admissions.

An Australian study (Wilson *et al*. 1995) is also reviewed by Weingart *et al*. (2000), and is a similar investigation of adverse events in Australia. Investigating the records of over 14 000 admissions to 28 Australian hospitals in 1995, it was discovered that an adverse event occurred in 16.6% of admissions, resulting in disability in 13.7% and death in 4.9%. Of these adverse events, 51% were thought to have been preventable [it is not stated by what definition].

In a similar study based in London (Vincent *et al*. 2001), it was found that 110 of 1014 (10.8%) patients experienced an adverse event (this study was not restricted to only adverse drug reactions), with an overall total of adverse events at 119 (11.7%). The means of identification of adverse events was the review of medical records from four different medical specialties.

The studies by Pirmohamed *et al*. (2004) and Wilson *et al*. (1995), as reviewed by Weingart *et al*. (2000), indicate that a large proportion of ADRs could have been avoided by increased medical vigilance. This creates an enormous potential for the prevention of ADRs, by developing and applying improved guidelines for good medical practice. Such guidelines can only be developed by promoting statistical focus on issues that directly relate to evidence synthesis of adverse events data.

2.2 Difficulties of investigating adverse drug reactions

2.2.1 Outline

Due to the nature of the events of interest, there are some inherent problems to be addressed when studying ADRs. These include:

1. Identification of such potential ADRs;
2. Causality;
3. Reporting of potential adverse events; and

4. Identification of studies including ADRs.

Each of these issues is discussed in detail below. A consideration of the merits of different data sources is also included.

2.2.2 Identification, causality and reporting of adverse drug reactions

Identification of ADRs depends on a series of factors.

- The ADR must be detectable to the clinician in terms of clinical signs or to the patient in terms of symptoms.
- Alternatively, the ADR must be detectable through other means, such as pathology or imaging tests, and these tests must be actually performed either routinely or due to clinical suspicions of an ADR.
- There must be a connection, however tentative, made between the potential ADR and the drug. This connection will usually be made on the basis of knowledge of the drug's pharmacological effects, previous ADRs, or the temporal association between administration of the drug and the onset of the ADR. At this stage, it is the making of the connection between the clinical event and the drug administration that is the key to the identification of a possible ADR.
- In the post-marketing phase, potential ADRs may be unrecognised by clinicians, or unreported by patients. One reason for this non-reporting may be the difficulty inherent in differentiating an adverse reaction to a medication from the symptoms of the original clinical condition that was being treated.

Identification of potential ADRs may therefore be complicated by a variety of factors. These include the possible ascription of symptoms, signs and results of pathological tests to the underlying condition rather than the drug therapy, a lengthy time delay between onset of treatment and onset of the adverse event (especially when the drug is taken over a long period of time and the adverse event progresses slowly to a detectable stage, such as some forms of cancer), or if there is no prior indication or information that a certain drug will result in a certain adverse reaction in some users.

With regard to causality, some putative ADRs may in fact be due to the underlying condition that is being treated or due to other causes such as transient viral infections. Most case reports of ADRs report suspected cases. Various methods have been put forward for the determination of the likelihood of a causal relationship between a drug and an adverse reaction. These include criteria based on the association in time or place between drug administration and the event, the drug's pharmacology, medical plausibility based on signs, symptoms and other tests, and likelihood or exclusion of other causes (*the* Uppsala Monitoring Centre 2000).

Further discussion of the definitions, terminology and issues surrounding causality with regard to adverse drug reactions is provided by Edwards & Aronson (2000) and Nebeker *et al.* (2004).

Reporting of potential ADRs may occur in many settings. In the drug testing phase, randomised controlled trials (RCTs) may be a source of ADR identification, but trials may not be specifically designed to focus on safety as well as efficacy. Therefore, potential ADRs may be overlooked, due to failure of the organisers to proactively elicit information from participants or to look for clinical signs of any ADRs.

Observational studies that look for ADRs following the licensing and marketing of a drug, whilst it is being used in treatment of patients, have issues regarding recall (for retrospective studies) and ascertainment of ADRs in cohort studies.

A third way in which ADRs may be reported is through the more informal systems of reporting when a drug is in general use, which rely on spontaneous vigilance from drug users and healthcare professionals. These systems include the 'yellow card' reporting system[1],run by the Medicines and Healthcare products Regulatory Agency (MHRA) and the Commission on Human Medicines (CHM). Yellow cards may be submitted by healthcare professionals or by drug users including patients or their parents or carers.

[1]Yellow Card Scheme - MHRA (2010). Available at [March 2010]: http://yellowcard.mhra.gov.uk/

In the USA, there is also a method for spontaneous reporting of potential ADRs, implemented by the Food and Drug Administration (FDA)[1]. The Adverse Event Reporting System (AERS) database includes over 4 million spontaneous reports of ADRs compiled from 1969 to 2009.

However, for spontaneous reporting, the potential ADR must first be identified as such, and then there must be a willingness to report the incident. It may be the case that incidents causing severe effects are more likely to be reported than more trivial ADRs, due to the possibility of receiving significant compensation from the drug manufacturer, which may be necessary to safeguard the future care of the patient or support of their family in the case of a fatality. There are also issues surrounding quality of documentation, duplicate reporting and coding when assessing case reports of potential ADRs (*the* Uppsala Monitoring Centre 2000).

A study investigating the identification of ADEs (including adverse events related to prescribing and dispensing errors) has been carried out by Jha *et al*. (1998), based in the USA. This study is interesting as it compares a specially-designed computer monitoring system for the identification of ADEs with a simulated voluntary reporting system, and a system of [drug] chart review. The investigation occurred over 8 months in 1994–1995, in nine hospital units. Out of 627 identified adverse events, the greatest number (398; 65%) was discovered by review of drugs charts. The computer monitoring system identified 275 (45%), whilst the voluntary reporting system yielded the smallest number of adverse events, just 23, 4% of the total.

Considering the overlap between the reporting methods, the overlap between computer monitoring and chart review was 76 adverse events (12%), whereas the overlap between computer monitor and voluntary report was lower at three events (1%). This study highlights the difficulties inherent in identifying adverse events, interestingly demonstrating that different methods of reporting pick up different events, with relatively little overlap between them. One of the most pertinent points is that voluntary reporting yielded the smallest number of adverse events. Therefore, the necessity of proactive detection methods for

[1] US Food and Drug Administration (2010). Available at [March 2010]: http://www.fda.gov/default.htm

adverse events is evident. The chart review system was the most labour intensive detection method in terms of staff hours spent in implementing the scheme. The chart review required 55 person-hours per week, while the computer monitoring system required 11 person-hours. As the computer monitoring system identified 69% of the cases identified by chart review, this system may be a time-economical method of identifying ADEs.

In a similar vein, another study from the USA compared two different methods for the detection of adverse events, including, but not exclusively, ADRs (O'Neil *et al*. 1993). In this investigation, two systems were used to detect adverse events, one involving retrospective scrutiny of the medical records by trained medical record analysts. The second method involved physician self-reporting of adverse events occurring in patients under their management. Using the first method, there were 85 adverse events related to medical management (from an initial patient group of 3128 admissions); using the second method there were 124 adverse events. When the physician-reported adverse events were further investigated, it was found that only 89 patients met the definition of an adverse event. The overlap between the 81 patients identified by the medical record scrutiny and the 89 patients identified by the physician self-reporting method was low, with 41 adverse events being identified using both methods. Of the 48 adverse events reported by the physicians (not identified by the medical record scrutiny), only 14 fulfilled the criteria of the medical record analysts. This study illustrates the difficulties in identifying adverse events of all descriptions, and the potential for errors and discrepancies in the collection of data.

The two studies described above were conducted in a clinical environment where the drugs were being used in a hospital setting, following licensing and marketing. Evidently, a specific drug will not be marketed unless it has passed through rigorous tests for efficacy and tolerability through the phases of clinical trials.

An investigation into a possible association between source of funding for a clinical trial and whether the drug being tested is recommended as a treatment of choice was conducted by Als-Nielsen *et al*. (2003). The trials funded by for-profit organisations were more likely to recommend the drug as treatment of choice compared to non-profit organisations with an odds ratio of 5.3 (95% CI 2.0; 14.4). This result did not, however, appear to be associated with lack of detection (or active suppression of reporting) of adverse effects. For-profit organisations reported significantly higher numbers of adverse events in the

treatment arm when compared to non-profit organisations. This may be related to quality of reporting of the studies and the increased emphasis placed by the drugs companies on Good Clinical Practice guidelines.

2.2.3 Identification of studies involving adverse events

The final difficulty in gathering data on ADRs and other adverse events is that of correctly identifying both primary studies and secondary reviews regarding such events.

Golder *et al*. (2006a) addressed this problem by developing search strategies in the online databases Medline and Embase, with the aim of identifying items related to adverse effects. Several different approaches were developed (to be used in conjunction with the name of the drug in question), including searching for specific adverse events, using thesaurus subheadings (for example for adverse events or toxicity), using search terms synonymous with adverse events, and using search strategies previously published (citing Badgett *et al*. 1999 and Loke *et al*. 2002), which use strategies involving the name of the study design in question. It was found that highly sensitive search strategies also had the disadvantage of low precision [therefore requiring significant user time to sort out the relevant records from the non-relevant records]. When excluding search terms based on specific adverse events (as these are often not known in advance of the search), the most sensitive Medline search strategy involved the use of 'floating' (not linked to the drug name) subheadings in conjunction with search terms synonymous with adverse events. In Embase the most sensitive search strategy involved subheadings linked to the drug name in conjunction with terms synonyms for adverse events.

Golder *et al*. (2006b) have also investigated searching methods for two of the main sources of information in the field of adverse events. These were Database of Abstracts of Reviews of Effects (DARE), the main focus of which is on medical interventions and their effects, and the Cochrane Database of Systematic Reviews (CDSR), which promotes systematic reviews of medical treatments with the aim of contributing to evidence-based medicine. The authors constructed several different search strategies for both databases and evaluated them by comparison to the 'gold standard' (GS), the relevant records identified in each database using both electronic searching and hand-searching of records not re-

trieved using the search strategies. The statistics used for evaluation of the search strategies were sensitivity (number of GS records retrieved/number of GS records indexed in database under investigation as a percentage) and precision (number of GS records retrieved/total number of records retrieved as a percentage). A low precision was found when searching CDSR (0–3%); when searching DARE precision was higher (16–71%). However, in both DARE and CDSR, a lower sensitivity tended to be associated with a higher precision. In DARE, sensitivity ranged from 4–85%, while in CDSR sensitivity ranged from 0–64%.

Observational studies are also a very useful source of data in identifying ADRs, but it is also highly likely that the same difficulties in retrieving relevant papers may be encountered. This is was addressed by Wieland & Dickersin (2005), who conducted several search strategies on Medline to ascertain the sensitivity and precision (using the same definitions as used by Golder *et al*. (2006b) above). The example used was a systematic review of oral contraceptive use in relation to breast cancer. There were 58 references in the systematic review. It was found that it was possible to develop appropriate Medline search strategies using both keywords (Medline subject heading (MeSH) terms) and text word searches that would retrieve all 58 papers (100% sensitivity). However, these strategies had very low precision (0.9 when using keywords and 0.8 when using text words). This was due to the large total number of references retrieved (6120 using MeSH terms and 7240 using text words). Such a large number of retrieved references is clearly inefficient and would require a large amount of time in manually extracting relevant papers based on titles, abstracts and full papers. Using MeSH terms, the highest precision was 11%, with a total of 424 references retrieved, of which 48 were included in the meta-analysis. The authors highlight two major problems in data retrieval. There was a significant problem with indexing procedures on Medline, with inadequate indexing of interventions used. Another problem was incomplete reporting in the paper of all data in the study, similar to selection bias in a clinical trial. The authors of the original meta-analysis had identified relevant studies by contacting principal investigators and colleagues in their endeavours to identify all relevant studies. Hence, improvements are required in indexing by Medline and in reporting of observational studies, so that all interventions and outcomes are included.

Setting aside the differences and limitations among search strategies, the most important determining factor in whether or not a relevant reference will be retrieved is the quality and accuracy of indexing with respect to adverse events. A study by Derry *et al*. (2001) investigated the quality of indexing of clinical trials with regard to adverse events. A sample of 107 papers that were known to report clinical trial data on adverse events were followed up in Medline and Embase. Of the 107 papers, 100 were indexed on Medline and 88 on Embase, with 81 papers being included on both databases. Considering the indexing of papers (which is used in keyword search strategies), 53 of the 100 papers were indexed correctly as having adverse event data, with a similar proportion of the Embase papers (43 out of 88) doing likewise. The inconsistency of indexing between Medline and Embase was apparent, with only 30 out of the 81 papers included in both databases being indexed with adverse events keywords in both databases. In many cases the paper was not correctly indexed in either database (25 out of 81 papers), while 26 out of 81 papers were indexed in only one of the two databases. [This finding highlights the desirability of conducting searches in both databases.] In addition to keyword searching, it is also possible to retrieve references by searching on text words in title and/or abstract. Only 62 out of the 107 papers made reference to adverse events in the title or abstract. Hence, manual searching of titles and abstracts would not have identified 45 of these references. Using a combined search strategy of keywords and free text searching of titles and abstracts, 82 of the 107 papers would have been found, leaving 25 that would not have been retrievable. The authors conclude that researchers will therefore need to manually check all published references on clinical trials to ascertain whether there is any data included on adverse events. This approach would be very time-consuming and hence not cost-effective.

The inability to identify such a large proportion of relevant references using online databases clearly calls into question the validity of any results drawn from systematic review and meta-analysis of the papers identified by these methods. An extension of the Consolidated Standards of Reporting Trials (CONSORT) statement, to improve the reporting of clinical trials in relation to safety and harms-related data has been proposed by Ioannidis *et al*. (2004).

2.2.4 Different data sources for adverse drug reactions

Taking a broader view of ADRs in general, when making decisions related to clinical guidelines, data are available from a variety of different sources. These include not only clinical trials, but also observational studies and spontaneous anecdotal reports. Results from observational studies may be quantitatively combined with data from RCTs, while spontaneous reporting may provide baseline information that will influence a prior 'belief' regarding a meta-analysis of quantitative studies done in a Bayesian framework. In this way, information from three sources may be brought together to strengthen the evidence for any particular conclusion.

An interesting question posed by Loke *et al.* (2004) is whether those ADRs most frequently reported in the clinical trials are also those most often reported in anecdotal sources, choosing journal articles and reports to WHO for investigation. Whilst it is impossible to calculate rates for ADRs based on spontaneous reports due to lack of a denominator (no formal data on numbers exposed), it is possible to rank ADRs by frequency, and compare these frequencies to those reported in clinical trials.

The drug of interest was amiodarone, which is used to treat cardiac arrythmias, and is known to have several adverse reactions. The ADRs of amiodarone were divided into eight groups, and data on ADRs collected from three sources: meta-analysis of data from placebo-controlled RCTs, data from the WHO Collaborating Centre for International Drug Monitoring in Uppsala, Sweden, and published case reports in journals. There was a wide discrepancy in the rank orders of frequency of ADRs between the three sources. The most common type of ADR in the meta-analysis of ADRs was cardiac problems, whilst in the WHO monitoring the most frequent ADR was thyroid dysfunction and the journal reports ranked respiratory disorders most highly. The authors put forward several explanations for these discrepancies including the fact that clinical trials cannot identify very rare disorders, may be restricted to specific types of patient (such as middle-aged males), and may only monitor certain types of adverse events (for example, as all patients receiving amiodarone would have cardiac disease, it is unsurprising that they would be closely monitored for cardiac dysfunction,

which is then demonstrated as the most common type of ADR). Hence, RCTs were unlikely to provide data on rare or previously unrecognised ADRs.

Spontaneous reports however, can provide data on a broader range of patients, and hence can identify very rare ADRs as well as those that only occur after taking a drug for a long period of time. However, there is the possibility of under-reporting based on a clinician's decision on whether to report an ADR and whether it will be published in a journal, based on an editorial decision on the interest of a potential ADR. There is also lack of homogeneity in the WHO reports from different countries based on different criteria for acceptance of a report. The authors conclude that both RCT data and spontaneous reporting are of use in fully understanding a drug's safety profile. It may be possible to extend this idea to the quantitative combination of data from different sources, which can then lead to clinical decision-making guidelines for prescription.

2.2.5 Other interventions and unintended outcomes

Despite focusing on ADRs, it is also interesting to consider unintended effects of other types of interventions. It is usually quite plausible to consider that drugs, which have a wide range of potential actions on human physiology, would also have a propensity to produce unintended outcomes. It is, however, harder to conceptualise this type of effect in relation to other interventions such as surgical procedures or public health programmes.

It would therefore be of interest to include a variety of case-studies of different interventions, to compare the potential requirements for different approaches to a statistical analysis, depending on the nature of the intervention, and possible unintended consequences.

2.3 Discussion

This chapter has discussed the difficulties in primary collection of data regarding adverse events in general, and has also illustrated the magnitude of the problem of adverse events in clinical practice. These issues demonstrate the need for focused statistical techniques in the area of adverse events, and highlight the point that statistical analyses (at the level of primary data) are dependent on

the quality of the primary data that are incorporated into them. Much of the discussion in this chapter has related to adverse events in general, rather than specific adverse events related to a defined intervention, and has considered observational studies and spontaneous reporting systems. These study types are of value in evidence synthesis, but the majority of the statistical techniques to be developed will rely on primary data from formal trials.

At the level of evidence synthesis, it is very important to retrieve as many primary studies out of the pool of literature as possible. Whilst not directly the focus of this work, it is useful to highlight the problems that are encountered in searching for adverse events literature, even by specialist information scientists. The review of adverse events meta-analyses discussed in Chapter 6 is based on searches performed by Golder *et al*. (2006b).

A possible extension to work on adverse drugs reactions would be to include case studies based on non-pharmacological interventions; this area may present a different range of statistical challenges that would be of interest to investigate.

2.4 Summary

The concept of an ADR refers to an adverse drug reaction that is not due to errors of prescribing or administration, or to failure of efficacy or withdrawal symptoms. The extent of the problem of ADRs in the UK and elsewhere is discussed, as well as the difficulties in collecting primary data on ADRs. Another pertinent issue for evidence synthesis for adverse outcomes is the identification of primary studies, usually using electronic databases, and this issue has also been addressed. Non-pharmacological interventions that may have unintended adverse or beneficial effects are also considered.

3

Overview of meta-analysis methods

3.1 Introduction

It is important to place the particular difficulties related to meta-analysis and evidence synthesis regarding adverse events data within the overall framework of meta-analysis methods. Meta-analysis methods have now been developed to a high level of complexity and there are many different statistical techniques allowing the combination of data from different sources. Such meta-analysis methods are discussed in detail by Sutton *et al.* (2000) and Borenstein *et al.* (2009). These techniques vary in their approaches to how to combine data from different sources. An outline is presented of these methods, with discussion of how they differ both statistically and philosophically.

In this chapter, the methods will be described in the context of combining odds ratios (ORs), although certain methods may be used for the combination of other outcome metrics such as risk differences (RDs). These methods are used for the combination of binary data, as this data format is commonly derived from clinical trials in the context of interest (in terms of whether or not a participant experiences an adverse event).

Whilst this is by no means a comprehensive account of meta-analysis methods, it is aimed to give an overview of the most important methods and how they relate to each other.

Included are:

1. basic methods for data combination;
2. standard fixed effect models;
3. 'exact' stratified methods;
4. standard random effects models;
5. regression models; and
6. maximum likelihood methods.

This chapter includes only frequentist methods for meta-analysis; Bayesian methods are discussed in Chapter 4. Chapter 5 includes discussion of meta-analysis areas that may be specifically problematic for adverse events analyses, and less well-known or well-used models are discussed there. Other atypical or non-standard methods, such as methods for individual patient data (IPD) meta-analysis, mixed treatment comparisons (MTCs), and methods for harm–benefit modelling are discussed in the relevant chapters where these methods are put into practice in case-studies (Chapters 8, 9 and 11 respectively).

In this chapter, the observed data with regard to the parameter of interest (for example, an observed OR from an individual study i) is denoted by y_i, the true underlying value of the parameter of interest is denoted by θ_i and the estimate of the true underlying parameter of interest is denoted by $\hat{\theta}_i$. The true underlying value of the pooled effect (across all studies) is denoted as μ, with the estimate for this effect denoted as $\hat{\mu}$. Unless otherwise stated, these notations refer to an OR. Regarding variance parameters, the standard deviation (or standard error) for an individual study is denoted as s_i, with the between-studies standard deviation denoted as τ. The study-level weighting used in a meta-analysis is designated as W_i. This notation is tabulated in Table 3.1. The convention for notation of numbers derived from a 2×2 table for each study, with the numbers of cases and non-cases for treatment and control groups, is also set out.

Table 3.1: Notation used in this chapter.

Notation	Concept
i	Refers to a statistic or parameter at the level of an individual trial
k	Total number of primary studies in a meta-analysis
j	Number of study groups within an individual study (usually two, treatment and control)
y_i	Observed value of parameter of interest (e.g. odds ratio) for study i (study-level data). Note that this can be on a logarithmic or natural scale
θ	True underlying value of parameter (across multiple studies)
θ_i	True underlying value of parameter for study i
$\hat{\theta}_i$	Estimate for true underlying value of parameter for study i
μ	True underlying value of pooled estimate for parameter
$\hat{\mu}$	Estimate for true underlying value of pooled estimate for parameter
s_i	Standard deviation or standard error for parameter at study-level (study i)
τ	True underlying between-studies standard deviation
$\hat{\tau}$	Estimate of true underlying between-studies standard deviation
ε_i	Random error factor for each study, assumed to be distributed normally with mean 0 and variance ξ_i^2
ξ_i^2	Variance associated with random error factor for each study, ε_i
W_i	Study-level weighting for each study within a meta-analysis
a_i	Number of cases in the treatment group in study i
b_i	Number of non-cases in the treatment group in study i
c_i	Number of cases in the control group in study i
d_i	Number of non-cases in the control group in study i
n_i	Total number of participants in study i

3.2 Basic methods for data combination

The simplest method for combining data from multiple sources is to amalgamate the data into one dataset, and ignore the fact that the data are from different studies. Effectively this method takes the data from several studies, using the 2×2 table from each, and pools them into one such table (a *pooled analysis* or an *unstratified analysis*), thus losing the concept of multiple data sources.

A marginal analysis is straightforward to apply to an outcome of the OR or relative risk (RR), also known as a risk ratio.

For an OR, the formula for the point estimate is (Borenstein *et al.* 2009):

$$OR = \frac{ad}{bc}, \tag{3.1}$$

with a standard error (on the log scale) of:

$$\text{Standard error (log OR)} = \sqrt{\left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\right)}. \tag{3.2}$$

Using Equations 3.1 and 3.2, it is then possible to create a 95% confidence interval (CI) for log OR, by using the assumption that the log OR is normally distributed, and taking the standard normal deviate of the 0.975 point as 1.96. Hence, the 95% CI is given by:

$$\text{log OR} \mp 1.96(\text{Standard error(log OR)}), \tag{3.3}$$

which can be exponentiated to give a 95% CI on the natural scale.

The RR is calculated as (Borenstein *et al.* 2009):

$$\text{RR} = \frac{a/(a+b)}{c/(c+d)}, \tag{3.4}$$

with a standard error on the log scale of:

$$\text{Standard error (log RR)} = \sqrt{\left(\frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}\right)}. \tag{3.5}$$

The 95% CI for the log RR can be calculated by using the same method as for the OR (with the same assumption of normality on the log scale), given in Equation 3.3.

The RD is discussed at greater length in Section 3.6. Due to the fact that the RD is not distributed normally, on either the natural or log scale, it is not straightforward to calculate the CI on this scale, and therefore it is not considered further.

This method is also known as a *marginal analysis* due to the fact that only the overall totals (the marginal results) are used rather than data derived from each study individually. The resulting analysis produces only one overall estimate of the effect, rather than producing multiple estimates (one for each study) which are then combined into one pooled estimate.

A marginal analysis is not usually to be recommended for the reason that information is being lost due to not accounting for the fact that multiple studies are being combined. One major danger is that when combining studies of different cohort sizes where the magnitude and/or direction of the treatment effect are different, an erroneous result can occur due to the failure to take into account the individual study results. This is known as Simpson's Paradox (Simpson 1951).

In effect, the overall data are treated as if they are derived from one study, rather than from several different studies. This is in contrast to treating each study as a separate exercise in producing an estimate of the underlying treatment effect. These estimates can then be combined using different methods, making different assumptions about the ways in which the underlying treatment effects from each study may be related.

Meta-analysis methods were then developed that would allow combination of data to include and account for the multiple studies, whilst varying the amount of 'weight' given to each study based on characteristics of the study, whilst retaining the underlying assumption of a common treatment effect.

3.3 Fixed effect meta-analysis methods

3.3.1 Principles of fixed effect meta-analysis methods

The FE meta-analysis methods have the assumption that all studies are reflecting the same underlying treatment effect. This may be incorrect for a variety of reasons. There may be differences in the study population, in the treatments applied (e.g. drug doses and regimes, or surgical techniques). The treatment effect may ostensibly differ between studies for non-clinical reasons such as the methods of data collection, for example using different criteria to classify an 'event'.

In general terms, an FE model can be thought of as a means of producing an overall estimate of the treatment effect by using a weighted average of the individual study level estimates to produce an estimated treatment effect for all studies combined.

A generic FE model for data where the value of 'no difference' between the two treatment groups is 0, and the data are normally distributed, can be described as:

$$\hat{\theta}_i = \theta + \varepsilon_i. \tag{3.6}$$

In Equation 3.6, the term ε_i is a random error factor, and is distributed normally with a variance of ξ_i^2, such that:

$$\hat{\theta}_i \sim \text{Normal}(\theta, \xi_i^2), \tag{3.7}$$

as described by Whitehead (2002).

3.3.2 Inverse variance method

Using the OR as an example, the formula for the study-level OR (y_i) is given by:

$$y_i = \frac{ad}{bc}, \tag{3.8}$$

with a variance (on the log scale) of:

$$\text{Variance } (\log y_i) = \left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\right). \tag{3.9}$$

The ease of computation of the variance of the OR on the log scale means that the IV method is best suited to a meta-analysis performed on the log scale (in effect, combining the study level log ORs, rather than ORs). Furthermore, transformation to a logarithmic scale improves the assumption that the outcome metric (log OR) is distributed normally.

Inverse variance computes a weighted average of y_i values, using the inverse of the within-study variance (s_i^2) as the study weight (Borenstein *et al*. 2009). In effect,

$$W_{iIV} = \frac{1}{s_i^2}, \tag{3.10}$$

where s_i^2 is the equivalent of ξ_i^2 in Equation 3.7.

To derive a pooled estimate, the following formula is used (where k is the number of studies):

$$\hat{\mu}_{IV} = \frac{\sum_{i=1}^{k} W_{iIV} y_i}{\sum_{i=1}^{k} W_{iIV}}. \tag{3.11}$$

Thus, the studies with smaller variance (usually those with greater numbers of participants) are given greater weight by this method. Hence, the variability of the pooled treatment effect estimate $\hat{\mu}_{IV}$ is minimised.

An estimate for the variance of $\hat{\mu}_{IV}$ is given by:

$$\text{Variance}(\hat{\mu}_{IV}) = \frac{1}{\sum_{i=1}^{k} W_i} \tag{3.12}$$

From this value for the variance, 95% confidence intervals and p-values for the

hypothesis test that the true underlying value of the pooled estimate μ is 0 can be calculated.

The inverse variance method can be used in many situations as it can be applied to any outcome metric with an associated variance. Another strength of this method is that it can be used for studies where the original 2×2 tables are not available, but there are stated treatment effects and confidence intervals, allowing calculation of the variance for each study-level parameter (Egger *et al.* eds. 2001).

Using standard meta-analysis software, for the variance of individual study ORs the Woolf method is used, for individual study RDs the variance is estimated using the Normal approximation (Bradburn *et al.* 2007).

3.3.3 Mantel–Haenszel method

The Mantel–Haenszel (M–H) method (Mantel & Haenszel 1959) was originally developed for use in combining studies where the outcome is reported as an OR, and is a weighted average of the individual study estimates of the OR.

Denoting the individual study OR as y_i the estimate for each study is as above in Section 3.3.2.

The weighting for each study, W_{iMH}, is calculated as follows:

$$W_{iMH} = \frac{b_i c_i}{n_i}. \tag{3.13}$$

Thus, the M–H pooled estimate $\hat{\mu}_{MH}$, is given by (Borenstein *et al.* 2009):

$$\hat{\mu}_{MH} = \frac{\sum_{i=1}^{k} W_{iMH} y_i}{\sum_{i=1}^{k} W_{iMH}}. \tag{3.14}$$

The Mantel–Haenszel estimate refers to the OR rather than the log OR (a major difference between the M–H and IV models) and hence is not symmetrically distributed, and cannot therefore be assumed to come from a normal distribution. Calculation of the variance (with the aim of calculating confidence intervals) for the M–H pooled estimate must be performed on the log scale (as for the IV model). The formula for the variance is complex, and involves calculating

multiple values based on the numbers in each individual study, summing these across all studies and then combining them across all studies. The full formulae for calculating the variance are set out in Borenstein *et al*. (2009). A continuity correction (this term is discussed more fully in Section 5.2.3) to calculate this variance is required only if there are zero events in the corresponding arm of *all* studies in the meta-analysis.

For the M–H method therefore, only studies with zero events in the control group will have a weighting of 0 (i.e. $c_i = 0$). As these studies will have an OR which is 'undefined', they will effectively be excluded from the meta-analysis (not contributing to the numerator of the pooled estimate nor to the sum of the weights). Studies with an OR of 0 (i.e. $a_i = 0$) will contribute to the pooled estimate, as they will contribute to the sum of the weights. Hence, a continuity correction for inclusion of individual studies with zero events in one of the two arms is not required. (However, meta-analytic software often automatically includes a continuity correction for all studies with only one arm with zero events.) If all studies in the meta-analysis have zero events in the control arm, then the sum of the weights will be 0, and hence it is not possible to calculate a pooled estimate without a continuity correction. Similarly, calculation of the variance (and hence CI) for the M–H estimate requires a continuity correction if there are zero events across all studies, or across all corresponding arms for either the control or treatment groups.

This contrasts with the IV method, whereby a study where either arm has zero events yields a variance which is undefined, and hence a weighting of 0. This means that a continuity correction is required for all studies with zero events in one arm to be included.

The M–H method has since been extended for use in cases where the outcome is not an odds ratio (Mantel 1963; Egger *et al*. eds. 2001), for example, a risk ratio or risk difference.

3.3.4 Derived statistics from fixed effect models

For an FE model a test of the null hypothesis that the treatment difference in all studies is equal to 0, the U statistic is compared to a chi-squared distribution with one degree of freedom (Whitehead 2002).

The U statistic is given by:

$$U = \frac{(\sum_{i=1}^{k} \hat{\theta}_i W_{iIV})^2}{\sum_{i=1}^{k} W_{iIV}}. \tag{3.15}$$

As the weights in this formula are as for the IV model (see Equation 3.10), this statistic applies to the IV FE model.

In order to test for statistical heterogeneity among the studies, the Q statistic is used, being compared to a chi-squared distribution with $k-1$ degrees of freedom (Cochran 1954; Whitehead 2002). The Q statistic is given by:

$$Q = \sum_{i=1}^{k} W_{iIV}(\hat{\theta}_i - \hat{\mu})^2. \tag{3.16}$$

The Q statistic is effectively a weighted sum of squares of the differences between treatment effects from the individual studies and the overall estimated treatment effect. If the treatment difference parameters are homogeneous (little or no heterogeneity), the Q statistic follows a chi-squared distribution with $k-1$ degrees of freedom.

Both the U and Q statistic can be used when estimates of summary statistics are calculated by other methods, such as the Mantel–Haenszel method.

Statistical heterogeneity is discussed more fully in Section 3.9.

3.3.5 Peto method

The Peto meta-analysis method (also referred to as the one-step method) is a modification of the Mantel–Haenszel method, and is used when the outcome measure is an OR (Borenstein *et al.* 2009).

The Peto method for combining ORs is highly relevant to the analysis of trials for adverse drug reactions (ADRs), as the baseline method for calculating the ORs can include studies with zero events in one arm of the trial without recourse to a continuity correction.

The Peto OR for an individual study is estimated as:

$$\exp(\hat{\theta}_{iPeto}) = \frac{O_i - E_i}{I_i}, \tag{3.17}$$

where $\hat{\theta}_{iPeto}$ is the log Peto OR, O_i is the observed number of events in the treatment group, and E_i is the expected number of events in the treatment group, calculated by:

$$E_i = \frac{(a_i + b_i) \times (a_i + c_i)}{n_i}. \tag{3.18}$$

The value for I_i in Equation 3.17 is given by:

$$I_i = \frac{(a_i + b_i) \times (c_i + d_i) \times (a_i + c_i) \times (b_i + d_i)}{n_i^2 \times (n_i - 1)}. \tag{3.19}$$

The value I_i is also known as the hypergeometric variance of the event count in the treatment group (Egger *et al.* eds. 2001).

The variance of the log Peto OR, θ_{iPeto}, for an individual study is:

$$s^2_{iPeto} = \frac{1}{I_i}. \tag{3.20}$$

The weighting for each study for combination of Peto log ORs in a meta-analysis is:

$$W_{iPeto} = \frac{1}{s^2_{iPeto}}, \tag{3.21}$$

so that in effect the weighting for each study is simply I_i. If there are no events at all in the study, then the value of I_i is 0, hence the study receives a weighting of 0.

Therefore, the combined estimate for a Peto log OR, $\hat{\mu}_{Peto}$, is:

$$\hat{\mu}_{Peto} = \frac{\sum_{i=1}^{k} W_{iPeto}\hat{\theta}_{iPeto}}{\sum_{i=1}^{k} W_{iPeto}}. \tag{3.22}$$

This can be expressed alternatively as:

$$\hat{\mu}_{Peto} = \frac{\sum_{i=1}^{k}(O_i - E_i)}{\sum_{i=1}^{k} I_i}, \tag{3.23}$$

The variance of $\hat{\mu}_{Peto}$ is given by:

$$\text{Variance}(\hat{\mu}_{Peto}) = \frac{1}{\sum_{i=1}^{k} I_i}, \tag{3.24}$$

or alternatively:

$$\text{Variance}(\hat{\mu}_{Peto}) = \frac{1}{\sum_{i=1}^{k} W_{iPeto}}. \tag{3.25}$$

The heterogeneity statistic, Q, is given by (Egger *et al*. eds. 2001):

$$Q = \sum I_i(\hat{\theta}_{iPeto} - \hat{\mu}_{Peto})^2. \tag{3.26}$$

The ability of the Peto method to lend itself to inclusion of studies with zero events in one treatment arm without recourse to a continuity correction appears to be an advantage where sparse data is an issue. However, the Peto model can lead to biased estimates if the study design is unbalanced, with different numbers of participants in the study groups (Greenland & Salvan 1990; Fleiss 1993). It has also been recommended to avoid the Peto method for non-experimental design studies (Greenland & Salvan 1990).

The Peto method can also be used as a means of meta-analysis for time-to-event (survival) data (Egger *et al*. eds. 2001). Time-to-event data can be combined either by calculating the individual hazard ratios (HRs) and combining using the IV method (Section 3.3.2), or by using the Peto method for ORs to calculate an estimate for the HR for each study. The overall HR, $\mu_{H\hat{R}Peto}$ is then calculated as a weighted average of the log HR values.

Dividing each of the i studies into j time periods we have as an estimate of the individual trial HRs ($y_{iHRPeto}$):

$$\hat{\theta}_{iHRPeto} = \exp\left(\frac{\sum_j O_{ij} - \sum_j E_{ij}}{\sum_j v_{ij}}\right), \tag{3.27}$$

where the v_{ij} value for each study is the study variance and each value O, E and v are summed across all time periods j for all studies i. The variance of $\hat{\theta}_{iHRPeto}$ is (on the log scale):

$$\text{Variance } (\log(\hat{\theta}_{iHRPeto})) = \frac{1}{\sum_j v_{ij}}, \tag{3.28}$$

and hence the weighting for each trial, i, is given by:

$$W_i = \sum v_{ij}. \tag{3.29}$$

The studies can then be combined as follows:

$$\hat{\mu}_{HRPeto} = \exp\left(\frac{\sum_i (W_i \log(\hat{\theta}_{iHRPeto})}{\sum_i W_i}\right). \tag{3.30}$$

The weights, Wi, are equal to the sum of the variances from the individual trials divided into their time periods, $\sum_j v_{ij}$.

This method is based on the fact that the log-rank statistic is calculated by a similar method to that of the Peto method for calculating an OR. Calculation of the O, E and v values for each study requires the use of individual patient data (IPD).

3.4 'Exact' stratified methods

The above methods of combining data and deriving an associated interval are asymptotic in that it is assumed that either the number of individuals in the study is large or that the number of strata, or studies, is large (Emerson 1994). Exact methods for the derivation of a combined odds ratio based on multiple strata have been developed, based on exact distribution theory (Emerson 1994). These methods are not dependent on the number of studies or participants.

Exact methods rely on deriving the true distribution of the test statistic. This requires permuting the observed data in all possible ways and comparing the observed data to what might have been observed, to determine an exact *p*-value.

One example is the method described by Mehta *et al*. (1985). The OR is estimated by evaluating all possible permutations of a conditional hypergeometric response and is an extension of Fisher's exact test which allows for separate success probabilities in each stratum. The StatXact® software produced by Cytel has two main procedures for tackling stratified data from multiple 2×2 tables. One procedure is a homogeneity test to determine whether all odds ratios across all strata are the same (Zelen 1971), the second is to provide an exact confidence interval for the combined odds ratio and to test whether this odds ratio is equal to one (Gart 1970).

Exact methods are used in Chapter 7, as one of the multiple methods used to analyse an adverse events dataset with sparse data (zero events in several treatment arms and studies).

3.5 Random effects models

3.5.1 DerSimonian & Laird method

Contrasting with an FE model, where the default assumption is that all studies are estimating the same true underlying treatment effect, a random effects (RE) model makes the assumption that the studies are not all estimating the same true underlying treatment effect. Rather, the default assumption is that all studies are estimating a different true underlying treatment effect, but these underlying treatment effects are connected in that all come from the same distribution. Alternatively, the true underlying treatment effects are said to be *exchangeable*, in that none of the studies 'stands out' prominently from the rest. Differences in underlying treatment effect may be due to clinical differences between studies, for example differences in the participant populations, or in the exposures (for example, different dosage regimes).

For an individual study in an RE meta-analysis:

$$y_i = \mu + \zeta_i + \varepsilon_i, \tag{3.31}$$

where μ is the true underlying pooled mean treatment effect across all studies, ζ_i is the difference between the true underlying treatment effect for study i, (θ_i), and ε_i is the random (sampling) error within study i, and represents the difference between y_i and θ_i (Borenstein *et al*. 2009).

Continuing to the level of the meta-analysis, some estimate of ζ_i must be made. It is usually assumed that the ζ_i values are distributed normally with a mean of 0, and variance τ^2. This parameter, τ^2, is known as the *between-studies variance*. In effect, τ^2 is the variance of the true difference in effect size between the individual study treatment effect (θ_i) and the underlying overall treatment effect (μ), which is ζ_i. Further, it is assumed that the ζ_i and ε_i values are distributed independently (Whitehead 2002). Alternatively, it can be envisaged that the θ_i values are distributed normally with a mean μ and variance τ^2.

Hence, for an RE meta-analysis, it is necessary to estimate the within-study variance (as for an FE meta-analysis) and the between-studies variance τ^2. The most common method used to accomplish this is to use the method described by DerSimonian & Laird (1986), which is also known as the method of moments (Borenstein *et al*. 2009), based on the method used to calculate τ^2. The method of moments is considered the simplest, and other methods are available including a restricted maximum likelihood (REML) method. Hence, the overall weighting for each study in an RE meta-analysis requires a combination of both within-study and between-studies variation.

To estimate the true underlying value of τ^2, the following formula is used (Borenstein *et al*. 2009):

$$\hat{\tau} = \frac{Q - (k-1)}{C}, \tag{3.32}$$

where

$$Q = \sum_{i=1}^{k} W_i y_i^2 - \frac{\left(\sum_{i=1}^{k} W_i y_i\right)^2}{\sum_{i=1}^{k} W_i} \tag{3.33}$$

and

$$C = \sum_{i=1}^{k} W_i - \frac{\sum_{i=1}^{k} W_i^2}{\sum_{i=1}^{k} W_i}. \tag{3.34}$$

This method is based on the method of moments. The above equations use the weighting W_i for each study, as calculated using an appropriate FE method, either the IV or M–H method (Egger *et al.* eds. 2001).

The next step is to calculate the pooled estimate for the overall parameter of interest ($\hat{\mu}_{DL}$).To accomplish this goal, each study needs to be weighted according to both its individual weighting W_i and τ^2.

Denoting the weighting by the DerSimonian & Laird method as W_{iDL}, and the appropriate FE model weighting for each study as W_{iFE},

$$W_{iDL} = \frac{1}{W_{iFE} + \hat{\tau^2}}. \tag{3.35}$$

The final step in this method is to calculate the overall pooled estimate, $\hat{\mu}_{DL}$:

$$\hat{\mu}_{DL} = \frac{\left(\sum_{i=1}^{k} W_{iDL} y_i\right)^2}{\sum_{i=1}^{k} W_{iDL}}. \tag{3.36}$$

This pooled estimate $\hat{\mu}_{DL}$ has a variance of

$$\text{Variance}(\hat{\mu}_{DL}) = \frac{1}{\sum_{i=1}^{k} W_{iDL}}. \tag{3.37}$$

A major difference between the fixed effect and random effects models is that the random effects model may produce a wider confidence interval for the pooled estimate. This is because the two sources of uncertainty reduce the relative weightings allocated to the larger and more precise studies, thus evening out the weightings and allowing the smaller and less precise studies to contribute more strongly (with relatively greater weights compared to the larger studies) to the pooled effect. Hence, this model is more conservative in the conclusions drawn.

When τ^2=0, which occurs when the Q statistic is equal to or smaller than its degrees of freedom (k-1), then the weights are the same as those used in the FE

methods and hence the results of the meta-analysis will be the same. In meta-analyses where there is little heterogeneity, the DerSimonian & Laird method has little to contribute beyond the FE methods.

In meta-analyses where there are few studies, and the effect size varies widely between studies, then the between-studies variation will be high, thus a random effects meta-analysis will be expected to have a wider confidence interval, and hence lower power, as compared with a fixed effect model (Borenstein *et al*. 2009).

Therefore, both clinical (in considering whether there are clinical reasons for heterogeneity) and statistical issues can be important in determining which model is preferred, although tests for statistical heterogeneity can help determine whether there would be any major differences in the conclusions between the two models.

The use of continuity corrections with the DerSimonian & Laird model follows the requirements for the IV and M–H models. A continuity correction is required for all studies with zero events in one arm for the IV model, whilst for the M–H model, a continuity correction is required only if the dataset has zero events in all control arms combined.

A further issue is that whilst an FE meta-analysis is making the default assumption that the true underlying treatment effect is the same, the assumption of the RE method is that there exists an underlying distribution of such treatment effects. This distribution can be described using a *prediction interval* (PI), which provides an interval in which it is expected that the mean treatment effect of a new study (selected at random within the population of primary studies) will fall (Borenstein *et al*. 2009). This PI is calculated in a similar manner to a CI (which considers the accuracy of the estimated mean treatment effect by relating it to its own variance). However, the PI includes two sources of variation, whereas the CI includes only one. The PI takes into account both the variance of the pooled estimate, but also the between-studies variance τ^2 (the actual formula requiring its estimate $\hat{\tau}^2$).

For the prediction interval the formula that can be used is (Higgins *et al*. 2009):

$$PI = \hat{\mu} \pm t^{\alpha}_{df} \sqrt{\hat{\tau}^2 + \text{Variance}(\hat{\mu})}. \tag{3.38}$$

Equation 3.38 allows for the fact that true values of μ and τ are unknown, and hence are estimated. In Equation 3.38, the t distribution is used rather than the normal distribution to derive the centile points for the desired interval based on α. The degrees of freedom (df) is $k-2$. This formula takes into account the variance of the true effects (τ^2) and the variance of the mean effect, Variance($\hat{\mu}$).

Hence, the PI is usually wider than the associated CI for an RE model (assuming $\hat{\tau}^2$ does not equal 0), and is a measure of the dispersion of the actual true underlying effect sizes.

3.6 Methods involving risk difference

The RD is a pertinent outcome metric when considering studies that may have zero event counts in one or both arms. If there are zero events in one arm only of a study, then no continuity correction is required to calculate either the RD or the individual study variance (and hence its weighting in a meta-analysis).

The RD is calculated as:

$$\hat{\theta}_{iRD} = \frac{a_i}{a_i + b_i} - \frac{c_i}{c_i + d_i}, \tag{3.39}$$

with an individual study variance s^2_{iRD} estimated as:

$$\hat{s}^2_{iRD} = \frac{a_i b_i}{(a_i + b_i)^3} + \frac{c_i d_i}{(c_i + d_i)^3}. \tag{3.40}$$

It can be seen that the RD is always 0 if there are zero events in one or both arms. If there are zero events in one arm only, the variance of the RD is still calculable without a continuity correction, and will be greater than 0. Only if there are zero events in both arms will the variance equal 0. Hence, if using the IV method to calculate the study weightings, then the weight for such a study will be undefined, and will require a continuity correction to be included in the study. Adding a uniform continuity correction will effectively weight each study according to the number of participants and the distribution of participants across the two arms.

Using the M–H method, each study is weighted according to the following formula (Egger *et al*. eds. 2001, citing earlier work):

$$W_{iMHRD} = \frac{(a_i + b_i)(c_i + d_i)}{n_i}. \tag{3.41}$$

Hence, the weighting of each study is greater than 0, and is based on the size of the study and the distribution of participants across the two arms. The lack of requirement for a continuity correction using the M–H method compared to the IV method may indicate that the M–H method is the preferred FE method when using RDs in meta-analyses where there are studies with zero events. The variance of the M–H estimator of the pooled RD ($\hat{\mu}_{MHRD}$) is calculated using a method that only produces a value of 0 when there are zero events in all studies (Egger *et al*. eds. 2001).

The RD is related to the Number Needed to Treat (NNT), in that the NNT is the reciprocal of the RD (Egger *et al*. eds. 2001, citing previous authors). This gives the number of patients required to produce one positive result of the treatment. Within the context of adverse events, a more appropriate outcome measure is the Number Needed to Harm, calculated in the same way, but indicating the number of patients required to produce one adverse event as a result of the intervention (Egger *et al*. eds. 2001, citing previous authors). The NNT has been demonstrated to be unsuitable for meta-analysis calculations, and it is preferable to use a probability [risk] difference to perform a meta-analysis, and then to calculate the NNT from the resulting pooled estimate (Whitehead 2002, citing previous authors).

3.7 Regression and meta-regression methods

Maximum likelihood logistic regression methods are a suitable FE method for aggregate data, where the 2×2 table is known for each study. Also, explanatory covariates can be added to the model if required. These models will give estimates similar to the M–H and IV models if sample sizes are large (Egger *et al*. eds. 2001). IPD could also be used for these models, whereby each patient effectively supplied the result of a single Bernoulli trial, with study as an

explanatory covariate in the model. Logistic regression methods are covered in detail in McCullagh & Nelder (1989).

The concept of *partial exchangeability* refers to a scenario whereby some of the heterogeneity between studies is due to random effects, whilst some is explained by systematic differences between studies (Higgins *et al*. 2009). Meta-regression can be used to explain these differences in the form of covariates in a meta-regression model.

It is more intuitive to use meta-regression in an RE model, but to develop the concepts, the FE model is dealt with first. Equation 3.6 is further differentiated as follows:

$$\hat{\theta}_i = \delta + \eta_i + \varepsilon_i, \tag{3.42}$$

where η_i refers to an explanatory variable x and its covariate, β (Whitehead 2002), where x is a continuous variable. Furthermore, η_i can refer to a binary explanatory variable, or could be expanded into multiple variables x_i with associated covariates β_i to account for a discrete variable with multiple levels. Also discussed by Whitehead (2002) are methods for estimation of the β_i values. If there are no explanatory variables in the model then δ is equal to θ in Equation 3.6.

To extend this model to incorporate random effects,

$$\hat{\theta}_i = \delta + \eta_i + \zeta_i + \varepsilon_i, \tag{3.43}$$

where ζ_i represents the ith trial's deviation from the mean of all trials with the same values for the explanatory covariates, as specified within the expression η_i (Whitehead 2002).

Logistic regression is one of the multiple methods used in Chapter 7, but meta-regression models are not pursued further.

3.8 Maximum likelihood methods

Maximum likelihood methods can also be used to combine odds ratios (Emerson 1994). For large sample sizes this method is the most efficient (Sutton *et al*. 2000).

A description of both unconditional and conditional maximum likelihood estimators is given by Hauck (1984). In this scenario, there are k studies, with two groups in each study denoted j. There are two possible asymptotic cases, the 'fixed-strata' case, in which the number of studies is fixed but the number of participants within each study increases in size, and the 'increasing-strata' case whereby the numbers in each study are fixed but the number of studies increases. It is assumed that the number of events in each study group (a and c in the 2×2 table) are independent binomial variables with parameters P_{jk} and N_{jk}, whereby N_{jk} is the number in each study and group, and P_{jk} is derived empirically as total number of events for each group divided by the total in the group, for each study.

The likelihood is given by:

$$L = \prod_{j=1}^{2} \prod_{k=1}^{k} P_{jk}^{X_{jk}} Q_{jk}^{N_{jk}-X_{jk}}, \tag{3.44}$$

where X_{jk} refers to the number of successes in the jth group of the kth study and $Q_{jk} = 1 - P_{jk}$.

This formula can then be used to derive the maximum likelihood estimate for the OR across all studies, $\hat{\theta}_{UML}$.

It has been shown that unconditional maximum likelihood estimation is not consistent for estimating the OR when the number of studies increases and the marginal counts remain fixed (Breslow 1981; Hauck 1984).

In this scenario, conditional maximum likelihood can be used, as described by Hauck (1984). This method provides an estimate of the pooled OR, $\hat{\theta}_{CML}$, that is consistent and asymptotically normal.

Maximum likelihood methods are not used further, as they are difficult to apply computationally, and may have problems when used in conjunction with issues commonly associated with adverse events meta-analyses, such as sparse data.

3.9 Heterogeneity within meta-analysis

This discussion refers to statistical heterogeneity, defined as variability in observed treatment effects across different primary studies within a meta-analysis that is greater than would be expected due to random error alone (Deeks *et al*. 2008). Other forms of heterogeneity such as clinical heterogeneity (differences in participants across primary studies) or methodological heterogeneity (differences in study design in primary studies) are not considered further, other than that these factors may be considered in the investigation of heterogeneity.

The test for heterogeneity based on the Q statistic has been discussed in Section 3.3.4. An alternative statistic relating to heterogeneity has been proposed (Higgins & Thompson 2002; Higgins *et al*. 2003). The Q statistic has low power to detect true heterogeneity, especially when the number of primary studies in the meta-analysis is low (but may have excessive power when there is a greater number of larger studies). It is also salient to note that in a meta-analysis, clinical and methodological heterogeneity may often lead to statistical heterogeneity. In the light of these issues, the authors developed an approach to heterogeneity that would provide a measure of the degree of inconsistency across the studies' results. They designated this measure I^2, calculated as follows:

$$I^2 = 100\% \times \frac{Q - df}{Q}, \tag{3.45}$$

where Q is the Cochran heterogeneity statistic (as described in Section 3.3.4). If I^2 is negative it is set to 0%, to give a value between 0% and 100%. An I^2 value of 0% indicates no heterogeneity, with high heterogeneity being at above 75%. The I^2 value can be interpreted as the percentage of total variation across the studies due to heterogeneity (rather than random chance), and is therefore a measure to quantify heterogeneity. It has the advantage that it does not depend on the number of studies in the meta-analysis, nor on the outcome measure.

The I^2 value can also be used to investigate heterogeneity among subgroups of studies.

It should be noted that when the Q statistic fails, for example when the variance of an individual study is incalculable due to zero events in one or both arms, then the I^2 statistic will also be incalculable across the full dataset of studies, unless a continuity correction is used. It is therefore difficult to assess the value of the I^2 statistic in such circumstances, as the use of a continuity correction may lead to bias.

Causes of statistical heterogeneity can be investigated using meta-regression, as discussed in Section 3.7, or by using subgroup analysis (either subgroups of studies or subgroups of individual patients). Subgroup analysis and individual patient data analysis are discussed in Sections 5.5 and 5.6.

3.10 Discussion and conclusions

Marginal analyses can give a basic impression of the overall numbers of a particular outcome and an immediate comparison between treatment and control groups.

An FE meta-analysis disregards any potential heterogeneity in treatment effect and estimates one overall study level treatment effect. For this reason it is very important to investigate any potential heterogeneity, to determine its existence and magnitude, and to discover why it may be occurring. Subgroup analysis, sensitivity analysis and meta-regression are all useful ways to approach this.

An RE meta-analysis is an option where there is heterogeneity that cannot be identified, but this method will result in a more conservative estimate of the pooled effect, with greater uncertainty. An alternative approach would be to place a random effects term in a regression model to allow for heterogeneity.

3.11 Summary

This chapter sets out several methods for combination of data, using frequentist methods. These range from the most basic methods, of simple marginal analysis to more sophisticated weighted averages (or meta-analyses), which take into

account differences across trials in their weighting systems. The most straightforward method is the fixed effect method, of which there are multiple methods of calculation, and which assumes that the same underlying treatment effect is being estimated in all studies. Slightly more complex are the random effects methods, which take into account exchangeability of treatment effect across studies. Meta-regression methods assist with investigation of heterogeneity between studies, which can occur when partial exchangeability is assumed across studies. Maximum likelihood, regression methods and exact methods are also discussed.

4

Bayesian meta-analysis

4.1 Introduction

The methods discussed in Chapter 3 have all been frequentist in nature, in that they regard the underlying 'true' value of a treatment effect to be an unknown but fixed quantity (or coming from a specific distribution with fixed but unknown parameters in the case of the random effects model).

The traditional frequentist approach to probability is based on the concept of the probability of an event occurring over a 'long-run' of repeated events. With this paradigm, hypothesis testing is based on the concept of collecting data, and then using the data to determine the probability of observing such data, on the assumption that the (unknown) parameter of an assumed distributional model takes a certain value. The aim of the frequentist analysis is therefore to determine the degree to which observed data *support* selected potential values for the underlying 'true' value.

By contrast, the Bayesian approach uses the observed data to make probability statements regarding the unknown parameters of the model being assumed. The underlying parameters are themselves regarded as unknown random quantities, and hence can be modelled distributionally.

This chapter discusses both the algebra underpinning Bayesian principles and some of the practical difficulties that, until recently, have prevented widespread use of Bayesian approaches to data analysis. Bayesian techniques are applied

in Chapters 7, 9, 10 and 11, and the methodologies used are discussed in more detail in each chapter.

4.2 Bayesian principles

The Bayesian approach to statistical inference is underpinned by Bayes' Theorem:

$$p(b|a) = \frac{p(a|b)}{p(a)} \times p(b), \tag{4.1}$$

where a and b are two events (that may or may not be independent) and $a|b$ refers to the probability of event a occurring conditional on the occurrence of event b (Spiegelhalter *et al*. 2004). In Equation 4.1, $p(b)$ is the prior probability for event b.

Whilst Bayes' Theorem may appear straightforward in concept, in actuality Bayesian methods are difficult to apply. In order to model the distribution of interest, that of the parameter of interest given the observed data, it is necessary to make distributional assumptions regarding the parameter of interest (known as the prior distribution). Once an algebraic model has been developed, making any necessary assumptions regarding unknown parameters, the prior distributions can be combined (according to Equation 4.1) with the observed data to create a posterior distribution for all the unknown parameters.

In a more complex scenario, there may be multiple parameters within a model, which are not of direct interest; these are known as *nuisance parameters*. These parameters must be accounted for in some way, in order to derive the *marginal* distribution for the parameter of interest, θ, in the light of the data y, or $p(\theta|y)$.

The *likelihood distribution* is the probability of producing the observed data conditional on certain values of the unknown modelling parameters, or $p(y|\theta, \psi)$ where θ is the true underlying value of the parameter of interest, ψ is a set of nuisance parameters, and y is the observed data). The true posterior distribution for θ is proportional to the likelihood distribution multiplied by the prior distributions on the unknown parameters; the prior distributions are known as they are selected by the user. Alternatively (Spiegelhalter *et al*. 2004),

$$p(\theta, \psi|y) \propto p(y|\theta, \psi)p(\theta, \psi). \tag{4.2}$$

This can also be expressed as:

$$p(\theta, \psi|y) = \frac{p(y|\theta, \psi)p(\theta, \psi)}{\int_\Theta \int_\Psi p(y|\theta, \psi)p(\theta, \psi)d\theta d\psi}. \tag{4.3}$$

The ultimate aim is to derive a *marginal* distribution for the parameter of interest, θ, in the light of the data y, where all nuisance parameters, ψ, have been *integrated out* or *averaged over*, to produce:

$$p(\theta|y) = \int_\Psi p(\theta, \psi|y)d\psi. \tag{4.4}$$

Using these methods, the Bayesian analysis is interpreted in terms of the 'degree of belief' regarding the true values of the parameters of interest. One important advantage of Bayesian methods (for all forms of analysis including meta-analysis) is the ability to incorporate within the model all forms of uncertainty, across all parameters and hyperparameters to which a prior distribution is applied. One disadvantage of the Bayesian paradigm is its lack of an easily interpretable test for statistical significance, such as a *p*-value. This issue is mentioned with regard to meta-analysis by Sutton & Abrams (2001), although the concept applies equally to other types of Bayesian analysis. To offset this drawback in the Bayesian approach, it is possible to make probability statements regarding the true underlying value of a parameter.

A major difficulty in implementing Bayesian analyses, which has until recently militated against their common usage, is that complex integrals may be required to produce the posterior distribution of the variable of interest, conditional on a continuous distribution of a nuisance variable. Referring back to Equation 4.1, the prior probability for event a, $p(a)$, is an integrating constant when generating the posterior distributions. In Equations 4.3 and 4.4, the integrals are more complex and unlikely to be easily solved algebraically.

To address this problem, computational methods have been recently developed to allow Bayesian modelling to be performed in a relatively straightforward manner. These methods are discussed below in Section 4.3.

One of the most significant differences between frequentist and Bayesian inference is in the way they conceptualise the unknown parameters of the model (both parameters of interest and nuisance parameters). In the frequentist approach, such parameters are considered to have a true underlying value that is fixed, but unknown. Alternatively, this true underlying value, whilst still taking a specific value, comes from some unknown distribution. For example, in a random effects (RE) model, the true underlying treatment effects for multiple studies investigating ostensibly the same quantity are connected by this distribution and hence considered to be 'exchangeable'.

By contrast, for the Bayesian approach, the true underlying treatment effect has not a specific value, but its own distribution of values, which the model is aiming to estimate. In order to achieve this, prior distributions must be placed on the distributions for the parameters of the model. These prior distributions can be based on external evidence (for example, from previous studies), or on beliefs, for example derived from 'expert opinion'. The ability to incorporate evidence external to the specific dataset being analysed at that time is one of the integral differences between frequentist and Bayesian methods.

4.3 Practical Bayesian analysis

4.3.1 Markov Chain Monte Carlo methods

Practical Bayesian analysis is usually accomplished by means of multiple sampling methods, based on Markov Chain Monte Carlo (MCMC) methods. Monte Carlo methods are used to evaluate integrals by means of simulation as opposed to using algebraic analysis (Spiegelhalter *et al*. 2004). Hence, these methods are useful for intractable or complex integrals involving several dimensions.

The concept of a Markov Chain refers to a sequence of random variables, for which the current value is dependent only on the immediately preceding value (regardless of all previous values); this property is known as the Markov property (Gelman *et al*. 2004). Markov chain simulation draws repeated values for a parameter of interest, θ, from approximate distributions, and on repeated sampling, the approximate distributions more closely approach the target distribution, $p(\theta|y)$.

By Markov Chain theory, following sufficient sampling, the samples will eventually come from an 'equilibrium distribution' which should accurately reflect the true posterior distribution for θ given the data y; (Spiegelhalter *et al*. 2004).

The MCMC approach involves sampling from a joint posterior distribution, $p(\theta, \psi|y)$, and with repeated sampling, large numbers of values for θ and ψ can be derived (Spiegelhalter *et al*. 2004). If there are t samples taken in total, these values can be denoted as (θ^1, ψ^1), (θ^2, ψ^2), ..., (θ^t, ψ^t), where t refers to the sampling iteration. Inferences about the true underlying value of θ can be derived from these sampled values, θ_i, where i refers to ith iteration. For example, the mean of these sampled values can be used as an estimate of the posterior mean, $E(\theta|y)$. A histogram of the sampled values can be used to represent the posterior distribution, $p(\theta|y)$.

By using this method, the sampled values for θ have been taken across a range of plausible values for the nuisance parameter, ψ, and hence the effect of this parameter has been eliminated.

There are various methods to achieve this, including Gibbs sampling (discussed below in Section 4.3.2), which is the method of choice for WinBUGS, the most prominent software for Bayesian statistical modelling.

4.3.2 Gibbs Sampling

Gibbs sampling (also known as *alternating conditional sampling*) is described in detail by Gelman *et al*. (2004). Let there be a vector of parameters, θ, divided into d subvectors, $\theta = (\theta_1, \theta_2, ..., \theta_d)$. The process of Gibbs Sampling involves multiple iterations, whereby at each iteration a sample value is derived for each of the d subvectors (i.e. for each iteration there are d steps). At each iteration, t, the d subvectors are placed in a specific order. For each of the individual subvectors, θ_j, a value for that iteration, θ_j^t is sampled. This sampling is conditional on the current values (from the current iteration if that subvector has been updated already, or the previous iteration if not) of all of the $d-1$ vectors of θ, excluding θ_j, and on the data, y. Alternatively,

$$p(\theta_j|\theta_{-j}^{t-1}, y). \tag{4.5}$$

In Equation 4.5, θ_{-j}^{t-1} refers to all the components of θ, except for θ_j, at their current values:

$$\theta_{-j}^{t-1} = (\theta_1^t, ..., \theta_{j-1}^t, \theta_{j+1}^{t-1}, ..., \theta_d^{t-1}). \quad (4.6)$$

The Gibbs sampling method is the simplest of the Markov chain simulation algorithms; more complex algorithms are described by Gelman *et al*. (2004). In this thesis, Gibbs sampling is the method of choice for all Bayesian analyses, due to its easy availability within the WinBUGS statistical package.

4.3.3 Other practical issues of Bayesian analysis

Other practicalities of performing Bayesian analysis include the selection of initial values, from which the sampling algorithm can be commenced. The choice of initial values may be important in facilitating convergence of the model, defined as the point at which the sampling algorithm has achieved the equilibrium distribution. Convergence can be verified by examining the history of the sampling algorithm, and if multiple chains are being run simultaneously, these can be checked for convergence using the methods of Brooks & Gelman (1998). To avoid incorporating samples from an unconverged posterior distribution, a certain number of initial samples can be discarded from the final distributions – this is known as the 'burn-in'.

Another issue is autocorrelation, whereby the sampling algorithm is not independent of previous values (other than the value directly preceding the current value). Ideally, the Markov chains should be 'mixing', or covering the full space of the joint posterior distribution from which they are sampling. If the chain(s) are mixing appropriately, then the autocorrelation from successive iterations will reduce as the 'lag' between iterations increases. If this does not occur, then less information is provided regarding the posterior distribution for each iterate, and therefore a larger sample is required for adequate coverage of the sample space (Congdon 2006). An alternative approach to autocorrelation is to thin the chain(s), by taking every nth sample and discarding the rest; this procedure is known as *thinning* the chain (Gelman *et al*. 2004).

In this thesis, convergence is assessed prior to determining a suitable period of 'burn-in', by means of the history trace, and the methods of Brooks & Gel-

man (1998), where there is more than one chain being sampled. Autocorrelation often occurs for some parameters within a model but not for others; to address this issue, generous sample sizes are used for all models (usually 50 0000 iterations), as discussed in the appropriate chapters (Chapters 7, 9, 10 and 11).

4.3.4 Prior distributions

Due to the potentially strong influence of the prior distribution in determining the posterior distribution, it is important to select a prior distribution for all parameters in the model that will accord with the current thinking regarding the model, the dataset and the research question. In many instances, it is desirable to make no prejudgement regarding the nature of the posterior distributions, hence any prior distribution should make as little impact on the posterior distribution as possible. In effect, the posterior distribution should reflect the data without being influenced by the prior distributions, or alternatively, the data should 'overwhelm' the prior distributions. Such prior distributions are referred to as 'vague', or non-informative.

Common distributions for a vague prior are the normal distribution, with large variance, or the uniform distribution, with a large range between its parameters. In situations where only positive values are feasible for a parameter, such as a standard deviation, a half-normal distribution can be applied. The half-normal distribution is derived from a normal distribution, centred on 0, with only the positive half of the distribution used, whilst the negative half is discarded. Alternatively, this model can be thought of as a normal distribution folded around 0 (Spiegelhalter *et al*. 2004).

Alternatives to a non-informative prior include the 'sceptical' prior, which is based on the concept that the null hypothesis is indeed true (and hence is 'sceptical' about the possibility that the true treatment effect demonstrates a difference in the treatment group compared to the control group). Conversely, a prior distribution that is based on the alternative hypothesis being correct is an 'enthusiastic' prior (in that it predisposes the posterior distribution to represent a difference between the comparison groups, presumably in the direction favoured by current thought). Prior distributions are discussed at greater length elsewhere (Spiegelhalter *et al*. 2004).

It is often difficult to specify suitably vague priors for a variance parameter; this issue has been discussed at length by Lambert *et al*. (2005). These authors make a comparison of 13 priors, 12 of which (B1–B12) are set out in full in Table 7.6 in Section 7.4.5, as they are used in a sensitivity analysis in Chapter 7.

The 13th prior distribution was based on the logistic distribution, and included the number of studies and the within-study variance for each study; this prior was based on previous work by DuMouchel & Normand (2000), as cited by Lambert *et al*. (2005). These priors were placed on the standard deviation or a function thereof, such as the variance, log of the variance, or precision (reciprocal of the variance). This study showed that differences in supposedly 'vague' prior distributions could lead to different conclusions. It is pointed out that as the dataset increases in size, the influence of the prior distribution is reduced, but, with specific relevance to adverse events data, the problem becomes more prominent when the between-studies standard deviation is close to 0. Due to the sparsity of events often found with adverse events, it may be the case that studies will exhibit little difference in their treatment effects.

In the study by Lambert *et al*. (2005), there was no specific prior distribution that performed well; the main finding from this study, in a practical context, was the importance of sensitivity analysis across multiple priors. Another recommendation from this study is to use previous empirical work to derive the prior distribution. It is also important to ensure that prior distributions encompass only realistic values for the standard deviation, and to check for convergence routinely. It was noted that convergence was a potential difficulty when the estimated between-studies standard deviation was close to 0, due to the enforced sampling [although in such cases a fixed effect model may have been a more appropriate option].

The prior distributions discussed (with the exception of that based on the logistic distribution) are evaluated in Chapter 7. The suggestion of using previous empirical work to inform a prior distribution is also acted upon with regard to the dataset being analysed; this approach to adverse events data is a novel one, but could be extended in several ways, for example by using observational data to inform a prior distribution for analysis of randomised studies. Comparison of prior distributions also forms part of the Mixed Treatment Comparisons (MTC) analysis of Chapter 10.

4.4 Bayesian meta-analysis methods

4.4.1 General meta-analysis models

The flexibility of Bayesian software allows the creation of meta-analysis models of varying degrees of complexity. The simple fixed effect (FE) model is easily constructed, as is the standard random effects (RE) model. The major difference between a frequentist FE model and a Bayesian FE model is that a prior distribution is placed on the true underlying value of the treatment effect θ.

For the RE model, as well as the prior on θ, the between-studies variance (τ^2) also has a prior placed on its distribution. This is in contrast to the assumption that this statistic takes a specific value, for which no accounting is made for potential variation – the usual assumption for a frequentist RE model.

One example of a frequentist approach that does allow for uncertainty between entities within a model has been proposed by Hardy & Thompson (1996). These authors put forward a likelihood-based method for calculation of the between-studies standard deviation that takes into account potential variation, unlike the method of moments estimator calculated by the more commonly-used DerSimonian & Laird method.

Returning to Bayesian methods, the prior distribution on τ^2 is a fundamentally important piece of information for the RE model. Its distribution can be chosen according to prior beliefs or based on empirical evidence, for example, from other primary studies. The importance of performing a sensitivity analysis across different prior distributions, and the inclusion of empirical evidence to form prior distributions on stochastic parameters, including τ^2, is demonstrated in Chapter 7.

The ability to develop models to a greater degree of complexity is a valuable asset to a Bayesian analysis. For example, the models can be extended to develop 'hierarchies' that may be based on factors such as treatment regimes or study types. The prior distributions placed on hyperparameters within the models can be manipulated to reflect the fact that some studies or treatment regimes may be more strongly related compared to others within the dataset. The use of hierarchical models, as discussed by Prevost *et al.* (2000), is included within the mixed treatment comparisons (MTC) analysis in Chapter 9.

Several advantages of Bayesian modelling for meta-analysis are discussed by Sutton & Abrams (2001). These advantages include the ability to incorporate all forms of uncertainty in the model for each parameter, and the ability to make predictive statements, including the degree of uncertainty.

Another advantage is the ability to make comparative statements between two treatments, for example the probability that one treatment has a better odds of success than a comparison treatment. With regard to clinical decision-making, the use of Bayesian modelling allows the incorporation of costs and utilities, which assist in making healthcare policy decisions. Due to the flexibility to create complex algebraic models and apply these with Bayesian methods, evidence from a variety of different sources can be included within the model.

In situations where many models can be produced of varying degrees of complexity, these models can be compared using the deviance information criterion (DIC) presented by Spiegelhalter *et al*. (2002), and the *pD* statistic, which is calculated as the 'effective number of parameters' and is a measure of goodness of fit that takes into account the number of model parameters, which the DIC does not. Both of these values are considered to indicate a better fit of model at lower values.

Also, the residual deviance (and the sum thereof) for a specific model can be calculated as a function of the model itself. These methods can be used in combination to select the most suitable model for a particular circumstance, and are applied in Chapters 9 and 10. The DIC and pD are applicable to Bayesian modelling in general, and are not exclusive to meta-analysis modelling.

The sum of deviance is of value in assessing the absolute degree of 'fit' for each model, whilst the DIC can be used to evaluate the relative 'goodness' of fit across multiple models. These values can be used in conjunction to determine the statistical value of each model. This evaluation can then be used alongside clinical considerations to select the most appropriate model for clinical requirements, with an understanding of its statistical merits.

4.4.2 Binomial meta-analysis models

As all of the outcomes considered in this thesis are binary, the binomial model is the model of choice. The basic fixed effect (FE) model is as follows:

$$\begin{aligned} r_i &\sim \text{Binomial}(p_i, n_i), \\ \text{logit}(p_i) &= \mu_{s_i} \text{ for control trial arms,} \\ \text{logit}(p_i) &= \mu_{s_i} + \delta_i \text{ for treatment trial arms.} \end{aligned} \tag{4.7}$$

In this model, r_i refers to the number of observed events in study arm i, p_i refers to the probability of an event occurring in study arm i, n_i is the number of participants in study arm i, μ_{s_i} is the log odds of an event occurring in the control arm i of study s, and δ_i is the log odds ratio of an event occurring in the treatment arm in comparison with the control arm for study s to which arm i belongs.

This model requires a prior distribution for the parameters μ_j and δ_i, where j indexes the study.

It is straightforward to extend this model to include a random effect on the between-studies variation. A prior distribution is placed on δ_i, for example,

$$\delta_i \sim \text{Normal}(d, \tau^2), \tag{4.8}$$

where d is the mean true underlying odds ratio, and τ^2 is the between-studies variance. Other distributions are also feasible. A prior distribution is also required for d, for which the normal distribution may also be used.

This model, or a variation thereof, is used in Chapters 7, 9, 10 and 11.

4.4.3 Meta-analysis in practice

Bayesian meta-analysis as a practical exercise also has several advantages, as pointed out by Sutton & Abrams (2001). The necessity to elicit prior beliefs enforces those posing the clinical question to consider the issues carefully, and determine precisely their true prior beliefs regarding what the 'answer' to the question should be. For example, in a clinical trials setting, the clinician needs

to determine the magnitude of difference between two treatments that would be clinically significant. However, it can also be argued that such reliance on prior belief and 'expert opinion' introduces too much subjectivity into any analysis. This issue can be addressed by use of a sensitivity analysis across a range of prior distributions. The influence of prior distributions can be especially strong where the data are sparse (in terms of few studies or few events within the studies); the use of priors with regard to sparse data is discussed by Lambert *et al*. (2005), see Section 4.3.4.

The use of prior distributions is one method, unique to Bayesian analysis, that can be used to synthesise evidence from randomised studies and observational studies. Although observational and randomised data can be pooled using frequentist methods, an advantage of using Bayesian methods for this type of analysis is that the analysis can take into account the different types of data.

For example, the observational evidence can be used to inform the prior for the analysis of the randomised evidence. Different approaches to this type of prior distribution construction are presented by Sutton & Abrams (2001). If the observational evidence is considered to be of high quality it can be used directly as a prior for the randomised data analysis. This is denoted as a 'naïve' prior, as this evidence is effectively being considered on an equal footing with the randomised evidence. An 'equivalent' prior is centred at the pooled estimate from the observational data, but has a variance taken from the meta-analysis of the randomised studies, τ^2, the between-studies variance. This method aims at constraining the influence of the often larger observational studies, which tend to be larger in numbers compared to randomised studies, and hence may have greater weighting in the meta-analysis. A 'sceptical' prior can downgrade the influence of the observational studies even further by artificially and arbitrarily increasing the variance on the prior around the centred estimate. This would be appropriate if there were concerns regarding serious biases in the observational studies.

An overall meta-analysis using all the evidence from all study designs analysed in one model produces similar results to those of a naïve prior (Sutton & Abrams 2001). One way of addressing the issue of different study designs is to use a model with different levels of heterogeneity. As well as using a between-studies heterogeneity parameter for different studies within a particular study type, an additional level of heterogeneity can be added to denote the

heterogeneity expected to occur between study types. This type of analysis can therefore provide an estimate for each study outcome, the overall pooled estimate and the pooled estimate across all studies of the same design (effectively a hierarchical model on study design). This model can be viewed as a form of random effects model with an additional level of heterogeneity.

Another advantage is that because of the iterative process of repeated sampling, the different units of analysis in the dataset (i.e. the individual studies) can 'borrow strength' in terms of reflecting the data of other studies in determining the results of an individual study. This process will tend to draw together the individual treatment effects, reducing extremity of results, and will also reduce uncertainty for each individual study's treatment effect.

This ability to 'borrow strength' across studies lends itself to hierarchical modelling (Prevost *et al*. 2000), in that certain datapoints (with a common factor such as study type or treatment allocation) can be linked together through an hierarchical model, and can then 'borrow strength' from more closely-related datapoints than from other datapoints in the overall dataset. Hierarchical modelling is discussed in greater detail in Chapter 10, within the context of an MTC model, evaluating the influences of class of drug, and drug dosage.

As a practical disadvantage in employing Bayesian methods, computational issues can mean that meta-analyes are potentially more time-consuming to perform (Sutton & Abrams 2001).

Other advantages of Bayesian methods include the lack of requirement of a continuity correction for studies with zero events in one or both arms. This problem of sparsity of events within a study is very pertinent to adverse events meta-analysis and is discussed more fully in Section 5.2.4.

A logical extension of the Bayesian model is to include both efficacy and adverse events data within the same model, and potentially cost-effectiveness data could also be combined to result in a framework of interest both clinically and in terms of health policy and decision-making. A case-study making use of these concepts is presented in Chapter 11. A Bayesian approach facilitates this form of modelling, with data regarding both adverse and positive effects of an intervention being incorporated into a single model, as it allows uncertainty to be propagated throughout the model, and hence to be expressed in the outcome metric of interest (Sutton *et al*. 2005).

4.5 Summary

Statistical principles of Bayesian inference are discussed, with reference to the more traditional frequentist models. The application of Bayesian modelling has many practical difficulties; however, these are now largely dealt with by the use of modern software. Practical issues include selection of prior distributions, checking for convergence of the samples to the equilibrium posterior distribution, and methods to compare the goodness of fit of multiple models of varying complexity. There are many advantages associated with meta-analysis modelling using a Bayesian approach, as well as some disadvantages. Some of the major advantages include the ability to incorporate external evidence into a model, by use of prior distributions, the fact that all aspects of uncertainty can be included into the model and that modelling within a Bayesian framework can lend itself to more complex models, compared to standard techniques.

5

Meta-analysis challenges with regard to adverse events data

5.1 Introduction

This chapter discusses issues that produce difficulty in the execution of standard meta-analysis methods, and that may be likely to arise in meta-analyses of adverse events data, although they are not exclusive to meta-analyses where the outcome is an adverse event.

These aspects of evidence synthesis that are relevant when considering adverse events data were outlined in Chapter 1 and are repeated here briefly, before being discussed in more detail below.

1. Sparse data (zero events in one or both treatment arms).
2. Heterogeneous data sources (e.g. observational and randomised studies).
3. Multiple outcomes.
4. Combining summary data with individual patient data (IPD).
5. Subgroup analysis.
6. Dose–response data.
7. Class effects.

8. Time–course effects.
9. Reporting bias.
10. Evidence synthesis of risks and benefits.

Also, the notation conventions set out in Section 3.1 are continued in this chapter.

Before entering into a discussion about the issues relating specifically to statistical methods, it is helpful to consider the wider context regarding systematic reviews of adverse and/or unintended events.

Systematic reviews of adverse event data have been the area of interest for several authors. For example, McIntosh *et al.* (2004), discuss their experiences of conducting systematic reviews of adverse event data, focusing on specific areas, namely the review question, the issue of differences in study design and difficulties in quality assessment. Although these authors do not specifically address any statistical issues, their work is relevant, because it should be remembered that meta-analysis cannot be performed without consideration of the surrounding methodologies. The importance of a focused review question, highlighted as being important for systematic reviews, can become even more essential for a quantitative evidence synthesis.

Difficulties or 'challenges' in systematic reviews of adverse events have been set out by Chou & Helfand (2005). They discuss three primary areas where the systematic review of adverse events is less than straightforward. The first of these is the identification and selection of information about harms. Heterogeneity of data sources is a fundamental issue. The challenges include excessive reliance on trials data, the caveat being that although trials are often the 'gold standard' for efficacy data, they often address harms outcomes inadequately, either due to the study design (for example, a small sample size or exclusion of patients at high risk of harm), or poor reporting of adverse outcomes. Randomised trials are often less suitable for surgical and device interventions as compared to drug therapies, due to difficulties in blinding, and ethical issues (McLeod 1999). Observational studies, such as cohort and case-control studies, may be more susceptible to bias and confounding than experimental studies.

Citing earlier work, Chou & Helfand (2005) point out that well-designed controlled observational studies and randomised trials can produce similar results

regarding effects; it has not yet been evaluated how observational studies produce different yet valid conclusions about harms. They discuss a report on surgical complications, and found that clinical trials reported a higher risk for adverse events, compared to observational studies, and speculate that this could be due to poorer assessment of harms in observational studies, or that observational studies may be less likely to be reported if they include deleterious results. Large databases such as pharmacoepidemiological databases may be useful in identifying adverse events (due to higher reporting rates than in other types of data). Practice-based databases are another source of data that may be applicable to patients in a community setting. Case reports can be of use in collecting data on longterm or uncommon adverse events. Pharmacological data can be used to shed light on the occurrence of adverse events in certain populations, where subgroup data are unavailable.

The second challenge is the assessment of the quality of reporting of adverse events. The data type could be randomised trials, observational studies, case reports or uncontrolled studies of surgical interventions, all of which present their own problems in quality assessment with regard to harms.

The final challenge is that of synthesising and displaying data from different types of studies. One of the main areas for concern in combining data from different study types is that observational studies carry a larger propensity for confounding and bias to distort the results. It is also important to balance the need for conciseness with full and transparent reporting. Citing earlier work, Chou & Helfand (2005) discuss methods of displaying relevant factors in summary tables conveying information about the studies and results. These authors also set out various recommendations for improving systematic reviews that assess harms; at the data analysis level their main recommendations include the avoidance of inappropriate combination of data and the investigation of any heterogeneity.

The use of systematic reviews for adverse effects is also discussed by Loke *et al*. (2007), who also set out a framework for such reviews. Regarding quantitative data combination, their assertion is that the data from heterogeneous sources cannot be pooled using standard meta-analysis techniques, due to differences in study design, study population, or data collection methods. Data from observational studies are often more prone to bias and often [statistically] heterogeneous, and should not be combined if this is the case. It may not be pos-

sible to directly compare harms and benefits. However, more recent techniques are being developed to address these issues, including the work in Chapter 11 regarding combination of data on harms and benefits.

It can be seen that, even prior to any analysis of quantitative results data, there are several issues that need to be addressed when synthesising data regarding harms and unintended effects. While these issues, such as data collection and quality assessment, are not the main focus of this work, they should be borne in mind, as they often impact on other aspects of the analysis, and will certainly influence the final conclusions, generalisability and degree of confidence in the results.

Chapter 6 presents the results of a systematic review of previous meta-analyses, where the primary outcome(s) was an adverse or unintended event. All of the areas of contention, set out above, were included in the review, with the exception of harm–benefit analyses.

5.2 Sparse data

The term 'sparse data' can have different meanings depending on the context, for example, within a meta-analysis sparse data may refer to primary studies where there are zero events in one of the study groups (for a comparative study). Alternatively, it could refer to a scenario where there are few studies of the appropriate characteristics to be included in a meta-analysis. In this section, the term 'sparse data' is used in the first context, with regard to comparative (two-arm) studies.

This issue has been considered to some extent in Chapter 3, whilst discussing different meta-analysis methods and how they perform with zero event counts. Different outcome metrics also behave differently with zero counts. For example, a ratio metric will be 0 if there are zero events in the treatment group, and will be undefined if there are zero events in the control group or in both groups. A risk difference can be calculated if there are zero events in either or both groups, but the variance of such a study will be 0 only if there are zero events in total (Section 3.6).

The different meta-analysis methods, such as inverse variance (IV), Mantel–Haenszel (M–H), Peto and the random effects DerSimonian & Laird method,

differ in their ability to cope with primary studies with zero events in one or both arms, as discussed in Sections 3.3.3, 3.3.5 and 3.5.1. The different models require the use of continuity corrections (discussed in more detail in Section 5.2.3) under different circumstances, although it is important to be mindful, when performing a meta-analysis, of the default application of continuity corrections by the software package being used.

The major issues presented by data with sparse events are:

1. choice of meta-analysis method;
2. choice of outcome metric in the light of meta-analysis method;
3. choice of continuity correction;
4. inclusion or exclusion of studies with zero events in total; and
5. use of prior distributions in Bayesian analyses.

These issues will be discussed in further detail.

5.2.1 Choice of meta-analysis method for sparse data

Choice of meta-analysis method initially requires a decision as to whether frequentist or Bayesian methods are preferable. Bayesian methods have certain advantages in that they do not require use of continuity corrections to include studies with zero events in one or both arms. A disadvantage is that choice of prior distributions on stochastic parameters may be difficult, as vague prior distributions may in fact influence the analysis (to be discussed in Section 5.2.5). However, the use of prior distributions can be advantageous in that data from previous studies (primary studies or meta-analyses, or even clinical opinion) can be used to frame a prior distribution.

If frequentist methods are used, standard meta-analysis methods may result in differences in the outcome and conclusions. One major contribution to this issue is by Bradburn *et al*. (2007), and it is helpful at this stage to discuss the methods of these authors, and refer back to their specific findings and conclusions as appropriate. The premise of the paper was to compare multiple meta-analysis methods in their performance with regard to datasets where sparsity of events was an issue. Whilst individual studies may lack adequate power to detect

rare events, a meta-analysis of several studies may increase power substantially, but there may be aspects of the results of the meta-analyses that differ across methods, therefore it is important to assess how the methods perform under different conditions and by different parameters.

The methods to be evaluated used both odds ratios (ORs) and risk differences (RDs) as the outcome measures. Seven methods were used for ORs:

1. inverse variance;
2. DerSimonian & Laird;
3. Mantel–Haenszel with continuity correction;
4. Mantel–Haenszel without continuity correction;
5. Peto;
6. exact stratified (by study); and
7. logistic regression maximum likelihood.

Note that the only random effects (RE) model is that of DerSimonian & Laird; all the others are fixed effects (FE). Another point to note is that the authors only used a continuity correction (0.5 in all cases) when absolutely necessary algebraically for calculation of a pooled estimate or variance. This may be at odds with some standard software packages, which will include a continuity correction by default for all studies with one count of zero events in either arm.

These authors use the Mantel–Haenszel model with and without a continuity correction (the M–H model is discussed further with regard to application of continuity corrections in Section 3.3.3). It is unclear exactly how Bradburn *et al*. (2007) addressed the issue of zero events for the uncorrected M–H method. If there are zero events in the control arm (when using the OR as the outcome measure) then the individual study weight is 0, so the study will not contribute to the pooled estimate (which is helpful as the OR would be undefined). A continuity correction for the variance of the pooled estimate is only required in cases where there are zero events for the corresponding arms of all studies, so in such cases a pooled estimate would be produced but no confidence interval (CI).

For the RD, the following meta-analysis methods were used:

1. inverse variance;
2. DerSimonian & Laird; and
3. Mantel–Haenszel.

Note that for the M–H method with an RD outcome, a continuity correction is not required for calculation of the individual study estimate or weight, but the variance of the pooled estimate will be 0 if there are no events across all studies. (The use of the RD in meta-analysis is discussed further in Section 3.6.) This situation did not arise as all studies with zero events in total were excluded from all analyses.

The Peto method (discussed in Section 3.3.5) has the advantage of not requiring a continuity correction, and automatically excludes all studies with zero events. Logistic regression methods do not allow non-integer counts and so do not allow non-integer continuity corrections. Also, logistic regression excludes trials with zero events across both arms, but can incorporate those studies with just one arm with zero events. However, they also encounter difficulties when all studies have zero events in one of the treatment arms. Exact methods are applicable in situations where there are zero events in one or both arms of an individual study, but cannot produce an estimate of the treatment effect when there are zero events in all corresponding groups of all trials, and cannot estimate statistical significance and confidence limits when all trials have no events in both groups.

Finally, unstratified methods (see Section 3.2) were used to pool the data across studies, using marginal totals, for both the OR and RD. This is the only model that can incorporate individual studies with zero events for the OR metric.

The differences in how meta-analysis methods address problems of zero events in one or both arms of an individual trial, and similar issues of zero events in all studies or in the corresponding arms of all studies, are important when selecting a meta-analysis method. These highlight the importance of scrutinising the dataset comprised of primary study results, to determine the extent of the problems of zero events.

These meta-analysis methods were compared using simulated data based on different scenarios. The scenarios differed according to the number of studies in each meta-analysis, whether the studies were balanced in terms of the numbers

in the two comparison groups, the baseline event rate (varying between 0.1%–10%), and the treatment effect (relative risk varying between 0.2–1 for the balanced studies, and 0.2–5 for the unbalanced studies).

The different meta-analyses methods were evaluated on four parameters:

1. bias: the difference between underlying treatment effect and mean observed treatment effect (pooled estimate), with bias for OR converted to an absolute risk scale;
2. coverage: percent of simulations for which the 95% CI of the pooled estimate included the true value;
3. statistical power: percent of simulations with a statistically significant result in the direction of the underlying effect; and
4. estimability: the percentage of simulations for which an estimate and significance test could not be produced or for which zero-cell corrections were required.

The only RE method used was the DerSimonian & Laird method, which performed poorly in the simulated tests. As the authors point out, the DerSimonian & Laird method is based on large sample theory, which is not suitable when there are rare events. Also, the continuity correction of 0.5 may introduce bias into the results (although this is not specifically mentioned by the authors). The DerSimonian & Laird method appeared to be biased for both OR and RD outcome metrics, with an increase in bias for lower baseline event rates, and greater treatment effects. This effect was seen for the balanced studies with both 19 and 5 primary trials.

Power was also poor at lower event rates and smaller treatment effects. For these reasons, the authors conclude the DerSimonian & Laird method is not suitable for sparse data analyses and do not include this method in simulations of unabalanced study designs. Overall, it was concluded that consideration of between-study heterogeneity was of little importance when analysing sparse data, a conclusion that was supported by the work of Sweeting *et al.* (2004), who found that in their simulations, despite including between-studies heterogeneity, it was rarely detected in meta-analyses.

Considering the FE models, the authors dismiss the IV method for similar reasons as the DerSimonian & Laird method. It is also based on large sample theory,

and showed bias for both OR and RD outcome metrics. Again, power was low for lower baseline event rates and less pronounced treatment effects. The IV method was also discarded for unbalanced study designs.

The Peto method appeared to be less biased at lower event rates, with small or moderate treatment effects, than other FE models, including the M–H, both with and without continuity corrections, logistic regression and exact methods. Regarding the M–H models, the model with the continuity correction appeared to be more biased than the uncorrected method. Power across the four models appeared to be comparable for the balanced study designs for the OR metric, although for the unbalanced studies, the M–H method with the continuity correction appeared to have greater power for ORs less than 1, and lower power for ORs greater than 1. The Peto method performed well across most simulations, apart from scenarios where the study numbers were unbalanced.

This study did not offer any evidence that exact methods were superior to other methods for sparse events. The Peto method has been shown to be biased in scenarios where primary studies are unbalanced in the sizes of their comparison groups (Greenland & Salvan 1990). These authors also argue that even when the study groups are suitably balanced, the bias introduced by the Peto method can be significant if the treatment effect is large, which may occur in meta-analyses of adverse events where the outcome is rare in untreated groups. In fact, the Peto method was designed with the aim of being suitable for meta-analyses methods where treatment effects are small.

Overall therefore, the Peto method may be a suitable option if two criteria are fulfilled; firstly, that all studies are reasonably balanced in the sizes of the groups, which is likely to be the case when performing meta-analyses with experimental trials rather than observational studies; and secondly, that the treatment effect is not excessive. Bradburn *et al*. (2007) argue that whilst an unstratified analysis is usually undesirable, in studies where events are rare, there is less likely to be major differences in proportions of events across trials, and therefore such methods will be at less risk of confounding by trial due to Simpson's paradox. Logistic regression and exact methods may also be worthwhile, whereas the M–H model may be biased if used in conjunction with a continuity correction.

5.2.2 Choice of outcome metric

From the work of Bradburn *et al*. (2007), despite the RD having certain advantages such as the ability to incorporate studies with zero events, all methods using this outcome metric were found to have low statistical power, and very conservative coverage (i.e. the CIs were too wide to be of value). These drawbacks make them unsuitable for studies with sparse events. The simulations were all performed based on the assumption of fixed effects, and if between-studies heterogeneity were to be included in the simulation process, then the RD would perform less well, due to variations in the relative treatment effect across different studies.

Despite this conclusion, an RD is valuable in assessing the absolute difference in risk, which may be more useful in clinical terms than a relative measure, especially if the baseline event rate is variable across studies. The RD is also helpful in that it can be used to calculate the Number Needed to Treat/Harm (NNT/NNH; discussed in Section 3.6). A comparison of both OR and RD would help to present a clearer picture for clinical decision-making.

5.2.3 Choice of continuity correction

Where there are studies in a dataset with a zero count in only one arm, the standard method for incorporating the data from this trial is to use what is termed as a 'continuity correction', although this phrase may be inaccurately used in this context. A continuity correction is added to a discrete variable, when that variable is being analysed by a method designed for the analysis of a continuous data (Yates 1934). For example, a continuity correction, usually 0.5, is added to a variable generated from a binomial distribution (a discrete distribution) when an approximation to a normal distribution (a continuous distribution) is being applied to analyse the data.

The continuity correction in conjunction with 2×2 tables is discussed by Plackett (1964). The use of the continuity correction of 0.5 may also have other benefits, for example, by lessening the bias in the estimator of the treatment effect and its variance (Spiegelhalter *et al*. 2004). The addition of a continuity correction will have a negligible effect on the treatment effect for large sample sizes.

In the context of meta-analyses of adverse events data, the data are usually in the form of counts and as such are naturally integers, but the methods for analysis of multiple datasets often cannot produce defined results if there are zero events in one or both study arms. A similar situation arises where there are no 'non-events' in either arm of the trial, but as this eventuality is unlikely in the context of adverse events it is not considered further.

Using standard statistical software (such as user-written commands for Stata®, e.g. **metan**), a continuity correction is automatically provided when using many standard commands to perform a meta-analysis using the IV or M–H method, or a random effects (RE) DerSimonian & Laird model. (Not all Stata® commands do in fact use a continuity correction by default; those that do not use a continuity correction do not appear to have any means of dealing with occurrences of zero events.) A continuity correction is provided for all studies with one arm with zero events. Those studies with zero events in total are excluded by default when the outcome metric is on a ratio scale.

A continuity correction effectively assists with three different issues; firstly, the calculation of the individual study estimated metric (to avoid a metric of 0 or 'undefined' when on a ratio scale); secondly, the calculation of individual study weightings (to avoid weightings of 0 or 'undefined'); and thirdly to calculate the variance of the overall pooled estimate. The calculation of the overall variance of the pooled estimate is assisted indirectly by the use of a continuity correction for the IV method (where the variance is the reciprocal of the sum of the study weights), and is only required for the M–H model where there are zero events in all corresponding arms across the primary studies.

The usual continuity correction applied is 0.5, which is added to all four cells of the 2×2 table, thus adding into the study two 'false' participants, and one 'false' event. Discussing the work of Agresti (1996), Sweeting *et al.* (2004), mention that smaller fixed value alternatives to a continuity correction of 0.5 have been considered; this approach would result in fewer 'false' participants added to each study.

The work of Bradburn *et al.* (2007) found that the M–H method with a 0.5 continuity correction was more biased than the same without the continuity correction. The DerSimonian & Laird method and IV method, which both require a continuity correction, were also associated with bias. Hence, it is

reasonable to conclude that the use of a continuity correction, especially with sparse events data, may result in bias.

The use of continuity corrections with the context of adverse events data, where it is likely that many studies may have zero events in one or both arms, presents two major challenges. The first is the choice of continuity correction. Although 0.5 is the standard default, any continuity correction would perform the required tasks. This issue has been investigated further by Sweeting *et al*. (2004). These authors evaluated a fixed continuity correction against two alternatives, one based on the reciprocal of the size of the opposite treatment arm, which would have the effect of influencing the final pooled estimate of the treatment effect towards the 'no effect' value. The other was based on the concept of using all studies without zero events to create an overall OR, which would then be used to derive a continuity correction such that the use of the correction would influence the final result (based on all trials including those with one arm with zero events). These methods tended to produce less bias than a standard continuity correction, varying according to the degree of imbalance in the treatments and the meta-analysis method.

The second is in the use of continuity corrections. For the IV and DerSimonian & Laird method a continuity correction is essential for calculation of the study OR if the number of events in the control group is zero, and for the individual study weightings, and hence the variance of the pooled estimate of the treatment effect. For the M–H method, however, the individual study weightings are calculable if the number of events in the treatment arm is zero (which will also result in an OR of 0); similarly, if the number of events in the control arm is zero, the OR is undefined but the study weight is 0, hence a continuity correction is not needed for these calculations. A continuity correction is only required if there are zero events in all corresponding arms of all studies.

The default for standard software (Stata®) appears to be to add a continuity correction for all IV, M–H and DerSimonian & Laird methods for all studies with a zero count in one arm. This continuity correction appears to be used for all subsequent calculations even when not specifically required. For example, the continuity correction is used for calculation of the OR when the treatment group has zero events, which without the continuity correction would simply have an OR of 0. It is interesting to note in the work of Bradburn *et al*. (2007)

that the uncorrected M–H method appeared to be associated with less bias, compared to the use of continuity corrections.

Overall, continuity corrections should be used judiciously, both in selection of the most appropriate correction, according to the study numbers and meta-analysis method, and in terms of when the correction is applied. Investigation of potential differences in results due to adding a continuity correction to all cells of a study where only one arm has zero events, and then using the corrected value for all subsequent calculations, could be compared to a scenario where only those calculations that require a continuity correction to avoid an undefined value actually use the corrected values. However, the use of different values for the corresponding cell of the 2×2 table in itself may introduce bias.

5.2.4 Dealing with studies with zero events

It can be argued that studies with zero events in total do not contribute to the pooled estimate of treatment effect size when using a ratio metric (Sweeting *et al*. 2004). Similarly Bradburn *et al*. (2007) excluded all studies with zero events in total from their meta-analysis method comparisons of ORs, with the exception of crude unstratified methods. Contrary to these standpoints is the intuitive argument that these studies convey some information relative to their overall size, and that ideally some form of inclusion into a meta-analysis would be desirable.

The Peto method and logistic regression methods automatically exclude studies with zero events. Exact methods however, are an option for including studies with zero events (unless all trials have zero events in one or both groups). Unstratified methods can also include studies with zero events in both arms, but will produce an estimate of the OR that is either 0 if all trials have zero events across the treatment groups or 'undefined' if all trials have zero events across the control group or both groups.

On an RD scale, studies with zero events are included in standard meta-analysis methods, although continuity corrections must be required for calculation of individual study weights when using the IV method, and there are zero events in both study arms. When using the M–H method, the individual study weights can be calculated if there are zero events across the whole study, but the variance of the pooled estimate will be zero in cases where there are zero events in both

arms in all studies in the dataset. Unstratified analyses can include studies with zero events on an RD scale.

One method to incorporate studies with zero events is to use continuity corrections where required to enforce inclusion of such studies into a standard meta-analysis model. However, this approach creates 'false' participants and 'false' events, to an extent determined by the size of the continuity correction. Considering the arguments both for and against the inclusion of studies with zero events, a suitable compromise may be to perform multiple meta-analyses including studies with zero events, and to compare these with meta-analyses excluding these studies. The exploratory meta-analyses performed in Chapter 7 include this approach.

When dealing specifically with studies related to adverse events, it is likely that studies with zero events will be a common occurrence. Intuitively, these studies contribute information to the overall dataset and it would be highly desirable to incorporate them in some way. An unstratified analysis is superficially the most straightforward way to accomplish this, but would be associated with concerns regarding bias.

Another approach to include studies with zero events is to use Bayesian methodology. Bayesian methods can include studies with zero events (in one or both arms) without recourse to a continuity correction, for both FE and RE models. There may be issues regarding the appropriate use of prior distributions, and this is discussed further in Section 5.2.5, whilst Bayesian methods in general are discussed in Chapter 4. The binomial model, which supports this ability to incorporate studies with zero events is set out in Section 4.4.2. Another advantage for Bayesian methods when using an RE model is that they can easily estimate the between-studies variance, by using an RE model, and placing a prior distribution on τ^2.

Bayesian methods are computationally suitable for an OR outcome, due to the fact that the normal distribution can be used to model the log OR, but there are issues for the relative risk (RR) and RD outcomes. However, suitable models can be created for both the RD and RR scales, allowing the actual values for the RD and RR to be constrained within appropriate values, using suitable prior distributions. For example, an RD may be constrained between -1 and 1 over a uniform distribution. For an RR, the range is bounded on the log scale between

$-\infty$ and the negative of the log of the risk in the control group (which occurs if the risk in the treatment group is 1, leading to the maximum possible value for the RR). A distribution can then be placed on the log RR for each study, derived from a normal distribution. The log RR can be replaced by the negative of the log of the risk in the control group, should the log RR sampled for a particular study exceed this maximum allowable value.

With regard to adverse events modelling, it is likely that the OR scale will be more satisfactory than the RR scale due to its mathematical properties and the fact that the OR will approximate the RR for the low event rates that are likely to occur. As pointed out earlier, the RD scale is likely to have lower statistical power compared to the OR scale, but this is less of a consideration for a Bayesian analysis, so the RD scale may be of value, if only as a comparison against the OR scale.

5.2.5 Choice of Bayesian prior distributions

When data are sparse in meta-analysis (both in terms of few primary studies and few events within those studies) there is a concern that any prior distributions used in a Bayesian model will overwhelm the available data and have an unacceptable influence on the results. For example, in a study where there are zero events overall, although the Bayesian model has the advantage of incorporating such a study without a continuity correction, the prior distributions placed on the estimates of the outcome metric(s) and variance parameter(s), or functions thereof, will exert their influence over the contribution of the study to the model output.

The work of Lambert *et al.* (2005) has highlighted this issue by comparing multiple supposedly 'vague', or non-informative, prior distributions. These prior distributions were placed on the scale parameter (variance or a function of the variance) of simulated RE meta-analyses. Ideally, a vague prior distribution should allow the data to dominate the posterior distribution across the range of values of interest, and therefore should treat all possible values of the parameter as being equally likely. These authors simulated multiple meta-analyses with different numbers of primary studies (five, ten and 30). However, these simulations did not include sparse events within each trial (each trial had 50% of patients

in the control group experiencing the event of interest, with an underlying OR of 1.38).

Hence, these simulations do not consider sparse data at the level of events per study. Using different prior distributions, there were differences in the 95% credible intervals (CrIs) for the log OR and the posterior distributions on the standard deviation. There were also difficulties when the true between-study standard deviation was close to 0, because the model enforced a positive value to be sampled at all times. This may imply that an RE model should not be attempted unless there is good reason to assume that an FE model would be inappropriate.

The influence of the prior distribution was more pronounced when there were fewer studies in the meta-analysis. With larger numbers of studies the data dominated the prior distribution to a greater extent. It may be reasonable to extrapolate this conclusion to a scenario where the number of studies is greater, but the events themselves are sparse, with zero events in one or both arms of a study. These authors concluded that there was no particular prior distribution that performed well (in terms of not influencing the data) in all scenarios, and hence that a sensitivity analysis across prior distribution should be performed. Certain prior distributions appeared to perform poorly with small numbers of studies, in particular those that were uniform on the variance scale.

Non-informative prior distributions can be valuable if they truly *are* non-informative, but may be dangerous should they exert undue influence on the results of an analysis. This phenomenon becomes more pronounced when using hierarchical models, in that the variance parameters may exert influence which is then propagated throughout the model.

Therefore, it could be appropriate to take a contrasting approach, and select a prior distribution that is expected to influence the results in the light of previous data on the same or similar areas, or based on clinical experience, pharmacological information or some other viable source of evidence. Using an informative prior distribution in this way would be analagous to the continuity correction developed by Sweeting *et al.* (2004), which was based on using data from those studies that did *not* require any continuity correction to calculate a continuity correction to apply to those studies that *did* require one. This is an area that would benefit from further investigation.

5.2.6 Alternative methods to address sparse data

Two novel approaches for dealing with sparsity of data have recently been developed (Rücker *et al*. 2009; Tian *et al*. 2009). The first of these methods (Rücker *et al*. 2009) discusses the use of the arcsine transformation, in terms of its application in meta-analyses where the primary studies include at least one instance of zero events in total.

The arcsine difference can be defined as:

$$\text{arcsine}\sqrt{\frac{a_i}{a_i+b_i}} - \text{arcsine}\sqrt{\frac{c_i}{c_i+d_i}}, \tag{5.1}$$

The arcsine difference takes the value of zero only if the values of both a_i and c_i are zero.

The variance estimate for an arcsine difference, for a two-arm trial is given by:

$$\frac{1}{4(a_i+b_i)} + \frac{1}{4(c_i+d_i)}. \tag{5.2}$$

This is not an exact estimate but has the advantage of producing a finite estimate even with zero values. A further method of calculating the variance is also presented, known as the analytical calculation of the variance, which has the effect of improving the approximation of the variance when there are few or no events. Using either of these methods of calculating the variance, the arcsine differences can be combined using an inverse variance method, to produce a pooled estimate and confidence interval.

The authors argue that this method compares favourably with the use of other metrics, such as the RD, RR and OR, and should be considered when performing a meta-analysis of studies where zero events occur. The arcsine difference avoids the need for arbitrary continuity corrections, and the bias they may introduce. Although the arcsine difference is not easy to interpret conceptually, the authors describe the metric using graphical methods to relate the arcsine difference to the RD for different baseline risks.

The arcsine difference is straightforward to calculate using standard inverse variance weightings for the individual studies. However, there is no simple formula

to convert the pooled estimate of the arcsine difference into an RD or other metric that is easily interpreted. In the light of this issue, the arcsine difference is not included in the case-studies in subsequent chapters. Furthermore, as discussed in Chapter 3, the RD is not usually the outcome metric of choice for data with sparse events.

Another alternative method for calculating a CI for a difference parameter, which does not require any continuity corrections, is demonstrated by Tian *et al.* (2009). This method derives a one-sided $100(1 - \alpha)\%$ CI (a, ∞) for each of k studies in a dataset. For any given value of η [presumed to be $1 - \alpha$], there are k one-sided η-level CIs for the parameter of interest (for example, the RD). The next step is to take all possible values of the parameter (Δ) and examine whether each one is the true value of Δ. If the selected value is the true value for Δ, then it should belong to η of the k CIs.

These authors provide a method to test the hypothesis that the interval (a, ∞) should include the selected value for Δ. This is achieved by setting up a variable, y_i, which takes the value 1 if Δ belongs to the observed (from the data) η interval from the ith study, and 0 if it does not. The selected value of Δ is included in (a, ∞) if:

$$t(\eta) = \sum_{i=1}^{n} w_i(y_i - \eta) \leq c. \tag{5.3}$$

In Equation 5.3, the value w_i is the study weight (for example, based on the sample size), and c is selected such that the P$(T(\eta) < c) \leq \alpha$, and

$$T(\eta) = \sum_{i=1}^{n} w_i(B_i - \eta). \tag{5.4}$$

The values of B_i where i refers to the ith study, are k independent Bernoulli random variables with a 'success' probability of η. Hence, the studies where Δ is included in the observed CI contribute positively to $T(\eta)$ according to their weight, whilst those where Δ is not included contribute negatively. This process is repeated for all possible values for Δ, based on the data. The lowest value of Δ that is found to be included in the CI becomes the value of a. This process can be repeated to discover the highest value of Δ that is found to be included

in the CI $(-\infty, b)$. Taking the CI (a, b), this is the $100(1-2\alpha)\%$ two-sided CI for the true value of Δ.

The authors provide a program to perform the calculations using the package R; this method is trialled in Chapter 7. Despite the concerns regarding the RD as a suitable parameter for adverse events datasets, there are also advantages to use of the RD.

5.2.7 Discussion of sparse data issues

The issue of how to include studies with sparse data in a meta-analysis is a complex one, and the various methods have been discussed above. Overall, it is reasonable to argue that an initial scrutiny of the dataset to examine the exact nature of the sparsity of the data is essential before embarking on any meta-analysis, and that multiple sensitivity analyses to evaluate the robustness of conclusions will be advisable in all circumstances.

As a final point to bring out of the study by Bradburn *et al.* (2007), the lack of power for all methods with low event rates was apparent. Even for the largest number of trials (19) in the meta-analyses, with balanced numbers across groups, the power for event rates of 1% was below 24% for an OR of 0.75, and below 78% for an OR of 0.5. At an event rate of 0.5%, the power was below 47% across all methods for an OR of 0.5. At the lowest event rate, 0.1%, the power even at an OR of 0.2 was highest at 30.9% (for the Peto method) and for smaller treatment effects was lower. For the smaller meta-analysis of only five balanced trials the powers were also very low for event rates of 1% and below.

This evidence indicates that for low event rates (which is a probable scenario for adverse events), the purpose of performing a meta-analysis is not so much to achieve a significant result, but more to provide an unbiased estimate of the pooled treatment effect, to obtain a CI, and hence to identify any signal from the data that may provide cause for concern, which can then be investigated further.

Issues related to sparseness of events data in the review of meta-analyses related to adverse or unintended events are discussed in Sections 6.3.2, 6.3.7 and 6.4.7.

5.3 Heterogeneous data sources

Regarding adverse events data, there are four main sources of data:

1. randomised controlled trials (and other forms of trial);
2. observational studies (cohort and case–control);
3. case series and case reports;
4. formal reporting systems for adverse events (e.g. the UK 'yellow card' system); and
5. 'anecdotal' information.

There are several potential ways to combine these forms of data. Comparative studies (trials and observational studies with at least two study groups) provide absolute risk estimates and relative risk estimates. Single-arm observational studies can also provide absolute risk estimates, while case series and reports can give an insight into frequencies and probability of event occurrence (if denominators of person numbers at risk are known), as can data from reporting systems such as the 'yellow card' (discussed in Chapter 2). Anecdotal evidence, such as clinical experience, cannot in itself provide numerical data but can provide qualitative information that may be used, for example, to form a Bayesian prior distribution.

The main concerns regarding combination of experimental (trial-based) data with comparative observational data are:

1. observational studies have more potential for confounding than trials;
2. observational studies may have longer follow-up than trials;
3. observational studies may have larger study populations than trials;
4. observational studies may not be representative of the entire population who have been exposed to an intervention, and hence not generalisable;
5. observational studies may have less accurate data regarding treatment regimes;
6. observational studies may be more prone to clinical heterogeneity; and
7. observational studies may be more susceptible to sources of bias in general.

The increased risks of bias and confounding, plus potential for inaccuracies and lack of detail regarding treatment regimes, indicate that observational studies would be less valid as data sources compared to trials. The arguments surrounding meta-analyis using randomised controlled trials (RCTs) and observational studies in combination are put forward by Borenstein *et al*. (2009), who conclude that these different study designs should be analysed separately, but can be combined in certain circumstances. Concerns regarding increased bias in observational studies compared to RCTs was the motivating factor in using hierarchical models with constraints to address this issue (Prevost *et al*. 2000; discussed further in Section 10.2).

However, the longer follow-up time of an observational study is an advantage when trying to discern adverse events that take a long time to develop and/or detect. Also, the larger study populations may result in more precise estimates and this may add to their weighting in a meta-analysis.

Hence, whilst RCTs are the 'gold standard' for effectiveness, observational studies may be relatively more valuable for adverse events, due to being based more in the 'real world'. As mentioned above, whilst an observational study may not encompass the full range of people exposed to an intervention, they may be capable of including a broader range of people than clinical trials. Furthermore, the environmental conditions in which an intervention is used may be more realistic than the more controlled conditions of a trial.

When combining experimental and observational data there are several options, including:

1. combine the study types on an equal basis;
2. as above, with sensitivity analyses by study type;
3. weight the studies in some way according to study design or other factors e.g. date of publication;
4. use Bayesian methods to place differential prior distributions on different study designs;
5. use Bayesian methods to explicitly model biases, which can then be adjusted for, within observational studies;

6. use Bayesian methods to represent the parameter of interest according to one study type in terms of a logical function of another study type;
7. use the observational studies as external evidence to provide prior estimates for parameters in a Bayesian model, with the trial studies as the main dataset for the analysis.

The third option is problematic, as it appears to be on a par with weighting by study quality; any weighting system may be arbitrary and subject to bias. However, empirical data could be used to derive the weights for the studies. A sensitivity analysis by study design would allow any differences in results between the two study types to be scrutinised, possibly returning to the original studies to determine potential clinical reasons for any discrepancy.

If different study designs are included, it is important to assess the results of the meta-analysis for heterogeneity and to fully investigate any heterogeneity found, for example by subgroup analysis (at the study level) or by meta-regression using study-level covariates if the required data are available. (Heterogeneity is discussed further in Section 3.9, meta-regression is considered in Section 3.7 and subgroup analysis is discussed in Section 5.5.)

Comparison between FE and RE methods in this situation would be advisable as a means of identifying whether any statistical heterogeneity is having an impact on the results. An RE method would be the preferred choice if this is the case, as trials and observational studies may be estimating differing underlying effects, and the RE method would yield more conservative results. However, if data are sparse, as has been discussed above in Section 5.2, achieving significance, or even a plausible CI may be a lower priority than producing an unbiased estimate of the pooled effect, so the choice of meta-analysis method may be driven by this consideration above statistical power or desire to incorporate between-studies heterogeneity.

The issue of how to model sources of bias, both internal (with relation to the study design) and external (with relation to generalisability), has been addressed by Spiegelhalter & Best (2003) and the work continued by Turner *et al*. (2009). The former authors considered how to derive an RE model for a parameter of interest, whilst taking into account sources of external and internal bias, and by assessing the degrees of bias, a quantitative evaluation of study quality can be derived (effectively a quality weight for each study). This quality weight can be

considered as the proportion of between-studies variability unrelated to internal bias (within each study).

The work of Spiegelhalter & Best (2003) is extended by Turner *et al*. (2009). The most prominent issue developed is the difficulty of disentangling sources of bias, and the importance of clinical input where necessary. Assessment of study methodology is the key first step in adjusting for bias using statistical methods. These authors then put forward ways to adjust for biases, differentiating between those that are independent of the underlying treatment effect (additive biases), and those that are proportional to the treatment effect. If all internal and external biases are adequately accounted for, there would be no remaining between-studies heterogeneity. Methods are presented to quantify both additive and proportional biases, to incorporate these into the overall meta-analysis.

Putting this issue into the context of adverse events, sources of internal and external bias can be adjusted for, allowing all sources of evidence to be incorporated into the dataset, with a specific system of weighting the studies according to defined sources of bias. However, the model requires the use of an RE model to incorporate external validity, which may not be the most appropriate approach for adverse events. With regard to adverse events it is likely that the most important source of external bias is in the study population, and it would be straightforward to exclude all studies that do not match the study group of interest.

In a meta-analysis where the outcome of interest is an unintended but beneficial one (Bonovas *et al*. 2005), the authors performed separate meta-analyses for observational studies and randomised trials. They then used a test of interaction (Altman & Bland 2003) between estimate of statin use on risk of prostate cancer and randomised controlled trials/observational studies to determine if there is any evidence that the results for the two study designs are different. They also performed an overall meta-analysis of all studies combined.

If Bayesian methods are preferred, constraints could be used on the prior distributions for different study types, for example to address potential differences in the degree of bias between the study types, as proposed by Prevost *et al*. (2000). This approach is utilised in the context of a hierarchical model in Chapter 10. Alternatively, different prior distributions could be placed independently on the

two types of study within the model, for example with the observational studies providing data to inform the prior distributions for the experimental studies.

Methods to include both comparative and non-comparative studies have been developed by Begg & Pilote (1991), who used the context of a clinical situation where comparative studies for two treatments were available, but also uncontrolled (single-arm) studies for each of the two treatments. The three study types were combined in ways that varied the weighting of the uncontrolled studies according to the between-studies variation. When this value is large, the uncontrolled studies receive low weighting in the estimate of the pooled treatment effect. Conversely, when the between-studies variance is 0, then the uncontrolled studies are weighted on an equal par with the comparative studies.

This approach is developed further by Li & Begg (1994), who point out that the previous methods required the assumption of a normal distribution for the baseline random effects and the treatment effects of each study. Also, it is assumed that the uncontrolled studies are not biased. In the light of these assumptions, a more general approach is developed, which does not require any assumptions regarding distributions.

With regard to adverse events meta-analyses, these methods would be of value when incorporating observational studies that have no control group, as they provide a quantitative means of incorporating the uncontrolled studies. They may also be of use for trials with no control group, for example, trials using a surgical intervention. However, the inclusion of the uncontrolled studies is in fact a function of the between-studies variance, and the uncontrolled studies are disregarded when the value of this statistic is 0. Given that for adverse events, the main aim is to detect any signal from the data and to generate a point estimate, it is hard to justify exclusion of uncontrolled studies simply due to low between-studies variance.

The case series, case reports and anecdotal reports (as well as non-comparative observational studies) can also be used as qualitative data, which is particularly appropriate when only the relative estimates are to be combined quantitatively. Such data can be incorporated into a narrative review, but the question of how to combine qualitative with quantitative data within a quantitative meta-analysis is more challenging. One way to bring the qualitative data to bear on the quantitative analysis is to incorporate such data into the prior distributions

for Bayesian analysis, effectively using such data as an 'expert opinion'. For example, the qualitative data (supporting the existence of an adverse event) could form an 'enthusiastic' prior distribution on the point estimate of the pooled estimate. The interpretation of such data is, however, a clinical issue and would ideally be performed by a clinical expert in the appropriate area, e.g. a pharmacologist or geneticist. With this information available, a meta-analysis can be performed that incorporates both quantitative and qualitative data.

An example of using both qualitative research and the prior beliefs of individual contributors has been provided by Roberts *et al*. (2002). However, in this case, the outcome referred to identifying the factors used by parents as reasons for their behaviour in relation to immunisation of their children. Therefore, qualitative research was likely to be of greater importance than for a physical outcome. Reviewers used both their subjective opinions (before evaluating the qualitative data), and the results of the literature. These were combined to yield a prior probability for each factor being of importance, and these prior probabilities were used in a Bayesian analysis. This approach is likely to be appropriate only for unintended outcomes that are subjective in nature.

Heterogeneous data sources in the systematic review of adverse and unintended events meta-analyses are discussed in Sections 6.3.1 and 6.4.1.

5.4 Multiple outcomes

When focusing on unintended and adverse events, it is probable that there will be several possible events of interest for each trial. Almost all treatments have the potential to produce multiple unintended outcomes and within a study they may all be reported. Indeed, failure to report on all unintended outcomes would constitute reporting bias. Another issue is the fact that some events may have been unforeseen and then had to be incorporated into the study outcomes, effectively turning the study into a form of 'fishing expedition'.

For an individual study there are several methods of adapting the analysis to account for multiple testing; some of these are outlined by Pocock (1997). Methods include use of the Bonferroni correction, defining a single primary outcome, amalgamating all test statistics into a global test statistic, and combining outcomes. Definition of a single primary outcome may be beneficial when the

outcomes are infrequent as is often the case for adverse events, but may be clinically undesirable.

Furthermore, combination of certain outcomes may be appropriate clinically when the mechanism of action of the drug or intervention is the same or similar, or for outcomes that are pathologically similar, but inappropriate in other circumstances. This is another area where clinical input is required prior to any statistical analysis. This approach is similar in concept to combining evidence from all outcomes into a single test statistic. It has been argued that this approach may enhance statistical power, discussing one method based on calculation of a standardised normal test score for treatment differences, and then combining these according to a weighting system for each outcome (Pocock 1997).

The problem of limited study size [and hence power], as discussed by Pocock (1997), is enhanced with the problem of small numbers of events; correction for multiple testing using a Bonferroni correction will play against the limited power by raising the threshold for significance, and indeed in many scenarios for adverse events data significance is secondary to receiving a 'signal' from the data. Indeed, Pocock refers to adverse events scenarios as "multiple outcomes gone crazy", and requires the lack of prespecified hypotheses which can lead to 'data dredging'. However, to contradict Pocock, the ratio of false positive to true positives is unlikely to be high due to low power for most studies for adverse events. A predefined primary outcome may not be feasible for clinical reasons, if an adverse event is not identified until after a study has commenced, and as mentioned by Pocock, reporting of all adverse events is required by regulations for clinical trials.

The difficulties presented by multiple outcomes for a single study are complex, but when placed in the context of multiple outcomes across multiple studies the problems are increased. For example, using the single global test statistic approach, the weighting systems within studies may lead to bias if not carefully considered.

The use of a random effects meta-regression model for the meta-analysis of multiple outcomes has been suggested (Berkey *et al*. 1998). The advantage of this is that multiple endpoints are usually correlated, and this should be accounted for in the model.

Despite the importance of this issue for adverse events meta-analysis, where it is possible that there will be many outcomes for each intervention, there is no further consideration in the case-studies presented. In Chapter 6, the review of previous meta-analyses in adverse and unintended events addresses issues of multiple outcomes in Sections 6.3.2 and 6.4.2.

5.5 Subgroup analysis

Subgroup analysis in this context is intended to refer to subgroups of individual patients within a single study, for example based on demographic factors such as age and sex, genetic factors, or on treatment factors such as different drug regimes. An alternative way to think of subgroup analysis is by subgroups of studies, for example, based on geographical area, date of study, study design or source of funding, effectively where all patients in the study have the same value for the covariate of interest. In some cases these two concepts may coincide, for example certain studies may have participants of only sex or in one age group.

It is usually easier to address issues based on subgroups at the study level, for example using methods discussed in Section 5.3 regarding heterogeneous data sources. It is more difficult to develop methods for use in subgroup analysis at the level of the individual patient, especially when information regarding the subgroups of interest is lacking (this topic is discussed more fully in Section 5.6). This lack of data could be in two parameters, firstly in terms of lack of data about the defining subgroup factor (for example, lack of data on age or sex of participants) and in terms of lack of outcome data for those subgroups (for example, presenting data for the total cohort but not broken down by age and sex).

Subgroup analysis may become an important part of a meta-analysis of adverse events data for different reasons. Firstly, if there is statistical heterogeneity present, then it would be advisable to investigate this further, and along with meta-regression, subgroup analysis would be an appropriate method. Also, there may be specific demographic subgroups within primary studies, such as age groups, that may be at differing risks of an adverse event. There may also be genetic subgroups at increased risk of an adverse event. Thus, there may be reasons why a meta-analysis of subgroups is required, which may be known

prior to commencing an overall meta-analysis, or becoming apparent whilst performing such a meta-analysis.

Therefore, the issue of subgroups is particularly pertinent to the analysis of adverse events data. Given that adverse events are often infrequent, this can mean that a study that is adequately powered for efficacy within subgroups may be underpowered for adverse events; hence the use of meta-analysis can be very valuable to bring together multiple studies and increase power. However, bearing in mind the low power of meta-analyses with low event rates (Section 5.2.7), such increases in power may not be substantial. Lack of power is one of the key disadvantages of subgroup analysis.

Another reason for subgroup analysis that is specific to adverse events data is that an intervention (usually a drug in this context) may be used for different indications, and it may also be the case that the adverse events profile will vary according to the indication. This highlights the desirability of combining primary studies for different indications (i.e. that will have efficacy data on different outcomes) when the same intervention is performed across studies, with the aim of combining data for common adverse events. Methods to investigate this phenomenon need not be limited to subgroup analysis, but it is one area where subgroup analysis may be usefully employed if a single study includes subjects receiving an intervention for different clinical conditions, or where different studies use the same intervention for different purposes.

Referring to individual studies, Brookes *et al.* (2004) point out that a separate analysis by subgroup leads to the risk of multiple testing and finding a positive result by chance. This approach also reduces power to detect a true treatment effect by subdividing the whole dataset into subsets. A test for treatment–subgroup interaction was considered the appropriate way to investigate whether the intervention differs across subgroups, using linear or logistic regression as appropriate. At the meta-analysis level, IPD would be required for patient-level covariates, but regression methods could be used for study-level covariates. As an alternative to subgroup analysis, a treatment–subgroup interaction could be used.

As IPD greatly facilitates subgroup analysis, with the ability to incorporate patient-level covariates, this method would be highly recommended in this area. However, IPD are often unavailable for all or some of the studies. Another

problem is that studies that report only aggregate data may not report the aggregate values for all subgroups of interest. Meta-regression by study-level demographic characteristics leads to the problem of the use of ecological data, which may not correspond to the true situation at the individual level.

The aspects of the systematic review of meta-analyses in adverse and unintended events with regard to subgroup analysis are to be found in Sections 6.3.2, and 6.4.2

5.6 Individual participant data meta-analysis

Individual participant (or patient) data (IPD) meta-analysis is regarded as being the 'gold standard' approach, superior in its resulting estimates to the use of summary data – some of the reasons for this are discussed below. In the ideal scenario, where IPD are available for all primary studies, there are two main approaches to meta-analysis of such data.

The first approach is simply to use the IPD to calculate a summary statistic for each individual study, which can then be used for a meta-analysis using standard techniques – hence, this method is known as the 'two-stage' method (Simmonds *et al*. 2005). This method has the advantage over using aggregate data, where the summary statistic for each study is presented, but not the 2×2 table, in that the choice of summary statistic is within the power of the performer of the meta-analysis. Also, if covariates are provided at the individual participant level, then the choice of which covariates to adjust for is available.

Another advantage of IPD meta-analysis is that it is more readily applicable to time-to-event data (survival data). Such datasets do not facilitate the break-down into simple 2×2 tables, which are the mainstay of standard meta-analysis methods. The exception to this is an extension of the Peto method, described in Section 3.3.5.

The second method available for IPD meta-analysis is known as the 'one-stage' method, whereby the data from all studies are combined into one dataset (Simmonds *et al*. 2005). This dataset can be analysed as if the data were all from the same study, or analysed using stratification by study, with the individual study treated as a covariate within the model . This method avoids the problem of

Simpson's Paradox, that besets a marginal analysis (although as has been mentioned, this issue may be less important for adverse events meta-analyses than for other forms of data).

If IPD are available, then covariates at the individual level (e.g. age), rather than the study level (e.g. mean age), can be included in a meta-analysis, as well as the study covariate. This allows a wider range of covariates to be included, since some covariates are relevant at the individual level rather than the study level. Also, the use of IPD allows 'borrowing strength' to occur across studies (Higgins & Whitehead 1996; Borenstein *et al*. 2009), whereby assumptions can be made that certain statistics for each study are the same. An illustrating example from Borenstein *et al*. (2009) is that of using age as a patient-level covariate (mean age across studies will be similar and so cannot be used) in a meta-analysis where weight loss is the outcome, and is related to age. Information could be borrowed across studies, increasing the power, for example by assuming that the standard deviation for weight loss is the same across all studies. Both FE and RE models can be implemented using IPD, but RE models are more difficult to apply in practice (Borenstein *et al*. 2009).

The IPD approach is particularly suitable when aiming to investigate patient-level characteristics with regard to the outcome of interest (Lambert *et al*. 2002); as such, it may be especially relevant to adverse events data. These authors used a simulation study to demonstrate that in many situations an IPD meta-analysis has greater power than its aggregate data counterpart, with the aim of detecting differences in treatment effect between high- and low-risk patients. Only when the number of subjects in each study was large (1000), with 20 studies in total, and the treatment effect also large, did the aggregate data analyses approach the power of the IPD analyses. In the meta-analysis of adverse event data, a large effect size is usually ruled out (it is to be hoped that drugs with potentially widespread adverse events would be excluded at an early stage of development). Hence, aggregate data meta-analysis would lack sufficient power to detect a treatment effect. For example, even with 20 studies of 1000 subjects each, with a small effect size, the meta-regression analysis detected a significant difference between the risk groups in only 15.6% of simulations compared to 90.5% of simulations for the IPD analysis. This study reinforces the advantages of IPD for adverse event data; indeed the benefits of IPD would seem to outweigh the disadvantages of time and cost involved in obtaining the data.

Chapter 8 discusses a meta-analysis of randomised controlled trials where IPD were available for all studies, with a time-to-event outcome. However, this situation is not common in practice, as IPD are not usually presented in published references. In many cases the data may not be available if the publication is some years old, or the authors may be unwilling to divulge the data for further analysis. In this case, if only a subset of studies have IPD, with the rest having only aggregate data, it is necessary to develop further methods to combine these forms of data.

Accepting the argument that IPD is superior to aggregate data, it is then necessary to develop ways to combine IPD and aggregate data in such a way that the additional information of IPD is retained and made best use of, while incorporating the aggregate data.

In a review of meta-analyses that combined aggregate data and IPD, Riley *et al*. (2007), found four main approaches to this methodology. By far the most common was the two-stage method, in which the IPD is used to produce a summary measure that is then combined with the aggregate data from the other studies. This approach is easy to apply but has the disadvantage of losing the patient-level data, which is one of the main benefits of IPD. It is therefore best suited to situations where only study-level covariates are of interest.

Another method involves the partial reconstruction of IPD from aggregate data; this method applies when the outcome is binary and the data are presented in a 2×2 table from which a series of individual datapoints categorised as 0 or 1 for the outcome of interest can be derived. This reconstructed data can then be combined in an overall meta-analysis with the provided IPD. It is, however, difficult to reconstruct patient-level covariates, therefore this type of analysis is usually most appropriately applied to situations where study-level covariates are of interest. There are also various types of bias that may arise when using reconstructed data as opposed to having available the original study dataset. For example, some participants may be excluded inappropriately, and importantly for adverse events, results may be selectively reported in the final publication, compared to the collected data. Also, it is often difficult to reconstruct patient-level covariates from aggregate data.

Another option is to use a multi-level modelling approach where the lowest level is an observation from an individual participant and the highest level is study-

level aggregate data. By including a dummy variable in the model to distinguish between aggregate data and IPD, the analysis can include both aggregate data and IPD for study-level effects, and include only IPD for patient-level covariates.

Finally, a Bayesian hierarchical model, known as Hierarchical Related Regression (HRR) has been proposed (Riley *et al*. 2007, citing earlier authors). This method uses Markov Chain Monte Carlo (MCMC) methods to simultaneously estimate two related regression models, one for IPD only and one for aggregate data only. These models have common parameters which are being estimated from both the IPD and aggregate data. By allowing both types of data to influence the estimate of the parameter, the problems associated with ecological bias of aggregate data and low power related to IPD are reduced.

The use of IPD and aggregate data has been extended to account for issues such as clustering by Sutton *et al*. (2008). These authors also investigate methods to use covariates when combining studies with IPD and aggregate data, where the latter has covariate data in terms of the proportion of patients within a study who have a certain characteristic. The aggregate data studies can be used in a meta-regression to estimate the slope of the regression line for the characteristic of interest. This slope is assumed to be equivalent to the coefficient for the interaction term between treatment and characteristic in the IPD analysis. Hence, these two data types can be used to estimate the effect of the covariate of interest.

An assessment and further development of methods for combining aggregate data and IPD has been conducted by Riley *et al*. (2008), using continuous outcomes. Both one- and two-step methods are considered, including the method involving the use of a dummy variable as described above, which allows the aggregate data and IPD to contribute to the overall pooled treatment effect, while the IPD studies also yield a study-level effect. These methods are used to investigate patient-level covariates, including within-trial and across-trials treatment–covariate interactions. Given the importance of identifying at-risk groups in the analysis of adverse events data, such methods have a strong potential for further development in this area. With regard to adverse events data, however, the majority of outcomes tend to be binary rather than continuous.

Focusing on time-to-event studies, aggregate data and IPD can be combined using a hazard ratio (HR) as the outcome measure. IPD analysis is particularly

appropriate when reporting time-to-event data. For example, time-to-event data may be reported using different outcome metrics in different publications; common outcome metrics include the HR, or log HR, results of the log rank test, and median time-to-event. Use of IPD can help to overcome the problem of diverse reporting methods across datasets.

Investigation of heterogeneity between studies is facilitated by the use of IPD. For example, Tudur Smith *et al*. (2005b) compared IPD and aggregate analysis of time-to-event data. To investigate heterogeneity they used a Cox proportional hazards regression model stratified by trial for the IPD, and for the aggregate data a meta-regression using the log HR as the outcome metric. The authors discovered that the aggregate data and IPD models differed in their identification of potentially significant variables and argue that the IPD method is 'safer' in its application. Chapter 8 discusses a time-to-event meta-analysis of IPD in detail.

In the context of adverse events data, the use of IPD can provide much-needed additional power (e.g. through borrowing strength across studies) to a meta-analysis and facilitates the inclusion of patient-level covariates for subgroup analysis. As many of the studies included in a meta-analysis of adverse events are often clinical trials it is to be hoped that IPD will be increasingly available.

Only two publications including IPD were retrieved for inclusion in the systematic review of meta-analyses with adverse and unintended outcomes; IPD analysis within the context of this review is included in Sections 6.3.2, 6.3.5, 6.4.2 and 6.4.5.

5.7 Dose–response data

From a clinical perspective, one of the major areas of interest when considering adverse events outcomes in relation to drug therapy is the concept of dose–response. As pointed out by DuMouchel (1995), there are two fundamental questions, namely, whether a dose–response relationship exists, and if so, what it is. It would be very interesting to evaluate a particular 'threshold dose' below which a certain adverse event would be very rare, and then to gain an understanding of how risk of an adverse event increased along with dose. Associated with this idea is also the issue of length of exposure (which results in increased cumulative dose as time progresses) as well as the dosage of the drug as used.

One of the major areas of difficulty in performing a dose–response meta-analysis is the fact that data on differing dosages may be derived from multiple studies (i.e. dose as a study-level covariate), rather than from the same study where the different dosage regimes are being compared over the same study population (i.e. dose as an individual-level covariate). For those studies where different doses are being compared, it would be very valuable to have IPD, especially if results are not presented for each dose regime separately. In such an instance the IPD methods discussed in Section 5.6 would be applicable.

Dose may be considered as a continuous covariate or a discrete covariate with multiple levels such as 'low', 'recommended' or 'high' dose (which occurs in the case-study explored in Chapters 9 and 10.). Evidently, if exact doses are provided, these can be categorised into discrete bands, according to clinical usage.

When considering the possibility of a causative relationship between an intervention and an adverse event, dose–response (or biological gradient) is one of the criteria for causality set out by Hill (1965), so the establishment of a dose–response relationship is useful clinically.

This fact is considered in a prominent study in this field, published by Tweedie & Mengersen in 1995. These authors discuss several meta-analysis methods that may be appropriate. The establishment of a dose–response relationship is a strong step in proving a causal association and once established it can be used to predict the level of risk for individuals at different levels of exposure. If data from different studies are to be pooled, it is important to first obtain a dose–response measure from each study (consistent across test statistic and quantity), and then to pool these dose–response relationships from individual studies using meta-analysis. Effectively, each study can be thought of as providing its own estimate of the dose-response curve.

Interestingly, Tweedie & Mengersen (1995) use as their example a dose–response relationship between lung cancer and environmental tobacco smoke, rather than a dose–response model for a drug and treatment effect, demonstrating that a dose–response relationship may occur in a variety of contexts. The use of observational data further blurs the bounds of a dose–response model. In the context of adverse drug reactions, epidemiological studies may have the advantage of

longer follow-up time and possibly a greater range of doses, but may also be disadvantaged by less detailed information regarding dosages.

Three methods for analysis of a dose–response relationship across different studies are discussed by Tweedie & Mengersen (1995), the first using a non-parametric test for equality of response across dosage levels, the second using an exponential model and the third using a linear model. The latter two then use a test of significance for the regression parameter. They then put forward three meta-analysis methods for the combination of the data derived from individual studies. The first is a non-parametric test for equality of responses based on combining the non-parametric test statistics derived from the individual studies. The second is a random effects model allowing for between- and within-study variation in dose response and the third is a fixed effect model with the assumption that there is no heterogeneity of dose response across studies. It goes without saying that data of a very high quality are required in this case, such as data on the exact treatment regimes, and if multiple treatment regimes are applied within the same study then they must all report outcome parameters in an appropriate format for meta-analysis at all levels.

The debate surrounding dose–response issues in meta-analysis is continued by DuMouchel (1995), who considers aspects such as whether a dose of zero should be included in a dose–response model, and measuring residual error, which appears to be more problematic when studies have fewer dose groups. Also discussed are issues regarding the random effects model, which are taken in the context of either an empirical Bayes or fully Bayesian model. These models differ in their evaluations if the number of studies is small.

Dose of drug is explored using Bayesian hierarchical models and mixed treatment comparisons (MTCs) in Chapters 9 and 10. In this case-study, IPD were not available. In situations that have IPD available for all studies, the logical approach would be to use meta-regression with dose as a continuous or discrete covariate, with the control being fixed at a dose of 0. This model would allow the addition of other covariates, such as age, to investigate any interactions. The methods of Tweedie & Mengersen (1995) are not employed in this thesis, as the nature of the dose–response data is not well-suited to the regression methods, as it is not continuous in the examples used in Chapters 9 and 10, and the non-parametric methods provide only a test for equality of responses rather than an estimate of treatment effect.

Dose–response issues are explored in previous meta-analyses of adverse and unintended events in Sections 6.3.2 and 6.4.2.

5.8 Class effects

This section is restricted to adverse events that are specific to drug therapy, and in particular to comparisons within and between specific classes of drug. Drugs are often classed according to their activity (such as beta-blockers or calcium channel inhibitors) and in clinical practice it is often the case that a specific condition is treated with drugs from more than one class.

It is therefore necessary to dissociate effects across drug classes to see if one class of drug has a different adverse events profile for the same indicating condition. Also, even within a class there may be some class members with a different pharmacological activity profile. It may be very important to disentangle differences in adverse events between drugs within the same class but with slightly different pharmacological properties, with the aim of selecting the drug with the best balance between efficacy and adverse events profiles.

These processes are effectively forms of subgroup analysis (discussed in Section 5.5) and risk–benefit analysis (to be discussed in Section 5.11 and more fully in Chapter 11).

A meta-analysis of efficacy of a variety of migraine treatments, some involving drugs and others involving non-drug therapy, is described by Dominici *et al*. (1999). Their approach hinges on the relative ranking of treatments both within classes and overall. This is achieved by performing indirect comparisons among treatments if they are not tested against each other in the same trial, and by fitting an hierarchical Bayesian grouped random-effects model. The authors used clinical data that suggested that efficacy was more similar within treatment classes than between classes, and used this information to deduce that when 'borrowing strength' within a model it was more important to borrow more heavily between treatments of the same class (Higgins & Whitehead 1996).

It would be feasible to use similar methods to those of Dominici *et al*. (1999) to investigate for adverse events across a range of drug classes, especially if there were existing clinical or pharmacological data regarding adverse events that could be used to inform the model.

The case-study set out in Chapter 9 involves an example whereby multiple drugs from the same overall class, but with one member that has a different pharmacological profile, are compared against each other using MTC analysis. The use of MTCs may be a particularly suitable approach for this problem, as it allows both direct (within the same study) and indirect (across separate studies) comparisons.

The systematic review of previous meta-analyses of adverse and unintended outcomes refers to class effects in Sections 6.3.2 and 6.4.2.

5.9 Time-course effects

Time-course effects in this instance refers to the making of repeated measurements of adverse events data at varying times following the commencement of treatment. This issue is particularly pertinent in the area of adverse events because some events may take a long time to develop (due to the pathological processes involved) and/or to become symptomatic and diagnosable. Also, some adverse events may require a greater cumulative dose and hence will only occur after a longer time period. Therefore, this issue may tie in with dose–response considerations.

This area is an issue for two reasons. Firstly, different studies will have different lengths of follow-up and/or report outcomes at different times, and therefore methods are required to account for this feature of the studies. Secondly, within one study, each individual may have the outcome (assumed to be non-fatal) reported at different times (depending on the nature of the outcome). As multiple observations are being made on the same individual at different times, this leads naturally to an hierarchical model, if IPD are available, or if the same outcomes were reported at corresponding times for different studies that have only aggregate data.

If IPD are available for all studies, but without a specific date of event for each patient, then time-course effects can be modelled within separate time-frames, where an event is the occurrence of an adverse event within a specific time-frame. If exact dates of events are known then a more accurate time-to-event analysis can be performed (this approach is used in the case–study in Chapter 8).

This approach was taken in the context of time-to-healing in a scenario primarily focusing on MTCs (Lu *et al*. 2007). A more detailed discussion of MTCs in general is included in Chapter 9. This study was complicated by the need to combine multiple follow-up times with direct and indirect comparisons. The meta-analysis was required to 'borrow strength' for effects across time periods and also to combine direct and indirect evidence on treatment effects within each time period. The authors developed a range of models using several Bayesian hierarchical models. To describe three of the models, these included a model using log HRs with homogeneous variance; the time aspect was addressed by allowing the baseline hazard to vary over time with the addition of a fixed term at each follow-up time. The second model involved a mixed effects baseline with a fixed effect for trial and time, and a trial-by-time interaction. The third model is a random walk model, which includes the assumption that for each trial the baseline effect during a particular time period should be closer for adjacent time periods than for non-adjacent time periods. The expectation of the log hazard for each time period is the log hazard at the previous time period. Assuming a normal distribution, the variance term for the later time period is estimated from the data, hence there is more variability for the earlier time period compared to the later one.

If IPD are available for some studies (with exact times known), but for other studies only IPD with date of events within a specific time-frame, or aggregate data with events within a specific time-frame, then there would be difficulties in combining such data. The easiest route would be to use specific time-frames and not use the time-to-event analysis, by using the IPD with individual times to separate events into time period of occurrence. Separate analyses could then be performed for different times (this approach is used in Chapter 8). Alternatively, duration of study could be used as study-level covariate (this method is used in the harm–benefit analysis in Chapter 10).

The occurrence of time-course effects in the meta-analyses of adverse and unintended outcomes is discussed in Sections 6.3.2 and 6.4.2.

5.10 Reporting bias

The wider issue of publication bias, whereby a study is less likely to be published if it shows a non-significant treatment effect, or alternatively an effect that is not desired by the study's organisers, is a large field in its own right. In the frame of adverse events, very few trials are performed with the aim of looking for adverse events, hence a study is more likely to be published based on the significance or otherwise of the treatment effect.

A more pertinent phenomenon for adverse events in a clinical trials context would be reporting bias, whereby adverse events are selectively not reported (whether intentionally or simply because it may not be considered of importance if there is no indication that an adverse event is causally associated with an intervention). It is, however, very difficult to formally assess or adjust for reporting bias. Observational studies are more likely to be specifically aimed at looking for adverse events, and hence fall foul of the risk of publication bias if no significant effect is found.

For these reasons, the associated issues of reporting and publication bias are not considered further in this work in order to concentrate on other areas.

Reporting bias is considered further in the systematic review of adverse and unintended events – see Sections 6.3.2, 6.3.3 and 6.4.3.

5.11 Evidence synthesis of risks and benefits

The analysis of an adverse events profile is clinically most useful when combined with an assessment of the efficacy of the intervention. Taking the efficacy and adverse events data as separate entities initially, they can be combined through a synthesised narrative of efficacy balanced against harms, which can be useful in developing clinical guidelines.

The next step is to combine quantitatively the risks and benefits in a single model, preferably one that can combine different study designs and can identify high-risk subgroups. By such a method it is to be hoped that any subgroups for whom a risk of an adverse event may be disproportional to the potential

benefit of the treatment can be identified and possibly counselled against the treatment where appropriate.

It is evident that there is still much work to be done in this area, which has the potential to be very informative clinically. The whole of Chapter 11 is devoted to a harm–benefit analysis, and the methodology is discussed more fully prior to implementation of the model within the case–study.

5.12 Summary

This chapter has described ten potential areas that may prove challenging should they occur when performing a meta-analysis where the primary outcome is an unintended or adverse event. The selected areas of difficulty are by no means the only challenges that may be encountered in this field and indeed many are not specific to adverse events data; datasets where sparse events are common occurrences, for example, may occur in other contexts. Some of these issues are explored in detail in later chapters, for example, issues of sparse data in Chapter 7, the use of IPD meta-analysis and time-course issues in Chapter 8, the use of MTCs and dose–response issues in Chapters 9 and 10, and a harm–benefit analysis in Chapter 11. It has not been possible to cover all of the areas discussed in this chapter, but they are included to demonstrate the range of potential problems and areas where the methodology is undeveloped.

6

Review of methods previously used in meta-analyses of adverse events data

6.1 Introduction

6.1.1 Background and aim of study

Having identified certain areas of meta-analysis methodology that may be problematic when applied to adverse events datasets (as discussed in Chapter 5), it was then of interest to investigate whether these areas had been addressed in previous meta-analyses, and in more general terms to discover the approaches and methods that were used in such meta-analyses.

The aim of this systematic review was to summarise the statistical methods used by published meta-analyses where the main outcome was an unintended (usually adverse) event; additionally other, non-statistical, information about each study, such as the nature of the intervention or source of sponsorship, was collated.

A review of systematic reviews of adverse drug reactions has been performed (Cornelius *et al*. 2009), and it is useful to compare the review by Cornelius *et al*. (2009) with the review presented in this chapter. Some of the areas encompassed by Cornelius *et al*. (2009) are included within this current review, but the scope was narrower, both in terms of the clinical remit (only adverse drug reac-

tions were included, as opposed to adverse events due to *any* clinical intervention), and in terms of the temporal range of publication (only reviews published in 2006 were selected, 43 reviews in total). The review by Cornelius *et al*. (2009) also examined statistical methodology used for meta-analyses within the systematic reviews being investigated. Aspects included were the model used, whether fixed or random effects were used, and and whether sparse events (with counts of zero) were present. Non-statistical elements investigated included the search strategies employed by the systematic reviews, use of quality assessment and source of funding. The findings of this review are discussed in greater detail in Section 6.5.10.

This current review focuses in more detail on the statistical methodologies of meta-analyses of adverse events, while also recognising that non-statistical issues are also of interest when describing these studies.

6.2 Methods

6.2.1 Reference retrieval

The dataset of meta-analyses for this survey was identified using a database of studies collected previously (Golder *et al*. 2006b). The aim of this previous study was to develop and evaluate search strategies for the retrieval of systematic reviews, which may or may not include a meta-analysis, where the primary outcome was an adverse event resulting from a clinical intervention. The databases searched in the previous study were the Database of Abstracts of Reviews of Effects (DARE) and the Cochrane Database of Systematic Reviews (CDSR). This study highlighted the difficulties of electronic database searching for systematic reviews of adverse events data. (Specific problems included lack of standardised terminology, poor indexing and variation in the interfaces of such databases. Handsearching revealed that systematic reviews had been missed in the electronic searches in both DARE and CDSR. Details of the search stratagy have been described.)

These search strategies yielded a total of 257 publications (246 from DARE plus 11 Cochrane reviews). References were published between 1994 and 2006.

The initial searches were updated, yielding a further 20 systematic reviews on adverse events (Golder *et al.* 2008).

This current review of statistical methods used for meta-analyses of adverse events data builds on the foundation of work in the information retrieval aspects of adverse events reviews (Golder *et al.* 2006b; 2008) by investigating the statistical methods that have been previously used in meta-analysis of adverse events data.

For ease of reference, the selected references for the review are referred to by number (from 1 to 166, prefaced by S) whereas other references are referred to by authors and year of publication. A full list of references included in the review is provided in Appendix A. To avoid lengthy lists of references in the text, where not provided as an adjunct to a specific result, in selected cases, a list of references supporting that result is provided in Appendix B.

6.2.2 Reference selection

From the references described above, it was necessary to select relevant studies according to the required criteria for this review of statistical methods. The criteria included:

1. some form of quantitative synthesis (or test for heterogeneity with intention to perform a quantitative synthesis if appropriate) must be performed using more than one observed estimate of effect;
2. the study group of interest must have received some form of clinical intervention with *intended* or *potential* therapeutic effect; and
3. the full study report must be available in English.

A quantitative data synthesis may take the form of a pooled estimate, a confidence interval, quoting a p-value only, or performing a meta-regression. Studies entailing only qualitative evidence synthesis, although having an essential role to play in the assessment of adverse events, are not included in this review, which is exclusively aimed at statistical synthesis methods. Meta-analyses of unintended or adverse reactions to non-interventional activities, for example recreational drug use, are excluded.

Table 6.1: Number of studies by year of publication.

Year of publication	No. studies	% studies[a]	Published in journal	Cochrane review	Other publication type
1994	4	2.4	4	0	0
1995	8	4.8	8	0	0
1996	9	5.4	9	0	0
1997	15	9.0	15	0	0
1998	16	9.6	16	0	0
1999	20	12.0	19	1	0
2000	11	6.6	11	0	0
2001	22	13.2	22	0	0
2002	21	12.6	17	2	2
2003	22	13.2	20	2	0
2004	13	7.8	10	3	0
2005	4	2.4	0	4	0
2006	2	1.2	0	2	0

[a]Out of 166 studies.

By the above criteria, studies were excluded due to being published in other languages than English, having no quantitative analysis of more than one effect estimate, for example systematic reviews or meta-analyses that present forest plots of only one study, or not relating to the effects of an intervention for clinical purposes. In total, 166 studies fulfilled all criteria and were included in the systematic review. Of these, 14 were Cochrane reviews (S1; S16; S20; S47; S78; S86; S91; S94; S131; S133; S136; S137; S148; S166), the others were published in a wide variety of medical journals or were reviews published by health agencies (S70; S103). Table 6.1 shows a breakdown of number of studies by publication year and type.

As can be seen, the number of Cochrane reviews appears to be greater in the more recent years.

6.2.3 Selection of relevant data for extraction

The following facets of each study were selected for further consideration:

1. general information;

2. statistical methodology aspects;
3. dissemination bias;
4. heterogeneity;
5. individual patient data;
6. quality assessment; and
7. sparse data.

These aspects of the meta-analyses will be discussed in greater detail below.

The information about the primary studies was recorded using a relational database created in Microsoft Access. The use of a relational database allows easy cross-referencing between tables of the dataset, which facilitates using the database to answer specific questions about the methodological aspects of the included studies. This is especially helpful in a situation where there is much interconnectedness between the different items of data recorded. Each table of the database includes a number of fields and by linking each table to a reference table (including the bibliographic details for each reference), through a unique key, it is possible to perform queries across the tables, allowing great flexibility for interrogating the database.

In addition to the fields described that are specific to each table, all tables included a Notes field for further information, explanation of terms, clarification of selected values for other fields and so on. Free text searching of this field would allow the retrieval of references with key terms.

6.3 Discussion of relevant aspects of meta-analyses

6.3.1 General information

General information refers to non-statistical information about the paper, such as its aim, the type of intervention used, the nature of the outcome and the source of sponsorship. The full list of items of information in this category is given below.

1. Primary aim of the study.

2. Intervention (primary intervention chosen from drug(s), surgery, diet, anaesthesia, diagnostic, device, immunisation, public health intervention, multiple or other).
3. Outcome (main outcome(s) of study).
4. Number of meta-analyses performed (the total number of pooled estimates across the reference). The purpose of this field is to give an idea of the size of the study. The number of meta-analyses are presented in bands (0, 1, 2–5, 6–10, >10) rather than as the actual number.
5. Number of estimated effect sizes derived from primary studies presented in bands (2–5, 6–10, 11–20, >20), with maximum and minimum numbers if more than one meta-analysis and numbers of primary estimated effect sizes in different bands.
6. Study types (selected from randomised controlled trials (RCTs), randomised trials (RTs), controlled trials (CTs), other trials (OTs), observational studies and mixed (including trials and observational studies) and other).
7. Graphs (refers to graphs for main meta-analysis, selected from forest plot, meta-regression plot, forest plot and meta-regression, other or none).
8. Sponsor (selected from academic, government, commercial or other).

The number of meta-analyses was determined by regarding an individual meta-analysis as being a unique combination of outcome and the set of primary estimates pooled. Repeated meta-analyses for the same outcome and set of primary estimates, using different methods, such as fixed and random effect(s) models, would not count as multiple meta-analyses but only as one. Referring to point 8 above, if more than one funding source was stated, a commercial sponsor would take precedence over government or academic sponsorship; likewise government sponsorship would take precedence over academic. This hierarchy was developed in order to be able to identify those studies with some commercial sponsorship (even in the presence of non-commercial sponsorship) and those studies with government sponsorship (including those with some academic sponsorship).

6.3.2 Statistical methodology aspects

In this section, basic meta-analysis methods are recorded, as well as the presence or absence of further aspects of the meta-analysis that would not be expected to occur in all meta-analyses. This could be due to lack of appropriateness, or possibly lack of thoroughness, or a decision not to extend the potential of the meta-analysis as far as it could be taken.

Fourteen specific aspects were included, as discussed below. For many of the areas the responses fell into a recurring pattern of standard responses comprising 'Yes' or 'No', sometimes with 'Not applicable' or 'Not stated' as appropriate.

Measure of effect

In total, 12 distinct outcome measures were individually coded, with 'Other' and 'Multiple' as appropriate. The main outcome measure was chosen for each reference, with 'Multiple' only being selected if there were two or more outcome measures, but not obviously a principal one.

The options are listed below.

1. Comparative measures (between interventions).
   (a) odds ratio (OR);
   (b) relative risk (RR);
   (c) risk difference (RD);
   (d) mean difference;
   (e) standardised mean difference (for example when using scoring systems for clinical conditions such as depression); and
   (f) percent difference;
2. Non-comparative measures (not making a comparison between interventions, either making a comparison within an intervention (such as 'before' and 'after') or non-comparative).
   (a) correlation;
   (b) percentage/probability;
   (c) mean difference; and

(d) percent difference.

3. Multiple outcomes.

   (a) multiple.

4. Other.

   (a) other.

In the list above, 'comparative' refers to the situation where two distinct interventions (one of which may be a control or no intervention) are directly compared within a resulting single figure such as a ratio or a difference. In some cases, a 'comparative' outcome metric such as mean difference could be used not to compare two separate interventions but to compare one intervention at different time points; this use of such a metric is classed as 'non-comparative'. The use of 'Multiple outcome' was intended to encompass situations where more than one outcome measure was used with roughly equal importance. In cases where there was clearly one outcome measure that was the primary measure used, this would be selected rather than using 'Multiple'. The 'Other' category was intended to cover situations where a non-standard or unique outcome measure was used.

Meta-analysis method

To allow easier classification, the methods were subdivided into categories, based on the statistical approaches. These included:

1. standard fixed effects (including inverse variance, Mantel–Haenszel and Peto methods as well as any other fixed effect model using standard or referenced methods);

2. other fixed effects (for example, using the author(s)' own method of weighting the individual study estimates such as by study size or some function thereof, using logistic regression methods or using a non-referenced method);

3. standard random effects (in practice this usually referred to the DerSimonian & Laird method for incorporating a between-studies heterogeneity value into the model);

4. marginal analysis (pooling of aggregate results from all studies without accounting for the fact that they came from different studies);
5. Bayesian methods; and
6. no pooled estimate (heterogeneity tests only).

Meta-analyses that used Bayesian methods may have been random or fixed effect(s) models but were classified as Bayesian as it was considered important to identify those studies that used any form of Bayesian methods. Studies that did not readily fit into one of these categories were classed as 'Other', 'Multiple' (if no primary meta-analysis method out of several used) or 'Unclear' as appropriate.

Fixed effect/Random effects

The aim of recording this information was to identify the use of fixed and random effect(s) models, leading on to reasons for the choice of model, if stated. The options for this field included fixed effect, random effects and 'More than one' (if both FE and RE were used). Alternatives were 'Other', 'Unclear', 'Not stated' and 'Not applicable'. If there was clearly a primary choice of either fixed or random effect(s), then this was stated, but in the case where both fixed and random effect(s) models were used with no clear precedence, then the 'More than one' option was brought into use.

Reason for choice of fixed effect/random effects

The reasons why fixed or random effects were used were also of interest, with several main reasons being repeatedly cited. These included:

1. heterogeneity present;
2. study types;
3. increased conservatism;
4. results similar;
5. larger studies contribute more to estimate;
6. to account for between-studies variation; and
7. no heterogeneity present.

The responses 'Other', 'Not stated' and 'Not applicable' were used as required.

Heterogeneity

The heterogeneity category referred to whether heterogeneity was considered in the meta-analysis, whether quantitatively or qualitatively, including the use of meta-regression or a subgroup analysis to investigate heterogeneity. A positive or negative response was recorded in each case.

Individual patient data

Use of individual participant (or patient) data (IPD) was a straightforward question of whether IPD was used in the analysis. It was always possible to classify without ambiguity the answer to this question, as those studies that used IPD always stated that they were doing so.

Sparseness of event data

Sparse data referred to the presence of zeroes in studies where binary outcomes were used. If this was clearly stated, then a Yes/No response was recorded. In some cases where sparse data may have occurred, it was not possible to determine from the reference if there were any instances of studies with zero events in one or both comparison groups, hence a 'Not stated' response was recorded. In some cases the nature of the outcome variable (i.e. a continuous variable rather than a binary or count variable) precluded sparse data, in which case a response of 'Not applicable' was recorded.

Multiple outcomes

The issue of multiple outcomes was considered important to investigate due to the likelihood of multiple adverse events occurring following an intervention. Also, the presence of multiple outcomes lends itself to the difficulties incurred in multiple testing. In all meta-analyses it was possible to determine if only one outcome or multiple outcomes were investigated, hence a 'Yes' or 'No' response was recorded for each study.

Subgroups

Subgroup analysis is also important where adverse events are concerned, as it is possible that certain types of patients may be at greater risk of an adverse event. In this context, a subgroup is defined as a function of patient characteristics within each study of the meta-analysis, such as age or sex, rather than as a function of the intervention, or of study-level characteristics such as year of

publication, country of study or study design, which may be investigated using sensitivity analysis. In all cases, it was possible to record positively or negatively the presence of subgroup analyses.

Dose–response

Applicable to drug interventions only, dose–response referred to whether any type of dose–response analysis had been considered, in this context referring to doses of the intervention drug rather than to duration of exposure. For drug interventions, a response of 'Yes' or 'No' was recorded, otherwise 'Not applicable' was recorded.

Dissemination Bias

Dissemination bias encompasses issues of publication bias, reporting bias and citation bias. In this case it referred to whether any attention had been paid to these issues of bias, at any level (including a proactive search for unpublished references which indicated that publication bias had been considered, even if not discussed further). It was always possible to record a positive or negative response.

Quality

Quality assessment received a positive response if there had been any discussion of quality issues in the meta-analysis. A 'Yes' or 'No' response was always recorded.

Class effects

Class effects referred to the investigation of more than one drug of the same class, such as non-steroidal anti-inflammatories (NSAIDs) or selective serotonin reuptake inhibitors (SSRIs). For drug interventions a 'Yes' or 'No' response was always recorded, while 'Not applicable' was recorded for all other interventions.

Time course effects

In this category the aim was to select meta-analyses that investigated the same outcome at different times following administration of the intervention to look for changes in the adverse events profile over time (for example, performing a meta-analysis at different times postoperatively). A 'Yes' or 'No' response was recorded for all interventions.

Some of these topics were selected for more detailed consideration, namely dissemination bias (Section 6.3.3), heterogeneity (Section 6.3.4), IPD (Section 6.3.5), methods for quality assessment of primary studies (Section 6.3.6), and issues surrounding sparseness of events data (Section 6.3.7).

6.3.3 Dissemination Bias

Dissemination bias encompasses both publication bias, and the slightly more subtle concept of reporting bias, the deliberate exclusion of outcomes that were non-significant or showed an undesirable outcome such as a significant increase in an adverse event due to a certain intervention.

There were nine aspects of data recorded with regard to dissemination bias. The first recorded whether or not publication bias was mentioned (including searching for unpublished results, which implied that publication bias had been considered in the study execution).

It was then recorded whether publication bias had been discussed only, without any formal graphical assessment or statistical tests, or whether a search for unpublished references had been made without further discussion of publication bias. An alternative option was the performance of a sensitivity analysis by publication status as opposed to statistical tests and graphs. The presence of a test with a p-value was recorded, as well as the name of the test where applicable.

The presence of graphs (funnel plot or none) was recorded, as was the adjustment of results for the presence of publication bias. For all studies (irrespective of whether publication bias was formally mentioned) the study sources were recorded. This category referred to the source of the actual studies used in the meta-analysis, so for example, a study that searched for unpublished studies but did not find any would be classed as published only, as would a study that intentionally did not search for unpublished studies. The options for data recording were 'Published only', 'Published studies with unpublished data' (obtained by contacting the authors), and 'Mixed' (published and unpublished studies), with 'Unclear' and 'Not stated' as appropriate. Finally, it was recorded whether reporting bias was mentioned.

6.3.4 Heterogeneity

The wide diversity of types of heterogeneity makes it impossible to do justice to this subject in its entirety. This review focuses primarily on statistical heterogeneity only (as opposed to clinical or methodological heterogeneity), and how it is identified and investigated in meta-analyses of adverse events data. Ten aspects of heterogeneity within meta-analysis are discussed.

The first of these records the type of heterogeneity assessment made for statistical heterogeneity. The options are 'Quantitative' (selected if there is any type of quantitative assessment of heterogeneity), 'Qualitative' (if heterogeneity is discussed in a qualitative manner only), 'Both' (for both quantitative and qualitative assessment), and options of 'Unclear' or 'No assessment' as appropriate.

Also recorded was the presence of an estimate for heterogeneity, with options of 'Yes', 'No', 'Unclear' or 'Not applicable'. This was followed by recording whether or not a test for presence of statistical heterogeneity had been performed, with the same response options as for the presence of an estimate. The names of any estimates or tests were also recorded. Estimates were recorded by name; however, if no estimate was performed but a test was present just 'Test' was recorded (it was decided not to name individually the many tests for heterogeneity due to the large number), along with 'Not stated', 'Not applicable', 'Other' and 'Unclear' when necessary.

If a test with a significance level was present, this significance level was recorded. If the significance level was not stated but p-values were given, then this was recorded. Options included, '0.05', '0.1', '0.2', 'p-value', and 'Not stated' and 'Not applicable'.

The presence of heterogeneity as determined by estimates, tests or qualitative assessment was recorded, as was the presence of subgroup analysis and meta-regression analysis. The use of subgroup analysis or meta-regression was not necessarily specifically intended to be for the purpose of investigating heterogeneity (if present) but both are recorded at this point for convenience. Finally, the presence of a qualitative discussion of potential causes of detected heterogeneity was recorded.

6.3.5 Individual participant data

Only three aspects of the use of IPD were included. The number of cohorts (bearing in mind that an individual primary study may include data from multiple cohorts of participants) with IPD was reported, along with the total number of cohorts within the overall meta-analysis. Also, the types of data included in the meta-analysis were recorded (IPD alone or in combination with summary data).

Finally, the method of incorporation of IPD, which could be a one-stage stratified, one-stage unstratified or two-stage meta-analysis, was recorded. A one-stage stratified meta-analysis refers to using the IPD as one overall dataset with stratification by study as a covariate, while a one-stage unstratified meta-analysis involves using all the IPD as one dataset but without taking into account the fact that data observations are derived from different studies by including study as a covariate, either in a fixed or random effect(s) model. A two-stage meta-analysis by contrast, takes the individual patient data from the original studies, calculates each study-level estimate, and then combines them using methods for aggregate-data meta-analysis.

IPD methods are discussed more fully in Section 5.6.

6.3.6 Quality assessment

Seven aspects of quality analysis were decided upon for data extraction. The number of assessors was recorded (with options of 'One', 'Two', 'More than two', 'Not stated' and 'Not applicable'). Resolution of disagreement (assuming two or more assessors) was also recorded, with responses of 'Agreement measure' (such as inter-rater agreement measures), 'Consensus', 'Additional assessor', 'Other', 'Not stated' and 'Not applicable'.

The use of a quality tool was recorded, and whether more than one of these was used, as well as the names of quality tools used. Additional quality aspects of the meta-analysis were noted, as well as whether the quality information was used in the meta-analysis, and if so, the means of incorporating it. Options for this field were 'Exclusion' (of poor quality studies), 'Subgroup' (dividing the studies into subgroups and performing analysis on studies by quality level), 'Other', 'Not used' and 'Unclear'.

6.3.7 Sparseness of event data

The presence of double-zero studies (primary studies with no events in both comparison groups, such as trial arms, exposed/non-exposed groups in a cohort study, or case and control groups in a case–control study), as opposed to single-zero primary studies with zero events in one comparison group only, was recorded. It was also noted whether such double-zero studies were excluded from the meta-analysis.

The use of continuity corrections was recorded, with options of 'Yes', 'No', 'Unclear', 'Assumed' and 'Not applicable'. (It was assumed that if double-zero primary studies were not present then the continuity correction would be applied to single-zero primary studies.) The primary continuity correction was noted, with responses of '0.5', '0.25', (these were the only two specified continuity corrections used), 'Not stated' and 'Not applicable'. It was also recorded whether there was a reason given for the choice of continuity correction; the subsequent field records the reason (where appropriate) with options of 'Minimise bias', 'Not stated' and 'Not applicable'. If more than one continuity correction was applied, it was recorded whether the results were sensitive to the choice of continuity correction.

Finally, any other methods for coping with zero events were recorded with responses of 'Peto' (the Peto method for meta-analysis), 'Risk difference' (an outcome measure which can cope with zero events as it is a difference rather than a ratio), 'Marginal analysis', 'Bayesian' (using Bayesian methods that can allow for zero events), 'Other', 'No' and 'Unclear'.

6.4 Results

6.4.1 General information

There was a wide range of types of adverse events investigated in the set of meta-analyses. To give an idea of the diversity of conditions, they included events such as mortality, cancer risk, fetal malformations, risk of infection, psychosocial outcomes, or physiological measurements such as changes in blood pressure, weight or bone mineral density. The outcomes could also be unexpected or

unanticipated as an adverse event, for example a reduced ability of screening tests to provide an accurate result.

The interventions being evaluated for adverse events were similarly diverse; by far the largest category was drug interventions with 69.9% (116/166). The next largest category was surgical interventions with 8.4% (14/166) of references. Anaesthesia interventions accounted for 3.6% (6/166 references); [S25; S100; S121; S143; S148 S166] while devices were investigated in 3.0% (5/166 references); [S32; S74; S75; S127; S164]. Diagnostic procedures were discussed in two references (1.2% of references); [S1; S17]. Public health interventions were also discussed in two references (1.2% of references); [S71; S79]. Immunisation was mentioned in only one reference [S128], or 0.6% of references. Dietary interventions were also investigated in only one reference (0.6%) [S4]. Multiple interventions (often multiple therapies for cancer being assessed simultaneously) were included in 5.4% (9/166) of references [S5; S18; S65; S80; S96; S105; S106; S109; S122] while other interventions accounted for 6.0% (10/166) of references [S16; S30; S39; S101; S114; S118; S119; S152; S153; S154]. Other interventions included blood transfusion, bone marrow transplantation, home birth, early postnatal discharge and preconception care.

Of the 166 references that were included for data extraction, all except one included some overall combined estimate. The one reference that did not included a meta-regression for dose–response with no overall pooled estimate of effect size [S95].

The number of meta-analyses performed by each meta-analysis study was also very variable. It was interesting to note that in many cases a large number of meta-analyses was performed with 44.6% (74/166) of studies having more than 10 meta-analyses. By comparison, 19.3% (32/166) of studies had 6–10 meta-analyses, with 28.9% (48/166) having 2–5 meta-analyses, 6.0% (10/166) had only one meta-analysis while 0.60% (1/166) had none (meta-regression only); [S95].

The number of cohorts (contributing individual datapoints) is shown in Table 6.2 (individual studies not referenced in Appendix B). This table indicates that many studies include meta-analyses with larger and smaller numbers of cohorts, possibly reflecting multiple outcomes or subgroup analyses. The number of cohorts was not applicable for one study [S95], which included meta-regressions

Table 6.2: Maximum and minimum numbers of contributing estimates for meta-analyses in the same references.

Maximum no. contributing estimates[a]	Minimum no. contributing estimates	No. meta-analyses	% meta-analyses[b]
2–5	2–5	17	10.4
6–10	2–5	34	20.7
6–10	6–10	6	3.7
11-20	2–5	50	30.5
11–20	6–10	5	3.0
11–20	11–20	4	2.4
>20	2–5	29	17.7
>20	6–10	12	7.3
>20	11–20	3	1.8
>20	>20	4	2.4

[a]A single primary study may contribute more than one estimate, hence number of contributing estimates may not be the same as number of primary studies.
[b]Out of 164 studies.

only, and not stated for one study [S35], hence 164 studies contributed to Table 6.2.

Table 6.3 shows the primary study types encountered in this review. The most frequent study type was some form of trial, the sole study type for 46.4% (77/166) of meta-analyses. The 26 studies (26/166, 15.7%) that included both trials and observational studies are worth additional scrutiny. It is of interest to know if these meta-analyses combined results directly from trials and observational studies or whether the data from the two study designs are combined separately.

Table 6.3: Study types incorporated within meta-analyses.

Study types	No. studies	Percent studies[a]
Randomised trials	70	42.2
[b] Other trials	7	4.2
Observational studies	56	33.7
Mixed (trials and observational studies)	26	15.7
Not stated	7	4.3

[a]Out of 166 studies.
[b]May include randomised trials but not specifically stated as such.

There was a wide variety of approaches taken to this situation, often reflecting the number of each different type of study. Some meta-analyses made no attempt to differentiate by study design [S13; S39; S129; S138]. In some cases there was only one RCT, all other studies were observational, and in one of these meta-analyses, the RCT was excluded, although its inclusion did not alter the results [S15]. In another instance with only one RCT, it was excluded from all meta-analyses, only the observational studies (of different designs) being included [S37]. In one meta-analysis [S53] the sole RCT was excluded due to no events being observed in one group of the study; similarly, in another meta-analysis with only one RCT, this study was excluded due to the small number of cases [S108].

The most common approach to mixed study types was to perform a sensitivity analysis by analysing all studies together and then dividing the studies by some element of study design. For example, one meta-analysis [S87] analysed all studies together and then case–control studies and cohort studies plus RCTs were analysed separately. Some variation on this theme was followed by several other meta-analyses [S25; S63; S98; S101; S103; S107; S121; S130].

Another approach was to avoid combination of estimates across study designs, by combining results from studies with similar designs. For example, one meta-analysis [S30] analysed cohort studies separately from RCTs, another [S111] analysed RCTs, case–control and cohort studies separately and a further meta-analysis [S156] analysed RCTs and observational studies separately.

Some meta-analyses took a more unique approach. For example, some meta-analyses combined all study designs for some outcomes or outcome measures, while for others only a subset of study types was included [S134; S137; S149]. In one case [S32] studies were assessed for quality score based on study design, and then analyses performed based on quality score, with the result that all studies were analysed together, with RCTs and higher quality observational studies being analysed as a separate subset of primary studies.

Another approach seen in only one study [S17] was that the authors developed their own Bayesian methods to combine cohort (including studies where the cohorts were randomised to their exposure, i.e. randomised trials) and case–control data. Their methods involved ways of including cohort studies without concurrent controls. In one study [S68] it was implied that some studies were

experimental and others observational, and it was not possible to determine whether different study designs had been quantitatively combined.

Graphical representations of data were used in the majority of meta-analyses. Forest plots were the only graph used in 53.0% (88/166) of references, while meta-regression plots were the only plot in 1.8% (3/166 references [S95; S118; S162]). Both forest plots and meta-regression plots appeared in 1.8% (3/166) of references [S64; S67; S142]. Other plots were used in 18.1% (30/166) of references, usually a plot of the individual studies but lacking a pooled estimate. Only 25.3% [42/166] of references produced no graphical representations of their results.

With regard to sponsorship, the largest number of studies were academically sponsored with 45.8% (76/166) of references. Commercial sponsorship accounted for 16.3% (27/166), while 30.1% (50/166) were sponsored by some form of government body. Other sponsorship sources provided funding for 7.2% (12/166) while funding source was not stated for one study [S4].

6.4.2 Statistical methodology aspects

Table 6.4 shows the proportions of studies using different effect measures. This is important because the choice of outcome measure may in itself influence the meta-analysis method and results.

In the above table the outcome 'More than one' was only selected in the eventuality that there was no obvious primary outcome metric, but instead at least two outcome metrics that appeared to receive largely equal importance in the meta-analysis. Otherwise, the primary outcome measure was recorded even if there were other outcome metrics used in secondary analyses such as sensitivity analyses. The diversity of outcome metrics reflected the nature of the data being analysed.

Only one meta-analysis had an outcome measure assessed as 'Other' [S8]. In this study the outcome measure was the relative proportion of incorrect screening diagnoses for breast cancer, comparing women using HRT with those not using HRT, for both specificity and sensitivity. Hence, this was a comparative outcome. Also, one study [S156] presented a percent change that was both comparative (percent extra change in outcome between patients receiv-

Table 6.4: Measure of effect.

Measure of effect	No. studies	Percent studies[a]
Comparative measures (between interventions)		
Odds ratio	55	33.1
Relative Risk	51	30.7
Risk difference	8	4.8
Mean difference	6	3.6
Standardised mean difference	7	4.2
Percent difference	2	1.2
Non-comparative measures (not between interventions)		
Correlation	2	1.2
Probability (or percent)	13	7.8
Mean difference	2	1.2
Percent difference	3	1.8
Multiple measures		
More than one	16	9.6
Other measures		
Other	1	0.6

[a]Out of 166.

ing the intervention and those not) and non-comparative (percent change for patients receiving the intervention and those not presented separately). This study was recorded as percent difference in the comparative section rather than non-comparative, as it was considered that the comparative outcome was of greater importance.

Many of the outcomes being reported in the primary studies were binary, thus lending themselves to analysis by odds ratio or relative risk, as seen in Table 6.4. Interestingly, the difference scale (for example risk difference), as opposed to a ratio scale, was chosen only infrequently, even when its use would have been an appropriate alternative option to a ratio scale. Only 9.6% of studies presented more than one outcome measure with a comparable degree of importance. This indicated that either more than one outcome measure was being analysed, requiring a different outcome metric, or that the authors had presented analyses for the same outcome on equal terms for different outcome metrics, possibly as an intended comparison between the two.

Table 6.5: Meta-analysis methods.

Meta-analysis method	No. studies	Percent studies[a]
Standard fixed effect[b]	54	32.5
Other fixed effect[b]	9	5.4
Standard random effects[b]	50	30.1
Marginal analysis	3	1.8
Bayesian methods	6	3.6
Multiple analysis methods	33	19.9
Other	1	0.6
Not stated	5	3.0
Meta-regression only	1	0.6
Heterogeneity test only	1	0.6
Unclear	3	1.8

[a] Out of 166.
[b] See text for definition.

Another major area of interest was the methodology used for the meta-analysis itself. Table 6.5 sets out the meta-analysis type along with numbers and percentages.

As seen from Table 6.5, fixed and random effect(s) models were used with roughly equal frequency. The term 'standard fixed effect' was used when the authors chose an accepted fixed effect model, such as the Mantel–Haenszel model, the inverse variance model or the Peto model. Also, if the authors used any referenced fixed effect model this was recorded as 'standard fixed effect'. In several cases the authors had used a method of combining data that would be considered as a 'fixed effect' model, but appeared to have either used a mean or weighted mean, or a logistic regression method, or had devised their own method for combining data, for example based on sample size of the studies. Such methods were non-standard and were recorded as 'other fixed effect'. Interestingly, all the random effects models were referenced standard models (referred to as standard random effects in Table 6.5, such as the DerSimonian & Laird model).

A record of 'Multiple analysis methods' was only selected when there was more than one method used on an equal basis. If there was an obvious primary method with additional supplementary methods the primary method was chosen.

The one study referred to as 'Other' [S48] used a novel approach developed by the authors of 'summary ranking' involving assigning a score to the rank order of toxicity in individual studies, and then combining the scores to provide an overall rank order for toxicity. One study included a meta-regression as the only quantitative analysis [S95], whilst another study performed a test for heterogeneity but no additional analyses [S8]. In only three cases was there insufficient detail regarding the methodology to allow the type of analysis to be determined [S13; S61; S90]. In five of the reviewed studies the meta-analysis method was not stated [S38; S45; S46; S128; S139].

A Bayesian approach was used by six meta-analyses, but in some cases the Bayesian model used in the meta-analysis was not fully described. For example, one meta-analysis [S89] referred to the use of a Bayesian analytic framework but did not explain this framework further; however, details of the MCMC implementation were provided. Both fixed and random effect(s) models were used initially but did not produce different results, so only the random effects model was presented. A heterogeneity test indicated non-heterogeneous results, but it was not stated whether this test was Bayesian or non-Bayesian. Another study [S103] also used a Bayesian data analysis framework, with both fixed and random effect(s). Meta-analysis using a random effects model with a Bayesian data analytic framework was the choice of one meta-analysis [S70]. In some aspects, this study provided the most detail regarding its Bayesian analysis, reporting that only non-informative priors were used, adequate convergence assured and stating the number of chains and simulations, although not in terms of model description.

A Bayesian hierarchical model using the authors' own weighting system which incorporated within-trial and between-trial variability was used in one meta-analysis [S150], based on the confidence profiling method (Eddy & Hasselblad 1990). The same approach was also used in an empirical Bayes method with a random effects model [S108].

Semi-Bayesian methods, incorporating a Bayesian use of a prior on the rate of the adverse event alongside standard frequentist methods were also used [S17]. This meta-analysis also experimented with the use of different prior distributions for the parameters. However, none of these studies presented graphical representations of the probability densities for the parameters.

Not taking into account the specific method of the meta-analysis, 74/166 (44.6%) used a fixed effect model. This included one study which stated a random effects model was used, but in the absence of heterogeneity the presented results were fixed effect [S43]. The novel method using rank summaries was also a fixed effect study [S48]. Two studies where the methodology was unclear [S13; S61], and one where the methodology was unstated [S45], were, however, able to be classed as a fixed effect model.

A random effects model was used by 53/166 (31.9%) meta-analyses including four of the Bayesian meta-analyses [S70; S89; S108; S150]. Hence, it is apparent that fixed and random effect(s) models were used with roughly equal frequency in meta-analyses of adverse event outcomes.

Both fixed and random effect(s) models were used in 28/166 (16.9%) studies, including one Bayesian study [S103]. In the other cases it was not applicable (5/166; 3.0%); [S8; S95; S118; S144; S145], or not stated (5/166; 3.0%); [S38; S46; S90; S128; S139].

It was unclear which method was used in one reference only, which stated that the Mantel–Haenszel method was used, and was then referred to as a random effects method [S87].

For references where it was clearly stated whether the method was fixed or random effects, or both, it was interesting to record the reasons why the authors chose that particular approach, and 69/166 (41.6%) studies did provide some explanation for the model choice. Reasons based upon heterogeneity (or between-study variation) were the most commonly cited, with 46/69 (62.3%). Increased conservatism (of a random effects model) was also frequently mentioned with 7/69 (10.1%) references alluding to this [S30; S35; S107; S108; S111; S123; S124].

Other reasons cited in support for a particular meta-analysis method included differences in primary study types, [S17; S152]; and so that larger studies would contribute more to the meta-analysis [S77]. Arriving at similar results from both approaches was also used to justify the chosen approach [S59; S85; S89]. Only one study offered multiple explanations [S151], while nine studies offered an explanation not mentioned above [S9; S12; S16; S20; S31; S55; S104; S113; S146].

Table 6.6: Statistical methodology aspects.

Category	Considered (no. (%))	Not considered (no. (%))
Heterogeneity (assessment, discussion or meta-regression)	138 (83.1)	28 (16.9)
Individual patient data	2 (1.2)	164 (98.8)
Sparse data	65 (39.2)	101 (60.8)
Multiple outcomes	104 (62.7)	62 (37.3)
Subgroups	33 (19.9)	133 (80.1)
Dissemination bias	89 (53.6)	77 (46.4)
Quality	70 (42.2)	96 (57.8)
Time-course analysis	18 (10.8)	148 (89.2)

Table 6.6 indicates the proportion of references (out of 166) that included a consideration of the aspect of data analysis mentioned. Note that it does not provide any information about whether a particular aspect of data analysis, for example heterogeneity, was in fact present in the analysis, merely that it was considered in some way.

The meta-analyses that discuss dose–response issues are not included in the table, due to the fact that only certain interventions lend themselves to a form of dose–response analysis. Drug interventions and anaesthetics are amenable to dose–response analysis, but also diets may be applied at different 'doses', and depending on the nature of the intervention, public health programmes may also be suitable for dose–response analysis. Of the 27 studies included, 26 were of drug interventions, with one public health intervention [S79]. Of all the studies involving a drug intervention, 26/116 (22.4%) included some form of dose–response analysis.

Similarly, the meta-analyses that referred to class effects potentially included studies with drug or anaesthesia interventions. Of the 75 studies that discussed class effects, all but three had a drug intervention. One study investigated anaesthetics [S100] while two had multiple interventions [S105; S106]. Of the 116 studies where the intervention was a drug, 72 (62.1%) included some investigation based on the class of drug.

Exploring any possible changes in the methodological aspects included in meta-analyses of adverse events data over time, choosing heterogeneity, quality and

Table 6.7: Percentages of meta-analyses by year including heterogeneity, quality and dissemination bias.

Year	Total no. studies	Heterogeneity (no. (%))	Quality (no. (%))	Dissemination bias (no. (%))
1994	4	2 (50)	1 (25)	1 (25)
1995	8	7 (88)	1 (13)	4 (50)
1996	9	8 (89)	1 (11)	6 (67)
1997	15	12 (80)	8 (53)	5 (33)
1998	16	14 (88)	3 (19)	10 (63)
1999	20	18 (90)	8 (40)	12 (60)
2000	11	9 (82)	4 (36)	4 (36)
2001	22	19 (86)	8 (36)	15 (68)
2002	21	16 (76)	13 (62)	8 (38)
2003	22	16 (73)	10 (45)	15 (68)
2004	13	11 (85)	7 (54)	5 (38)
2005	4	4 (100)	4 (100)	3 (75)
2006	2	2 (100)	2 (100)	1 (50)

dissemination bias, Table 6.7 sets out the number of studies incorporating each aspect, published between 1994 and 2006.

Some of the aspects of data analysis mentioned above, including heterogeneity, publication bias, individual patient data, sparse data and quality are discussed in greater detail below. Those not discussed further, such as multiple outcomes, subgroup analysis, class effects, and time course analysis are not disregarded due to lack of interest, and indeed, some are covered in case studies in later chapters. However, data extraction for these subjects was often very complex and not easily reduced to simple categories, hence these aspects were not pursued further in this essentially quantitative review.

6.4.3 Dissemination Bias

Publication bias was mentioned in 89/166 (53.6%) of references. Such a 'mention' may not have been specifically described as a consideration of publication bias; for example, some studies performed searches for unpublished studies, indicating that publication bias was within the awareness of the authors when performing a meta-analysis even if it was not taken any further than searching for such studies. Similarly, many studies restricted their consideration of publi-

cation bias to a brief mention within the discussion section of the paper, usually with effect that the authors did not consider publication bias to be a problem in their study design, often accompanied by some explanation of why this should be so.

The next stage in recording data about how publication bias was dealt with in meta-analyses of adverse events was to investigate how many studies performed some type of quantitative analysis and how many offered a discussion only. The latter option was preferred by 44/166 (26.5%) studies (44/89; 49.4%). A quantitative analysis was performed by 31/166 (18.7%) studies (31/89; 34.8%). A sensitivity analysis by publication status was the preferred method of investigating publication bias for one study [S123].

Out of 31 references with some form of quantitative analysis 12 (38.7%) used a test with a *p*-value. The other 19/31 (61.3%) used an alternative means not resulting in a *p*-value. The most commonly used tests were Egger's test [S30; S41; S45; S111; S121; S129] and Begg's test [S30; S41; S66; S111; S129]. Kendall's tau test was mentioned by three meta-analyses [S51; S62; S83].

One study investigated any possible association between number of incidences of the adverse event and the magnitude of association between the adverse event and the intervention, quoting a *p*-value for a significant difference between the two study types, based on number of cases [S22]. Another non-standard test for publication bias involved investigation of correlation between study size and RR [S53].

In one case, the Trim and Fill method was used to adjust for publication bias [S103]. This was the only study that adjusted for publication bias; one study adjusted for selection (but not publication) bias [S31], whereby outcomes with significant or desired results are reported while other outcomes are not.

Graphical methods (funnel plots) were used to investigate for publication bias in 29/166 references (17.5%), or in 28/89 (31.5%) of studies that discussed dissemination bias (a funnel plot was also used in study S31, which discussed selection bias).

The vast majority of meta-analyses used only published studies (129/166 or 77.7%). Published studies with unpublished data (obtained through contact with the authors), were used in 20/166 (12.0%) of meta-analyses. Both pub-

lished and unpublished studies were used in 14/166 (8.4%). In the other cases the study source(s) was either unclear or not stated. In six cases where published studies only were included, it was made clear that unpublished data had been sought [S20; S32; S54; S73; S76; S109].

Reporting bias was rarely mentioned with only 3/166 (1.8%) discussing this issue [S157; S159; S161].

6.4.4 Heterogeneity

Heterogeneity was considered in some manner by 138 of 166 references (83.1%). A quantitative assessment was performed in 124/166 of these meta-analyses (74.7%). Qualitative assessment of heterogeneity (for example inspection of forest plots or noting heterogeneous results) was made in 10/166 (6.0%) or 10/128 (7.8%) references. Six studies included both quantitative and qualitative aspects of heterogeneity (6/128; 4.7%); [S14; S19; S44; S45; S86; S129], while four studies included only qualitative assessment [S72; S74; S110; S156]. Meta-regression was included in nine studies that had no other assessment of heterogeneity, whilst one study discussed issues regarding combination of primary studies with different criteria, but did not do a formal qualitative or quantitative analysis of heterogeneity [S61].

Considering quantitative analysis methods, 121/166 (72.9%) studies included some form of statistical test for heterogeneity, although with variation in the chosen critical *p*-value for significance. The chosen significance value was 0.05 for 28 studies (23.1% of the 121 studies that performed a test), while 23/121 chose a more liberal *p*-value of 0.1. (19.0%). Only one study [S119] chose 0.2 as the cut-off *p*-value. In many cases the actual *p*-value was quoted without reference to a particular threshold (51/121 studies (42.1%) did this). In the other cases no *p*-value or significance level was stated.

An estimate for heterogeneity was performed by 16 studies of the 121 with a quantitative analysis (13.2%). The most frequently-used estimate measure was the I^2 statistic (Higgins & Thompson 2002; Higgins *et al*. 2003). This estimate measure was used in 13 meta-analyses. Alternative estimate measures included the between-studies variance [S26; S155]. One study [S44] used another estimate measure, the R(I) statistic (Takkouche *et al*. 1999). Only

one meta-analysis used multiple estimate measures [S133]; the estimates used included I^2 and the Q parameter (Berlin *et al.* 1989).

Heterogeneity was found to be present in 82 of the 128 (64.1%) references that performed any type of analysis for its presence.

Two ways to investigate the causes of heterogeneity are subgroup analysis and meta-regression. Subgroup analysis was performed in 27 of the 128 (21.1%) references that included an analysis of heterogeneity, and in two references that did not formally assess heterogeneity [S107; S157]. The actual subgroups used were a mixture of generic groups that would be applicable in many situations such as age and sex, and those that were more specific to a particular intervention, such as a pre-existing comorbidity. None of the reviewed studies investigated genetic subgroups. In many cases where subgroups were investigated, heterogeneity had been found to be present, although it was not always the case that any subgroup analysis was explicitly to investigate heterogeneity; while in some cases there was no evident heterogeneity.

Meta-regression was used in 27 studies in total (27/166; 16.3%). In nine of these studies no formal assessment of heterogeneity had been performed [S2; S18; S88; S95; S107; S123; S144; S145; S157]. The covariates used in the meta-regression analyses were often very specific to the nature of the intervention or outcome being considered. For example, in one study they included type of anticoagulant, type of prosthetic heart valve and position of valve [S18]. In other cases meta-regressions were used to investigate the influence of more general demographic characteristics such as age or sex, or study characteristics such as quality score or year of publication.

A qualitative investigation of sources of heterogeneity was carried out in 17 studies of the 138 that considered heterogeneity in some way (12.3%). This includes one meta-analysis that discussed potential sources of heterogeneity without making a formal assessment of its presence in the primary studies [S61]. The authors discussed the appropriateness of pooling studies due to different entry criteria for the studies, but did not make any assessment of heterogeneity prior to performing the meta-analysis.

6.4.5 Individual patient data

Very little use was made of IPD in the reviewed meta-analyses. Only two studies of the overall total (1.2%) included IPD [S7; S27]. Of these two, all primary studies included had IPD available (so there was no requirement to combine IPD and summary data). Both meta-analyses used a one-stage method for the meta-analysis. In one case the analysis was stratified by trial [S27] and other factors including centre within study for multicentre studies and age divisions. In the other case it was not clearly stated whether the analysis was stratified or not [S7].

6.4.6 Quality assessment

In contrast to IPD analysis, quality assessment of primary papers was frequently performed by the meta-analysis authors. Considering number of assessors, out of 70 references mentioning quality, one study (1.4%) had only one assessor [S114], 37/70 (52.9%) had two assessors, and 6/70 (8.6%) had more than two [S16; S19; S64; S94; S100; S104]. In the case of the other studies it was either not stated or not applicable (where a study discussed quality issues but did not have a formal scoring system).

Where required, resolution of disagreement among assessors was resolved by consensus in 26/70 cases (37.1%). Recourse to an additional assessor was the chosen method in 3/70 cases (4.3%); [S24; S115; S131]. An agreement measure (for inter-rater difference) was used in 4/70 studies (5.7%); [S90; S110; S117; S135]. One study [S58] resolved disagreement between two initial assessors by consensus or by recourse to a third author.

A quality tool was used in 56/70 (80.0%) studies. More than one quality tool was used in only eight studies [S25; S101; S111; S118; S134; S137; S144; S145]. For trials the most commonly used established quality tools included the Jadad score (Jadad *et al*. 1996), among others (Chalmers *et al*. 1981; Schulz *et al*. 1995). Alternative methods were used for observational studies (Stroup *et al*. 2000), including the Newcastle-Ottawa Scale.[1]

[1]Ottawa Health Research Institute (Undated). *The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses*. Available online [February 2010] at: `http://www.ohri.ca/programs/clinical_epidemiology/oxford.htm`

Other methods included those devised by the Cochrane collaboration and the US Preventive Services Task Force [S103].

Quality scoring systems appear to have been used appropriately. For example, of 18 meta-analyses that used the Jadad score as a quality assessment tool, it was used in 13 meta-analyses that included only randomised trials [S35; S47; S64; S100; S104; S112; S123; S124; S135; S144; S145; S146; S159]. The Jadad score was also used to assess randomised studies in meta-analyses that included randomised and non-randomised studies, whilst alternative quality scoring systems were used for the non-randomised studies [S25; S101; S111; S134; S137]. The Newcastle-Ottawa score was only used in one meta-analysis [S67], which included only observational studies.

There were several methods by which the quality assessment could be used in the meta-analysis. In 13 cases (13/70; 18.6%) poorer quality studies were excluded, while in 16 studies (16/70; 22.9%) some form of subgroup analysis was performed. Some other way of incorporating quality information (such as investigation for a relationship between quality score and effect size or year of publication) was found in 12 studies (17.1%), while in 28 (28/70; 40.0%) cases no further use was made of quality information, and in one case the inclusion of quality data was unclear [S97].

6.4.7 Sparse data

The issue of sparse data, whereby statistical methods were required to allow the inclusion of primary studies where the outcome was a count of zero, or a percent of zero, occurred in 65 of the 166 meta-analyses (39.2%). Such statistical methods may be required to allow incorporation of such a primary study into an overall pooled estimate or for calculation of confidence intervals. It was initially anticipated that this issue would be restricted to two-arm studies with a relative outcome; however, some meta-analyses with sparse data did not fall within this category. These are discussed first, followed by the main discussion of this area as it relates to two-arm studies with relative outcomes.

In only one study [S25], the outcome was a percent (not a percent change), in this case a non-comparative percent referring to risk of an adverse event on different treatments in single-arm studies. In some cases there was a risk of zero (with no events in that treatment arm). In this meta-analysis, a specific

referenced method (described by Ho *et al*. 2002, as cited in S25) was used to calculate the upper limit of the 95% CI for such studies. Continuity corrections were not used to facilitate the calculation of the CIs, which would have been an alternative approach.

In two studies (both of two-arm trials), [S134, S137], there were no events in either arm (across all trials) for one type of adverse event where the outcome would have been incidence rate (thus lending itself to a relative risk or odds ratio outcome). Hence, a marginal analysis was performed with the use of a Poisson method to calculate the upper limit of the 95% CIs for the incidence rate. In these two references an alternative but related outcome method was chosen as an alternative analysis, with an outcome of comparative (between treatments) mean difference. This alternative analysis was the predominant analysis for each study, thus replacing the intended primary analysis due to lack of events.

In the other 62 studies where sparse data was an issue, the outcome measure was always a comparative measure, such as a relative risk, odds ratio or risk difference. For the other studies, sparse data issues were not anticipated due to not having a binary relative outcome or else primary studies with zero events were either clearly not present in the dataset, or it was unclear whether they were present or not.

Of the 64 studies that considered the issue of sparse data in two-arm studies, 41 (64.1%) had datasets involving double-zero (zero events in both arms) primary studies. In the remaining 23 meta-analyses, only single-zero (zero events in only one arm of a two-arm study) primary studies were present, or it was either unclear or not directly stated whether any double-zero studies were included within the dataset. It is helpful to discuss all the meta-analyses with zero events in at least one arm of one primary study at first, and then focus more clearly on those studies where there were primary studies with zero events in both arms.

Of these 64 studies, 30 (46.9%) presented their outcome as an odds ratio, 24 (37.5%) as a relative risk, and two as a risk difference [S38; S61]. As mentioned above, two studies [S134; S137] presented their outcome (where sparse data were incorporated) as an incidence rate.

In six studies there was more than one outcome with roughly equal importance in the meta-analyses [S47; S62; S75; S133; S136; S160]. In one meta-analysis [S62] the outcomes were a relative risk and absolute risk difference; in this case

it was unclear whether a continuity correction was also applied to the absolute risk difference as well as the relative risk.

In another meta-analysis [S75] the odds ratio was used for unadjusted analyses, with the relative risk for adjusted analyses (this outcome appeared to be the principle outcome within the meta-analytic study). The unadjusted analyses used the data from the supplied 2×2 tables for each study and individual outcome. Exact methods and conditional likelihood were used for the unadjusted methods, which could have included the cohorts with zero events in any cell. Adjusted studies used the relative risk or odds ratio directly from the primary study, as well as the standard error reported by the study or derived from the reported confidence interval. This approach necessitated the use of a method based on the inverse variance method (see Section 3.3.2) which would have required a continuity correction if primary cohorts with a single zero, in one or both arms, were to have been included, although it appears that such cohorts were not (outcomes that had zero events in one arm did not have a reported odds ratio or relative risk). A test for homogeneity based on the χ^2-test was used to determine whether the adjusted metrics were estimating the same underlying association between exposure and outcome. If this criterion was fulfilled, the pooled estimate was produced. Adjustment was performed for various factors including age, ethnicity, and several others; adjustment factors varied across studies.

In one study [S160] both a relative risk and risk difference were used as outcomes, the relative risk analysis including double-zero studies and the risk difference analysis including both single- and double-zero studies. The odds ratio and relative risk were also used in another study [S47]. For the relative risk analyses, only single-zero studies appeared, while for the odds ratio analyses there were both single- and double-zero studies, although the double-zero studies were excluded for the meta-analyses.

Finally, for two studies [S133; S136] there were two prominent outcome measures, although one of these was a continuous measurement, and therefore not usually susceptible to issues of sparse data. The other outcome measure was a risk difference, used for data where the outcome was measured as counts.

Continuity corrections were used as a means to address zero counts (in one or both arms of a study) that resulted in the impossibility of making pooled esti-

mates for studies where the outcome metric was a ratio. Continuity corrections are also required to calculate the variance (and hence confidence intervals) for a risk difference. However, it was very difficult to determine an accurate picture of how continuity corrections were used.

Some studies clearly stated that continuity corrections had been used (15/64, 23.4%). In 17/64 (26.6%) studies continuity corrections were not used. In 32 cases it was not clearly stated whether continuity corrections had been used or not, although in some of the cases it would be reasonable to assume that a continuity correction had been used in the absence of any other obvious means of incorporating primary studies with zero events in at least one study arm.

The most popular primary continuity correction was 0.5, used in 14 of the 15 references that stated their continuity correction. Only one meta-analysis used an alternative continuity correction with 0.25 being the chosen value [S59]. Only one study [S106] performed a sensitivity analysis across different continuity corrections, using 0.5, 0.1 and 0.01, and reporting that the continuity correction did not alter the results. Only three studies provided a reason for their choice of continuity correction [S11; S35; S62], and the only reason cited was to minimise bias.

Of the 15 meta-analyses that stated that a continuity correction was used, the primary outcome measure was an odds ratio in 10 cases [S11; S15; S35; S39; S40; S52; S100; S104; S106; S114]. In four cases a relative risk was used as the primary outcome measure [S30; S59; S74; S112]. Only one study used two outcome measures with roughly equal prominence [S62]; these outcomes were an absolute risk difference and relative risk, although it was unclear whether the continuity correction also applied to the absolute risk difference.

For many of the studies with single- or double-zero studies it was impossible to accurately determine how they had been incorporated into the meta-analysis, in terms of whether or not a continuity correction was applied, or if other methods had been employed.

Considering all the studies with some form of sparse data (including both single- and double-zero primary studies), a variety of other methods (other than continuity corrections) were used to analyse sparse data. The most frequently used was the Peto method, employed by 12 studies. The use of a difference metric rather than a ratio as the outcome measure was used to circumvent problems

with zeroes in seven studies [S38; S61; S62; S84; S133; S136; S160], although calculation of confidence intervals with such methods would be problematic. Seven studies resorted to the use of marginal analysis [S46; S81; S134; S137; S144; S145; S160]. Bayesian methods were used to tackle sparsity of events in only two studies [S17; S103].

Double-zero studies were included in 17 of the 41 meta-analyses where double-zero primary studies were clearly present within the dataset. In two cases [S40; S52] they were included in a sensitivity analysis. Double-zero primary studies were clearly excluded in 18 meta-analyses. Such an exclusion was either a deliberate decision by the authors, shown by excluding the primary study in forest plots of the meta-analysis, or was done by default, the primary study being shown on a forest plot, but being given a weighting of zero. The four remaining studies [S46; S84; S99; S129] were unclear as to whether or not these double-zero primary studies were included.

Of the 19 meta-analyses where double-zero primary studies were included in some way, seven made explicit use of continuity corrections [S11; S40; S52; S62; S74; S104; S106]. The primary outcome most commonly used in these meta-analyses was the odds ratio [S11; S40; S52; S104; S106]; alternative outcome measures included relative risk in two studies [S62; S74], and possibly the absolute risk difference as well in one meta-analysis [S62].

In some of the other meta-analyses it was not clearly stated whether continuity corrections had been used. To discuss the various methods used by some of these meta-analyses, in two cases a risk difference was used for count data [S133; S136]. A continuity correction was not mentioned, as it would not be required for the point estimate; however, some form of continuity correction would have been required for the calculation of the variance and hence the confidence intervals for those meta-analyses with zero events across all studies, but such a continuity correction was not discussed . In one meta-analysis, which used both relative risk and risk difference outcomes, it appeared that a marginal analysis was used for both outcomes, although this was unclear [S61].

The two studies described above, where there were no events in the entire dataset, also used a marginal analysis [S134; S137]. For one meta-analysis the methodology was very poorly described, with double-zero studies apparently used in meta-analyses where the outcome was a relative risk, but without any

detail as to how they were incorporated [S160]. A risk difference was also used for some individual meta-analyses reported by this reference.

In two meta-analyses, the majority of primary studies had zero events. These were combined using marginal analyses, with an odds ratio outcome, and the results compared with those of the Peto method, which excluded trials with zero events [S144; S145].

In other meta-analyses either an odds ratio or relative risk was used as the primary outcome, but no specific mention of how double-zero studies were included [S34; S40; S41; S115; S119]. For example, in one of the meta-analyses where double-zero primary studies were included as part of a sensitivity analysis, the Peto method of meta-analysis was employed with OR as the primary outcome measure, implying the necessity of using a continuity correction for the inclusion of studies with zero events in both arms, but there was no mention of using continuity corrections in conjunction with the Peto method [S40].

6.5 Discussion and conclusions

6.5.1 General overview

It is gratifying to see that a wide range of clinical interventions have been investigated using systematic methods and meta-analysis, not just for their efficacy but also for potential adverse effects and problems they may cause. Indeed, the many interventions that have been recognised to be potentially harmful, and thus investigated with deleterious effects in mind, justifies the need for more attention to be paid to the statistical techniques required to approach such outcomes.

This study aims to document previous practice and further understanding of statistical methods used in adverse events meta-analyses, by bringing together, using a systematic approach, information regarding the nature of the methods used and about the types of interventions and associated adverse events.

One striking issue that was immediately noticeable on reviewing meta-analyses of adverse events data was that in many cases, the description of statistical methods was incomplete or unclear. For example, many studies did not consider elements important to any meta-analysis such as how to investigate and account

for heterogeneity or possible publication bias. Methods to consider how to include studies where events are sparse (with zero events in at least one study group) also required more consideration in order to make best use of available data.

The ideal situation would be to achieve an outcome that would be as unbiased as possible, and with as much precision as possible, whilst not excluding potentially useful information. Increased detail and clarity in the description of statistical methods would make it easier to evaluate the conclusions of meta-analyses of adverse events data.

An increase in the number of meta-analyses using non-trial primary studies where available would be beneficial, as it would allow a more complete picture of the situation of an intervention with regard to adverse events to be developed. For example, many RCTs have a limited length of follow-up, and this prevents the collection of data on long-term adverse events, such as certain types of malignancy. Indeed, such adverse events may not be associated with the intervention at the beginning of the RCT. This would also be an area for research into the most appropriate methods for combining different types of data, with a particular emphasis on minimising bias in such analyses.

6.5.2 Meta-analysis method with no direct comparison group

In this review, only one meta-analysis [S17] developed methods for addressing primary studies with no direct comparison group, which is also an area for further consideration.

6.5.3 Graphical methods

Graphical methods for data representation were also not used as frequently as they might be, or in the most appropriate manner. Forest plots for all major analyses would be extremely helpful, as would tabulation of all major results, making clear the number of primary studies, their number of participants, outcomes and results of any heterogeneity investigations. Such transparency of reporting would allow reproducibility of results, and facilitate updating of meta-analyses.

6.5.4 General meta-analysis methods and heterogeneity

As expected, a large proportion of studies made use of standard meta-analysis methods, both fixed and random effects. In many cases however, there was no clear reasoning, whether statistical or clinical, behind the choice of meta-analysis method and whether to use random or fixed effects. Heterogeneity was usually assessed quantitatively when it was considered, most commonly with a test for heterogeneity. However, the thresholds of significance for such *p*-values for such tests were often inappropriately low, possibly resulting in the erroneous conclusion that heterogeneity was not present.

The use of an estimate for heterogeneity, such as the I^2 statistic, occurred in a few studies, notably Cochrane reviews, and the use of estimates rather than a test with an arbitrary cut-off point and low power may be a valuable method that will be used more in future work. It is also important to uncover the causes of heterogeneity, and in many cases there was room for additional investigations, such as subgroup analysis, meta-regression or sensitivity analysis.

6.5.5 Bayesian meta-analysis

Bayesian methods were used in a surprisingly small number of meta-analyses, when considering the fact that Bayesian methods are now easily implemented. This is clearly an area where further research would be both timely and beneficial, especially in the light of many of the difficulties surrounding meta-analysis of adverse events data, which Bayesian methods may be able to address, such as inclusion of primary studies with sparse events (Carlin 1992; Ashby *et al*. 1993; Ashby & Hutton 1996; Sutton & Abrams 2001).

6.5.6 Publication bias

The way that many studies approached the issue of publication bias often left much to be desired; for example, many did not perform any quantitative analysis for the existence of publication bias, or attempt to adjust for it as a form of sensitivity analysis. It was also disappointing that so few studies considered the issue of reporting bias. Search strategies that also looked for 'grey literature'

and unpublished studies would also help to address the problem of publication bias.

6.5.7 Subgroup analysis and meta-regression

It was interesting to note that when subgroup analysis and meta-regression were employed, the nature of the subgroup or the covariate for the meta-regression was often highly specific to the nature of the intervention. This indicates that in order to produce a clear picture of the adverse events profile, it is highly important to consider specific factors that may influence adverse events. A straightforward aggregation of all data may be inadequate to determine additional risk factors or certain vulnerable groups. Unfortunately, one of the major limitations of meta-analysis of aggregate data is the lack of adequate data for subgroup analysis and meta-regression. It is to be hoped that in future there will be more IPD available to facilitate investigation of the causes of heterogeneity. The use of IPD also militates against the danger of ecological bias that can occur when using aggregate data.

A recent review (Koopman *et al*. 2007), including 171 IPD meta-analyses and 102 meta-analyses using aggregate data that addressed similar research questions, found that subgroup analyses were performed in 80% of IPD meta-analyses compared with 45% of aggregate data meta-analyses, correlating with a risk difference of 34% (95% CI 23%, 46%). Interaction tests were also seen more frequently in the IPD meta-analyses compared with aggregate data meta-analyses. However, even the IPD meta-analyses used interaction tests in only 28% of studies.

Possibly reflecting the fact that many meta-analyses are based on primary studies that were performed several years earlier, and would therefore not have included genetic data, there was no meta-analysis found that included genetic subgroup analysis. As the sphere of pharmacogenetics develops, such data may become more readily available, hence allowing possible correlation of adverse events with specific genetic factors.

6.5.8 Quality of primary studies

Although there appears to be a widespread understanding that the quality of primary studies is a major issue when conducting a meta-analysis, with adverse events data no less than with any other type of outcome, there appeared to be a limited understanding of how data on quality can be incorporated quantitatively within a meta-analysis and indeed whether such use is appropriate. The correct use of scoring systems was often not followed up by any inclusion in the meta-analysis, with the quality issue being considered in a narrative way by discussing the overall quality of the dataset.

There is evidence to indicate that quality data should not be included quantitatively in a meta-analysis. For example, a study of Cochrane meta-analyses compared the results of meta-analyses where the studies were divided by means of a quality score into 'high' and 'low' quality (Herbison *et al*. 2006). None of the 45 quality scoring systems succeeded in selecting high quality trials in such a way that a meta-analysis of smaller studies agreed with the results of a large trial about 70% of the time. The outcomes had to be binary, and the majority of meta-analyses included in the review were largely concerned with efficacy, although some meta-analyses did include adverse events within a range of outcomes. The authors concluded that quality scores do not provide a useful means of introducing quality of primary study into a meta-analysis, and that other quantitative methods are now required. Such methods for studies of adverse events outcomes would also be very useful when conducting meta-analyses.

It is another area for further research to investigate how quality data may be used, beyond a straightforward sensitivity analysis, whereby poorer quality studies are excluded. Also, quality assessment scores specifically for use with primary studies (both trials and observational) for adverse events data would be very beneficial.

With regard to quality scoring for adverse events studies, it would be important to assess issues such as:

1. whether the adverse events to be recorded were determined *a priori*;
2. whether there was a protocol for including any adverse events that arose unexpectedly;

3. whether participants who were lost to follow-up due to adverse events were recorded as such;
4. whether adverse events were reported for specific subgroups, e.g. by age or sex; and
5. method of case definition for adverse events.

These aspects of a primary study could be used in conjunction with other quality issues to form a scoring system or facilitate qualitative judgement.

6.5.9 Dealing with sparse data

Possibly the most disappointing aspect of the review was that methods for dealing with studies with sparse events were often poorly described. Many studies with double-zero cohorts were excluded from the meta-analysis, which is an appropriate statistical approach when using a ratio outcome measure. In other cases, studies with zero events in total were apparently 'forced' into the analysis by the use of continuity corrections. This inappropriate use of continuity corrections calls into question the validity of results, especially where the outcome variable occurs infrequently. From a statistical perspective, creating spurious events where none previously existed, is a dubious practice when performed in studies where zero events occurred, and from a clinical perspective, if an adverse event is extremely rare in a control arm, the use of continuity corrections for a control arm with zero events then becomes problematic.

Continuity corrections are intended to be appropriately used in situations where there is one arm of a study with zero events, compared to the other arm with one or more events. Although a continuity correction can be applied to both arms of a study where there are zero events in both arms, to enforce the generation of an outcome estimate (such as an odds ratio) and the associated confidence interval, there are philosophical reasons why this use of continuity corrections is inappropriate.

In the adverse events meta-analyses being reviewed here, it appeared that some primary studies with double-zero events had been inappropriately included by means of continuity corrections. Some meta-analyses had apparently (either stating that this method was used, or leaving it as an assumed method in the absence of any other obvious way of including double-zero primary studies)

used a continuity correction on cells of the 2×2 table to enforce the inclusion into the meta-analysis of one or more double-zero studies. However, in many meta-analyses reviewed, the exact methods used to address the issues of sparse data were very difficult to decipher from the methods described and it is clearly invalid to make assumptions regarding the statistical methods that may have been used.

Further developments in how to incorporate data from trials with zero events across the whole study would be highly desirable in the field of adverse events, where many outcomes are inevitably uncommon but can have serious clinical consequences when they do occur. Such methods would enable all information regarding a rare adverse event to be used, without the exclusion of studies with zero events, which may provide additional valuable information regarding the overall picture of adverse events in a particular clinical situation.

In several meta-analyses, the description of the statistical methods made it difficult to determine whether continuity corrections had been used, or whether alternative methods that could allow an analysis of studies with zero events in only one arm, without a continuity correction, such as the Peto method, had been used. If, however, a random effects model was used, it would be reasonable to assume that a continuity correction must have been employed as there are no random effects models that can cope with single-zero studies without the use of a continuity correction.

Development of methods for the accurate inclusion of data from studies where no events occur is a priority area for meta-analysis of adverse events data, where sparsity of events is a commonplace occurrence. These methods could involve continuity corrections, use of outcome metrics that are not ratio measures (and can therefore provide estimates of outcomes when there are zero events without recourse to adding continuity corrections, although such methods may be required to calculate standard errors and confidence intervals), or Bayesian methods.

An evaluation of statistical methods for meta-analysis of rare events found propensities for bias with some of the more standard methods, such as the inverse variance and DerSimonian & Laird methods (Bradburn *et al.* 2007). The Peto method was the least biased at the lower event rates that would be anticipated for adverse events. However, when considering trials with un-

balanced treatment arms, other methods such as logistic regression, Mantel–Haenszel without correction for cells with zero events, and exact methods, in fact performed better than the Peto method for lower event rates, indicating the necessity of considering the trial numbers when selecting an analysis method. These results are discussed further in Chapter 5.

The use of different forms of continuity correction for studies with one arm having zero events has also been investigated (Sweeting *et al*. 2004), concluding that studies with zero events in total should be excluded from analyses as they do not add any information to an analysis with a ratio outcome. The use of Bayesian methods would also be valuable for use with this sort of data (Spiegelhalter *et al*. 2004). Although Bayesian methods have been used in some studies, there is wide scope to extend Bayesian methods where appropriate, for example by using differing prior distributions for the parameters of interest.

6.5.10 Other reviews of previous meta-analyses

A systematic review of reviews and meta-analyses of primary studies of adverse effects of a drug intervention has been conducted elsewhere (Cornelius *et al*. 2009, introduced in Section 6.1.1). A total of 43 reviews, published in 2006, were retrieved. Of these, 15% assessed quality of primary studies, compared to 42.2% of meta-analyses in this review. Of the 43 reviews included, only 24 performed a meta-analysis. As seen in this review, there was some poor reporting of the methods used for pooling data, but 83% did report the method used for pooling data and exploring heterogeneity. With regard to funding source, 23% (of the 43 revews) had pharmaceutical funding, compared to 16.3% of meta-analyses in the current review that had commercial funding.

Of 22 meta-analyses reviewed by Cornelius *et al*. (those with case–control studies were excluded), 11 (50%) included studies with zero events in one study arm. Of these 11 meta-analyses, seven did not report what continuity correction was used [it is unclear if Cornelius *et al*. assumed that a continuity correction was in fact being used, as other methods could be used in such a circumstance, such as the Peto method, as discussed in Section 3.3.5 and Section 5.2.1, or if it was stated in the review that a continuity correction was used, but the precise value was not provided.]. Two studies used a continuity correction of 0.5, and a further one used a continuity correction of 0.25. One meta-analysis used a continuity

correction proportional to the inverse of the size of the opposite study [arm]; it is assumed that this refers to the continuity correction developed by Sweeting *et al*. (2004). Of eight meta-analyses including primary studies with zero events across both arms, five were excluded from the analysis, one was included with a continuity correction of 0.25 assumed, and for two meta-analyses it was not reported whether the studies with zero events were included. By comparison, sparse data was considered by 39.2% of all studies in the current review (sparse data issues are discussed further in Sections 6.4.7 and 6.5.9). A graphic means of representing the results was used by 75% of the meta-analyses, similar to the proportion of meta-analyses that used some form of graphical representation in the current review. The review by Cornelius *et al*. (2009) is of interest as it highlights the importance of improved reporting in all aspects of meta-analyses regarding adverse events data.

A recent systematic review of systematic reviews and meta-analyses of data derived from animal experiments found that simple quantitative methods of data combination were being used, such as an unweighted mean or median seen in 12 out of 46 meta-analyses (Peters *et al*. 2006). By contrast, in this review of 166 meta-analyses only one study used a simple unweighted mean [S2]. Both this review and the review of meta-analyses of animal data found that many studies took a casual attitude to publication bias, often failing to investigate this issue thoroughly. In both reviews of meta-analyses, the identification and investigation of heterogeneity was not well addressed in many of the meta-analyses being scrutinised, possibly indicating that this is a methodological area often not well understood by many researchers, or else the importance of making an analysis of heterogeneity may not be sufficiently appreciated.

A review of meta-analysis methodologies found that the quality of meta-analysis conduct had improved during the period of the review, from 1993 to 2002 (Gerber *et al*. 2007). There is no specific mention of the type of outcome required, but as it was a criteria that at least five controlled trials should be included, it seems probable that the outcomes were primarily if not exclusively based on efficacy. By studying a total of 272 meta-analyses, largely in general medicine but including some journals from medical specialties, the authors found improvements in search strategies, for example increased numbers of databases being searched, more studies using hand-searching methods, and more studies making an effort to look for grey literature.

Also increasing over time were the use of quality assessment, in particular concealment of allocation, and the inclusion of a test for heterogeneity. The inclusion of IPD did not however, appear to increase over time, nor did the use of sensitivity analyses or data extraction by more than one reviewer. The use of fixed and random effects appeared to fluctuate over time with no clear pattern in their usage. The authors also found that the quality of meta-analyses was higher in general medical journals rather than the specialist journals.

Another review also supports the argument that the quality of meta-analyses is improving over time (Wen *et al*. 2008). Including a random sample of 161 systematic reviews and meta-analyses to be found on Medline and published between 2000 to 2005, it was found that the mean QUOROM (Quality of Reporting of Meta-analyses; Moher *et al*. 1999) score increased over time from a mean of 10.5 (95% CI 8.8; 12.1) in 2000, to 13.0 (95% CI 12.2; 13.8) in 2005. Specifically regarding quantitative synthesis it is reassuring that in this sample of studies, the proportion that fulfilled the QUOROM criteria either completely or partially increased over time. The authors also noted that Cochrane reviews appeared to have higher QUOROM scores than journal articles, with a mean of 14.2% (95% CI 13.9; 14.5) compared to 11.7% (95% CI 11.3; 12.1). The above review included only meta-analyses of RCTs, so it is likely that the majority of outcomes were efficacy-related. Indeed, the majority of their studies were regarding treatment, prognosis or prevention, with only 5.6% of outcomes categorised as 'Other', which may include adverse events.

Although this present review has not scored the studies using a system based on or similar to the QUOROM system, it is useful to refer back to Table 6.7, to see if there are any noticeable changes over the time of the review in the proportion of meta-analyses considering specific facets of methodology. Overall, heterogeneity appears to have been considered in a high proportion of meta-analyses with no obvious increase across the time of this review, whereas there does appear to have been some increase in the proportion of meta-analyses that performed some type of quality assessment in later publications.

As regards dissemination bias, there is little evidence of an increase in consideration over time, although it should be borne in mind that the number of meta-analyses performed in each year is relatively small and that in later years many of the studies were Cochrane reviews which are usually performed to a high standard.

Overall, the field of meta-analysis methods for adverse events data has many aspects of intrinsic statistical interest, and can be developed with the aim of promoting better use of available data to support clinical practice. It is possibly in the most fundamental areas of meta-analysis such as the choice of meta-analysis model and the examination of heterogeneity and publication bias that improvements in methodology may lead to increased validity of results.

Besides the areas concentrated on in this review, there are of course many other areas where further research and improvements to existing methods would be valuable. For example, class effects would be one area where Bayesian methods would be particularly useful in allowing borrowing of strength across drugs from the same class. Other areas that would be strong candidates for further investigation include dose–response analysis, how to cope with multiple outcomes, and time course effects when appropriate, as these issues appear frequently and also have a high relevance to clinical practice and decision-making.

6.6 Summary

This chapter reviews 166 meta-analysis studies where the primary outcome is an adverse event resulting from some form of clinical intervention. Many aspects of these studies are recorded and assimilated, for example, information about the types of intervention and adverse events, also statistical information regarding meta-analysis methods, use of graphs, handling of issues such as publication bias, heterogeneity, sparse data, IPD and quality assessment.

This review has highlighted areas where there is room for further development of meta-analysis methods for adverse events data, especially selection of outcome metric, selection of methods and model, sparse data issues, use of IPD, quality assessment, issues surrounding retrieval and selection of primary references, and aspects of dissemination bias.

7

Comparison of multiple meta-analysis methods using a dataset with sparse events

7.1 Introduction

To investigate the influence of meta-analysis methods on the results obtained, multiple analyses of the same dataset can be performed to contrast the results from different methods. The following aspects of meta-analysis are of interest in this chapter:

1. comparison between different outcome metrics;
2. comparison between fixed effect and random effects models;
3. comparison between different fixed effect models;
4. comparison of models including studies with zero events in total against those excluding them;
5. comparison of different continuity corrections to allow inclusion of sparse events, within and between meta-analysis models;
6. comparison between standard meta-analysis methods and generalised linear methods;

7. comparison between Bayesian and frequentist approaches; and
8. evaluation of different elements of a Bayesian model, such as use of different prior distributions.

These analyses can be used to compare the results of different analysis methods. Whilst it is difficult to determine which models and methods may be most suitable under different circumstances, due to the lack of a 'gold standard' against which to evaluate the different methods, such an approach provides insight into the range of potential results, and any specific patterns regarding method and outcome.

7.2 Clinical example

Many of the issues surrounding adverse events meta-analysis can be illustrated by the use of a dataset from GlaxoSmithKline (GSK). This dataset comprises data from 19 trials of the antidepressant paroxetine, one of the selective serotonin reuptake inhibitors (SSRIs).

Concerns regarding the safety of SSRIs in several areas have been raised since the early 1990s, in particular regarding suicidal ideation and behaviour (Teicher *et al*. 1990). In particular, the potential association between SSRIs and suicidal behaviours in children and adolescents was of particular cause for alarm (Gunnell *et al*. 2005; Gibbons *et al*. 2006; Hammad *et al*. 2006). In response to these concerns, the manufacturers of paroxetine (marketed as Seroxat® in the UK and Paxil® in the USA) were encouraged to publish the results of multiple trials involving paroxetine. These results are available on the GSK website[1].

The concern regarding paroxetine and the possibility that information was not being made available from clinical trials led to a Panorama investigation by the BBC, first broadcast on 29 January 2007[2].

This clinical example is appropriate as there are multiple trials with zero events, hence it is very important to analyse the data using methods to ensure that as much information as possible is included in the model. Should information be

[1] GSK (2006). Available [January 2010] at: `http://www.gsk.com/media/par_current_analysis.htm`

[2] BBC: (2007). Available [January 2010] at: `http://news.bbc.co.uk/1/hi/programmes/panorama/6291773.stm`

lost, or included inappropriately, then incorrect conclusions could be drawn from the model, with deleterious consequences, such as the risk of suicidal ideation or behaviour in patients, or alternatively, the risk that a potentially beneficial treatment may be denied patients unnecessarily.

7.3 Methods

7.3.1 Data extraction

Each of the 19 trials is described in the GSK report on their website (Footnote 1, page 150), with relevant information set out for each trial set. The dataset used in these analyses is based on the data for patients diagnosed with Major Depressive Disorder (MDD), with outcomes of Definitive Suicidal Behaviour and Ideation. The data are displayed in Table 7.1, for a total of 19 trials. These trials include adults only; however, the youngest age range for many of the trials is 18–24, indicating that the results for this age group may be relevant to adolescents if not pre-adolescent children.

7.3.2 Statistical methods

Standard meta-analysis methods were the first line of approach to this dataset. A marginal analysis (Section 3.2), using the relative risk (RR) and odds ratio (OR), to act as a benchmark comparator for the results of the other meta-analysis methods.

Comparisons were made using a variety of outcome measures, the RR, OR and risk difference (RD), and using different fixed effect (FE) methods, including inverse variance (IV), Mantel–Haenszel (M–H) and Peto (all of these outcome metrics and methods are described in Chapter 3). To contrast with the FE models, random effects (RE) models were also used with the RR, OR and RD. The standard continuity correction (where required) was 0.5; when added to all four cells of the 2×2 table this can be considered as adding two extra participants, who are effectively 'false' participants.

For comparison across different continuity corrections, a smaller value of 0.05 was used, and a continuity correction based on that described by Sweeting *et*

Table 7.1: Dataset of GSK trials for adult suicidality analysis. Outcomes: Definitive Suicidal Behaviour and Ideation. Indication: Major Depressive Disorder.

Trial no.	Paroxetine			Placebo		
	No. in arm	No. with event	No. without event	No. in arm	No. with event	No. without event
279	21	2	19	10	1	9
2	170	1	169	171	2	169
9	421	5	416	53	0	53
3	241	0	241	244	2	242
115	283	5	278	117	3	114
128	357	8	349	140	2	138
251	125	2	123	129	0	129
448	212	3	209	103	0	103
449	223	1	222	110	0	110
487	214	2	212	109	0	109
625	112	1	111	117	0	117
785	197	1	196	105	0	105
810	306	0	306	148	1	147
276	20	0	20	21	0	21
274	22	0	22	23	0	23
1	25	0	25	25	0	25
442	41	0	41	48	0	48
NKD20006	124	0	124	125	0	125
874	341	0	341	180	0	180

al. (2004), discussed further in Section 5.2.3, in this example with the overall number of 'false' participants across both trial arms summing to two. The continuity corrections were used in their appropriate context, to allow the inclusion of studies with zero events in one arm, and also used inappropriately, to enforce inclusion of studies with zero events in total into the meta-analysis. Only the 0.5 continuity correction was used in this way, and for OR outcomes only.

Generalised linear models (GLMs) were also used, with unconditional likelihood and a logit link (as the outcome was binary). GLMs were intended as a comparator to standard meta-analysis methods as they allow the inclusion of studies with zero events without manipulation of the data, and also allow an estimate of treatment effect without weighting of the studies, which may be illuminating in itself.

The novel 'exact' method developed by Tian *et al*. (2009), described in Section 5.2.6, is also included for comparison purposes against standard methods.

Finally, Bayesian methods (as discussed in Chapter 4) were used to perform a meta-analysis, again, with the advantage of being able to incorporate primary trials with zero events without recourse to continuity corrections or other artificial manipulation. All Bayesian analyses were conducted using WinBUGS 1.4. One chain was used, with an initial 'burn-in' period of 10 000 iterations, followed by a sample size of 50 000 iterations. The history plot was checked for convergence prior to discarding the initial 10 000 iterations, as well as checks for autocorrelation.

A standard Bayesian meta-analysis of the 19 studies was performed, with the intention of using non-informative prior distributions. In the light of concerns regarding the ability of the prior distribution to unduly influence the posterior distribution in cases where events are sparse (Lambert *et al*. 2005, discussed further in Section 4.3.4), multiple prior distributions were placed on the between-studies variance (or a function thereof), with 12 of these being derived from the same reference (Lambert *et al*. 2005). (These 12 prior distributions are set out in Table 7.6 in Section 7.4.5).

Finally, prior distributions were derived from previous studies using paroxetine that were conducted in children, with an indication of major depressive disorder. These studies were used in a meta-analysis by Kaizar *et al*. (2006) regarding suicidality (suicidal behaviour and ideation) and anti-depressants in children, also

Table 7.2: Dataset of trials using paroxetine in children with major depressive disorder, extracted from Kaizar *et al*. (2006)* Figure 2.

Trial no.	Paroxetine			Placebo		
	No. in arm	No. with event	No. without event	No. in arm	No. with event	No. without event
329	97	4	93	89	1	88
701	106	2	104	103	1	102
377	186	6	180	97	2	95

*Kaizar *et al*. (2006). Do antidepressants cause suicidality in children? A Bayesian meta-analysis. *Clinical trials* 3(2), 73–98.

Table 7.3: Results derived from dataset of trials using paroxetine in children with major depressive disorder, extracted from Kaizar *et al*. (2006)* Figure 2.

Outcome	Mean	Standard deviation	Median (95% CI)
Log OR (μ)	0.9169	0.8171	0.8933 (-0.3928; 2.393)
Standard deviation	0.3936	0.8356	0.1811 (0.0286; 1.991)

OR: odds ratio; *Kaizar *et al*. (2006). Do antidepressants cause suicidality in children? A Bayesian meta-analysis. *Clinical trials* 3(2), 73–98.

using Bayesian methods. These studies were selected out of the many primary studies used by Kaizar *et al*. (2006), as they used paroxetine in participants with pre-existing depression. The primary studies themselves are referenced by Kaizar *et al*. (2006); the dataset, comprising just three studies, is set out below in Table 7.2.

Using these three studies, values could be derived using a Bayesian meta-analysis (with non-informative priors, using a burn-in of 10 000 iterations and a sample of 50 000 iterations) for the underlying value for the mean log OR (μ) in the treatment arms compared to the controls, and the between-studies standard deviation (for the underlying mean log OR). The relevant results are given in Table 7.3.

These results can be used to inform a set of priors on both μ and on the between-studies standard deviation. These can be used together or with standard non-informative prior distributions on the parameter not being derived from the dataset.

The prior distribution on μ is based on a normal distribution with a mean of μ as derived from the dataset in Table 7.3, which is 0.9169. The standard deviation of μ is 0.8171, so the variance is 0.8171^2.

Hence, we have

$$\mu \sim \text{Normal}(0.9169, 0.8171^2). \tag{7.1}$$

For the prior distribution on the between-studies standard deviation (τ), a half-normal distribution was selected, centred on 0. From the analysis of the studies in children, the median of the standard deviation was 0.1811. For a random pair of ORs selected from a distribution, the median ratio of the maximum to minimum OR is exp(1.09τ). Accordingly, the difference between two log ORs randomly selected will be 1.09τ (Spiegelhater *et al*. 2004). Therefore, the median of τ will be 1.09c, where c is the standard deviation of τ. Hence, c=0.1811/1.09=0.17, using the result from Table 7.3. The variance of τ is equivalent to c^2.

This gives a distribution on τ as follows:

$$\tau \sim \text{Half-normal}(0, (0.1811/1.09)^2). \tag{7.2}$$

This method of calculating the standard deviation of τ is intended to produce a half-normal distribution which has the same median value as derived from the previous meta-analysis by Kaizar *et al*. (2006).

These prior distributions can be used together or with alternative, non-informative, priors on the other parameter, for example, on μ we have

$$\mu \sim \text{Normal}(0, 10\ 000), \tag{7.3}$$

and on τ

$$\tau \sim \text{Half-normal}(0, 1). \tag{7.4}$$

7.4 Results

7.4.1 Initial data inspection

There are 19 trials, the smallest of which has 31 participants, and the largest 521 participants. Across all trials, there were 3455 participants who received paroxetine and 1978 participants who received a placebo (with 5433 participants in total). In the paroxetine arms there were 31 events (corresponding to 0.90%), and the placebo arms only 11 (0.56%). It is immediately evident that suicidality events were rare in both groups, but in terms of raw percentages, and without taking into account the different trials, these events are almost twice as frequent in the participants receiving paroxetine. However, with such small numbers, the play of random chance militates against the discovery of a strong association between paroxetine and suicidality.

Using a marginal analysis, the RR for the paroxetine group compared to the placebo group was 1.613 (95% CI 0.813; 3.203). For the OR, the equivalent result was 1.619 (95% CI 0.812; 3.228). In this example, the OR and RR are similar, due to the small number of events.

As can be seen from Table 7.1, there are six studies with no events across the two arms, while nine studies have zero events in one arm. Only four studies have at least one event in both arms. This situation immediately requires some thought prior to analysis. If all studies with no events are excluded, then this effectively excludes data from 995 subjects out of a total of 5433, which is 18.3% of the data. Such an exclusion could produce highly misleading results; however, inaccurate or invalid measures taken to address this situation could lead to equally false results. Similarly, if all studies with at least one treatment arm with zero effects are excluded, then data from 4164 subjects will be lost, which is 76.6% of the data. The loss of such a large proportion of the data would clearly be deleterious.

A method or methods are required for incorporating the data from trials where one or both arms have zero events, and that can produce valid conclusions from the analysis. This produces an immediate question: how can the concept of validity be determined in this context? Given that different analysis methods will produce different results, how is it possible to evaluate which method is

producing results that are the most 'valid' in terms of most closely approximating an 'underlying' reality?

Unfortunately, there is no obvious 'gold standard' for methodology regarding datasets with sparse events. A simulation study (similar to that of Bradburn *et al*. (2007), discussed in Section 5.2.1) based on this dataset may be the most appropriate means of addressing this question.

7.4.2 Standard meta-analysis results

Table 7.4 shows the results of all standard frequentist meta-analyses, with a variety of outcome metrics, meta-analysis methods and continuity corrections.

The forest plots associated with selected models are shown in Section 7.7.1.

7.4.3 Regression results

The results of the two regression models performed are shown in Table 7.5. The first method is unstratified, in that the individual trial is not taken into account in the model. The second method includes trial within the model, and so allows for the influence of trial on the outcome.

For visual comparison purposes, the results of the standard meta-analyses as set out in Table 7.4 and the regression models set out in Table 7.5 are set out in forest plots (note that the pooled estimates and individual 'study' (in this case pooled analyses) estimates are not of interest); see Figures 7.10–7.12, for the OR, RR and RD results respectively. For the OR and RR outcomes, the results of the marginal analyses are included.

7.4.4 'Exact' results

The method of Tian *et al*. (2009), described in Section5.2.6, yielded a 95% CI of -0.0072; 0.0082 for the RD, with an associated p-value of 0.827, using inverse variance weightings for the individual studies.

Using StatXact®, the exact test for homogeneity of ORs (Zelen's test; Section 3.4) yielded a p-value of 0.5658, which effectively indicates no evidence

Table 7.4: Results for frequentist analyses of GSK trials for adult suicidality analysis. Outcomes: Definitive Suicidal Behaviour and Ideation. Indication: Major Depressive Disorder.

Analysis ID	MA method	Outcome metric	Continuity correction	0-events studies excluded	Estimate	95% CI
A	IV	RR	0.5	Yes	1.085	0.557; 2.116
B	M–H	RR	0.5	Yes	1.124	0.609; 2.072
C	IV	OR	0.5	Yes	1.090	0.552; 2.152
D	M–H	OR	0.5	Yes	1.126	0.605; 2.094
E	Peto	OR	No	Yes	1.324	0.676; 2.596
F	IV	RD	No	No	0.001	-0.003; 0.006
G	M–H	RD	No	No	0.002	-0.004; 0.008
H	D&L(IV)	RR	0.5	Yes	1.085	0.557; 2.116
J	D&L(IV)	OR	0.5	Yes	1.090	0.552; 2.152
K	D&L(IV)	RD	No	No	0.001	-0.003; 0.006
L	D&L (M–H)	RR	0.5	Yes	1.085	0.557; 2.116
M	D&L (M–H)	OR	0.5	Yes	1.090	0.552; 2.152
N	D&L (M–H)	RD	No	No	0.001	-0.003; 0.006
O	IV	OR	0.05	Yes	1.034	0.435; 2.463
P	M–H	OR	0.05	Yes	1.321	0.652; 2.676
Q	D&L(IV)	OR	0.05	Yes	1.034	0.435; 2.463
R	D&L (M–H)	OR	0.05	Yes	1.034	0.435; 2.463
S	IV	OR	SS	Yes	1.142	0.561; 2.326
T	M–H	OR	SS	Yes	1.274	0.673; 2.411
U	D&L(IV)	OR	SS	Yes	1.142	0.561; 2.326
V	D&L (M–H)	OR	SS	Yes	1.142	0.561; 2.326
W	IV	OR	0.5	No	1.065	0.569; 1.994
X	M–H	OR	0.5	No	1.100	0.617; 1.962
Y	D&L(IV)	OR	0.5	No	1.065	0.569; 1.994
Z	D&L (M–H)	OR	0.5	No	1.065	0.569; 1.994

CI: confidence interval; D&L: DerSimonian & Laird; IV: inverse variance MA method; M–H: Mantel–Haenszel MA method; MA: meta-analysis; Peto: Peto MA Method; OR: odds ratio; RD: risk difference; RR: relative risk; SS: Sweeting *et al.* (2004) continuity correction.

Table 7.5: Results for regression analyses of GSK trials for adult suicidality analysis. Outcomes: Definitive Suicidal Behaviour and Ideation. Indication: Major Depressive Disorder.

Regression Model	GLM type	Link	ML type	Group factors	Outcome metric	Estimate	95% CI
1	Binomial	Logit	Unconditional	Treatment	OR (treatment)	1.619	0.812–3.228
2	Binomial	Logit	Unconditional	Treatment, trial	OR (treatment)	1.345	0.660–2.741

CI: confidence interval; OR: odds ratio.

for a difference in odds of suicidality between control and paroxetine treatment groups.

The exact inference of the common (pooled) OR was 0.7443, with a 95% CI of 0.3296; 1.565. Again, there is no evidence to support any difference in risk of suicidality across the two treatment groups.

7.4.5 Bayesian results

The results of the Bayesian analyses (B.1–B.15) are set out in Table 7.6. For all analyses, the prior distribution on the underlying log OR (μ) was:

$$\mu \sim \text{Normal}(0, 10^9), \tag{7.5}$$

apart from analyses B.13 (as above with variance set at 10^4) and B.14 and B.15 where the prior on μ was

$$\mu \sim \text{Normal}(0.9169, 0.8171^2). \tag{7.6}$$

For all analyses, the posterior densities on the between-studies standard deviation are shown in Figures 7.13 and 7.14. The posterior densities on the values of μ on selected analyses are shown in Figure 7.15. These figures are in Section 7.7.3 below. Some evidence of autocorrelation was seen for certain parameters; however, the large sample size would help to reduce the effects of autocorrelation on the sample statistics.

7.5 Discussion

7.5.1 Comparison of different outcome metrics

Three different outcome metrics were considered in the standard meta-analyses: the RR, OR and RD. It was expected that the RR and OR would be very similar, given the small number of events in all studies. This result was borne out in practice; for example, for Analysis A, the RR was 1.085, compared to the

Table 7.6: Results for Bayesian analyses of GSK trials for adult suicidality analysis. Outcomes: Definitive Suicidal Behaviour and Ideation. Indication: Major Depressive Disorder. Prior on mean underlying log odds ratio μ is $\mu \sim$ Normal$(0, 10^9)$ unless otherwise indicated.

Analysis	Prior distribution scale	Prior distribution	Median OR (95% CrI)	P(OR>1)	Median standard deviation (95% CrI)
B.1	Precision	$\sim \Gamma$ $(0.2, 0.2)$	1.591 (0.595; 8.120)	0.831	0.816 (0.288; 3.676)
B2	Precision	$\sim \Gamma$ $(0.1, 0.1)$	1.562 (0.612; 6.511)	0.826	0.726 (0.221; 3.121)
B.3	Log variance	$\sim$ Uniform (-10, 10)	1.412 (0.687; 3.525)	0.829	0.087 (0.008; 1.688)
B.4	Log variance	$\sim$ Uniform (-10, 1.386)	1.347 (0.649; 3.146)	0.797	0.105 (0.008; 1.535)
B.5	Variance	$\sim$ Uniform (0.001, 1000)	2.326 (0.333; 131.9)	0.839	2.479 (0.428; 10.43)
B.6	Variance	$\sim$ Uniform (0.001, 4)	1.668 (0.591; 6.143)	0.836	1.182 (0.220; 1.953)
B.7	Precision	$\sim$ Pareto (1, 0.001)	2.314 (0.341; 100.1)	0.841	2.482 (0.412; 9.631)
B.8	Precision	$\sim$ Pareto (1, 0.25)	1.667 (0.578; 6.075)	0.829	1.197 (0.214; 1.954)
B.9	Standard deviation	$\sim$ Uniform (0, 100)	1.702 (0.547; 35.24)	0.836	1.061 (0.066; 8.272)
B.10	Standard deviation	$\sim$ Uniform (0, 2)	1.577 (0.612; 5.205)	0.822	0.761 (0.042; 1.904)
B.11	Standard deviation	$\sim$ Half-normal (0, 100)	1.315 (0.693; 2.819)	0.797	0.071 (0.004; 0.224)
B.12	Standard deviation	$\sim$ Half-normal (0, 1)	1.480 (0.609; 4.354)	0.829	0.536 (0.022; 1.871)
B.13*	Standard deviation	$\sim$ Half-normal (0, $1/(0.1811/1.09)^2$)	1.340 (0.646; 2.976)	0.782	0.106 (0.005; 0.371)
B.14**	Standard deviation	$\sim$ Half-normal (0, 1)	1.634 (0.788; 3.979)	0.901	0.526 (0.034; 1.812)
B.15**	Standard deviation	$\sim$ Half-normal (0, $1/(0.1811/1.09)^2$)	1.504 (0.801; 2.930)	0.888	0.121 (0.006; 0.370)

*prior on μ is $\mu \sim$ Normal$(0, 10^4)$; **prior on μ is $\mu \sim$ Normal$(0.9169, 0.8171^2)$; CrI: credible interval; OR: odds ratio.

equivalent analysis using the OR (Analysis C), which was 1.090. The meta-analysis method for both of these was the inverse variance (IV) method. A similar scenario was found when the Mantel–Haenszel (M–H) method was used. Analysis B, using the RR, yielded a result of 1.124, compared to an OR from Analysis D of 1.126. It is notable that for both the IV and M–H methods, the RR yielded a lower value (closer to 1) compared to the OR. It would be interesting to investigate whether this finding represents a tendency for the RR to be lower (regardless of whether greater or less than 1) or whether the tendency is for the RR to be closer to 1 (so that the RR would be greater than the OR if both were less than 0).

The RD has a much narrower range of values, for given dataset and meta-analysis methods, compared to the relative outcome metrics. The lack of requirement for a continuity correction to calculate the outcome metric is a benefit, as it allows the inclusion of studies with zero events. Depending on the method used, a continuity correction may be required for studies with zero events in total, or if there are zero events in the overall dataset. Also, the RD is not dependent on baseline risk for its interpretation. This gives the RD some important advantages over the outcome metrics on a relative scale.

The method of Tian *et al.* (2009) yielded similar results for the RD compared with the IV, M–H and DerSimonian & Laird methods, and has the advantage of not requiring a continuity correction to calculate the study variances. However, as discussed in Chapter 5, the RD has some distinct disadvantages when used with sparse data.

7.5.2 Comparison of different fixed effect models

It is notable that the meta-analysis method appears to have a stronger influence on the outcome metric than the choice of metric across different methods. For example, for the RR calculated by the IV method (1.085) is closer to the OR calculated by the same method (1.090) than it is to the RR calculated by the M–H method (1.124) and this occurrence is seen also for the OR. For theoretical reasons the IV method is considered less suitable for sparse events data (Higgins *et al.* eds. 2008) due to its incorporation of the variance into the study weighting, but where events are rare, this method is less appropriate due to high variances associated with fewer events.

The OR calculated by the Peto method (which does not require a continuity correction) is higher than either the IV or M–H methods which do require a continuity correction, although the associated 95% CI is still wide. Bearing in mind that the Peto method has been found to be less biased than other methods (Chapter 5.2.1) it is interesting that it yields the highest OR than any of the other standard meta-analysis methods. The Peto OR is also similar to that produced in Regression Model 2, shown in Table 7.5.

It is notable that the IV method, which is less suitable for sparse events, yielded outcome metrics closer to 1 than the other FE models. It would be interesting to investigate whether the same phenomenon would be seen if the data were reversed to yield outcome metrics that were less than 1 rather than greater than 1.

One relevant point to note is that for all FE models, whether using the Peto method without a continuity correction, or using one of the other methods with a continuity correction, is that the pooled OR (or RR) is consistently greater than 1. This applies also to the analyses that include the studies with zero events, when enforced into the analysis by means of a continuity correction. However, the exact pooled OR was 0.7443, which is clearly less than 1. Despite the fact that the 95% CIs are wide for all methods, the fact that the exact method, which includes all studies with no continuity correction, yields a pooled estimate of less than 1, is significant.

7.5.3 Comparison of fixed effect and random effects models

The results of the DerSimonian & Laird RE models coincide with the results of the IV (FE) model, regardless of the method used to calculate the between-studies variance (IV or M–H). In effect, the weightings are not changed by the use of fixed or random effect(s). This possibly relates to the fact that there are not extreme differences in the sizes of the studies (although the change in order of magnitude between the smallest and largest is approximately 17), and the degree of variability between them.

This example therefore is not a particularly useful one to highlight possible differences in fixed and random effect(s). A dataset with greater variation in the number of participants across studies, and the treatment effects between

studies, would provide a more useful example. Alternatively, simulation studies may also be of benefit.

7.5.4 Comparison of continuity corrections

Three different continuity corrections were considered, using the OR as outcome metric across the IV, M–H and RE models. These continuity corrections were 0.5 added to all cells for a study with zero events in one arm (resulting in two 'false' participants per study), 0.05 added to all cells for a study with zero events in one arm (resulting in 0.2 'false' participants per study), and the Sweeting *et al*. (2004) continuity correction, whereby an arbitrary number of 'false' participants per treatment arm is divided between the 'events' and 'non-events' within each arm. In this example, two 'false' participants per trial was used, analagous with a continuity correction of 0.5 per cell, but dividing the 'false' participants unequally between all four cells of the 2×2 table, but proportionately to the percentage of participants in each arm of the trial, such that the arm with the greater proportion of participants receives the higher number of false participants (see Section 5.2.3).

The 0.05 continuity correction produced a reduced OR for the IV method and all RE models, compared to 0.5 in the same models (1.034 compared to 1.090). However, this result was not repeated for the models using the M–H model, which produced an OR of 1.321 with a continuity correction of 0.05 compared to 1.126 using 0.5. This result appears to be anomalous, and cannot be easily explained.

When using the continuity correction proposed by Sweeting *et al*. (2004), the OR was higher for the IV method and all RE models compared to a continuity correction of 0.5 (1.142 compared to 1.090). This result appears to be consistent with what would be expected, in that fewer 'false' participants are added to the control arm, and more specifically to their number of cases, thus increasing the OR. For the M–H method also, this effect was seen, with an OR of 1.274 using the Sweeting *et al*. (2004) method and an OR of 1.126 with the 0.5 continuity correction. Thus, the Sweeting *et al*. (2004) continuity correction appears to be more consistent in its results, possibly, as argued by its proponents, reducing the bias inherent in using a continuity correction.

7.5.5 Inclusion and exclusion of studies with zero events

The inclusion of studies with zero events is highly debatable for relative outcome metrics. One argument states that studies with zero events do not have their own estimate of the outcome measure and hence contribute nothing to a pooled estimate, and should be excluded (Whitehead & Whitehead 1991). This argument is supported by Sweeting *et al*. (2004), who back up this position by the use of a simulation study using Bayesian methods, which concluded that studies with zero events in total contribute nothing to an FE meta-analysis.

The intuitive counter-argument to this is that these studies contain relevant information and add to the number of participants whose data can be combined, and therefore should not be excluded; to do so would be to discard potentially relevant information in a way that could lead to false conclusions and be detrimental to patients. This argument has been discussed by Sweeting *et al*. (2004), referring to previous studies (Cook *et al*. 1991 and Sankey *et al*. 1996).

In these analyses, the studies with zero events are forced into the analysis by adding 0.5 to all four cells of the 2×2 table. This is commensurate with adding a continuity correction of 0.5 to all cells when one arm has zero events. In this way, a further six studies including 995 participants can be included in the meta-analysis.

Using the IV method, and all RE models, the OR was 1.065, compared with an OR of 1.090 when these studies are not included (but using a continuity correction of 0.5 on studies with zero events in one arm). In this example, the OR is reduced towards 1, which is as expected, since the number of 'false' participants added to the events in the control arms is proportionately higher than those added to the events in the treatment arms. For the M–H model, the OR was 1.100 compared to 1.126 when studies with zero events are excluded. In this example, the differences are not great, but it is feasible that in some cases the exclusion of studies with zero events could make a stronger difference when using relative metrics.

Using the RD, regardless of model, the studies with zero events are included without use of a continuity correction, and the RD values are similar across all models. Interpreting the relative scale and difference scale outcomes in

combination may be the most useful approach to the inclusion and exclusion of studies with zero events.

7.5.6 Standard meta-analysis models compared with generalised linear models

Using logistic regression, all studies can be incorporated in the model, including those with zero events. Another advantage is that the model can include study-level covariates, if such data are available. It is interesting to note that when the trial effect is ignored and only treatment is included in the model (not advisable due to loss of randomisation effects and lack of weighting of trials), that the OR is higher than when the trial is included (1.619 compared to 1.345, 95% CIs as shown in Table 7.5). This is probably due to the smaller trials with higher ORs being 'weighted' equally with larger trials with smaller ORs. The logistic regression model with no trial stratification corresponds to a marginal analysis (yielding identical results).

Comparing these results with those of the standard meta-analyses, the GLMs appear to produce generally higher ORs. However, the OR derived from the model using both trial and treatment closely approximates the results from the Peto model (Analysis E in Table 7.4) and the M–H model using a continuity correction of 0.05 (Analysis P). The 95% CIs are within similar parameters. However, the major difference between the GLM analysis and Analyses E and P is that studies with zero events are included within the GLM, so this may imply that the Peto method and M–H method with a continuity correction of 0.05 are (at least with this particular dataset) producing results that are closer to those using all studies.

7.5.7 Differences between Bayesian models

The results of the Bayesian analyses are displayed in Table 7.6. Of the 15 Bayesian analyses, the median ORs ranged from 1.315 (Analysis B.11) to 2.326 (Analysis B.5). In all cases the 95% credible intervals (CrIs) were wide, with the highest median ORs being associated with the widest CrIs (due to the extremity of the upper bound); these are seen in Analyses B.5 and B.7.

All models have a high probability that the true underlying OR for suicidality in the treatment arms compared to the control arms is greater than 1. The lowest probability was 0.782 (Analysis B.13), with the highest probability being associated with Analysis B.14 at 0.901. It is interesting to note that the two highest probabilities (0.901 and 0.888) were associated with Analyses B.14 and B.15 respectively; these analyses were performed using the prior distribution on the underlying mean log OR (μ) derived from the results of the studies in children, extracted from Kaizar *et al*. (2006). However, inspection of posterior densities of μ (selected densities are shown in Figure 7.15) does not provide any indication of why this might be. Also, these analyses were not associated with the highest median ORs.

The highest OR, seen in Analysis B.7, was associated with a probability that the true underlying OR is greater than 1 of 0.839, which was the highest probability of those distributions not using the data derived from Kaizar *et al*. (2006). Similarly, the three analyses with a median OR of less than 1.4 (Analyses B.4, B.11 and B.13) were associated with the three lowest probabilities that the true underlying OR was greater than 1, all less than 0.8.

It is difficult to discern any specific pattern between the posterior densities of the standard deviations (shown in Figures 7.13 and 7.14) and the results of the ORs. Analyses B.5 and B.7 have similar densities, and also similar ORs, as do Analyses B.6 and B.8. However, the two lowest ORs, produced by Analyses B.3 and B.11, have very different densities, that of B.3 being very narrow and B.11 being much wider. Analysis B.11 is seen to have the wider 95% CrI of the two, which is difficult to explain.

7.5.8 Bayesian models compared with frequentist models

All the Bayesian models fitted are RE, and incorporate all studies including those with zero events. The logical frequentist model against which to compare the Bayesian models is Analysis Y (Analysis Z produced identical results). This model yielded a median OR of 1.065 (95% CrI 0.569–1.994). However, this model utilised a continuity correction of 0.5 across all studies with at least one arm with zero events, which is not required by the Bayesian models. This addition of 'false' participants (and 'false' events) would introduce bias to the

model. This may be the reason why all 15 Bayesian models produced a higher OR than this value (the lowest OR being 1.315 from Model B.11).

In fact, the FE Peto model (Analysis E), which does not use a continuity correction, and Analyses P and T (which use the 0.05 continuity correction and the Sweeting *et al*. (2004) continuity correction respectively), produce the highest ORs of the frequentist model and therefore most closely approximate the Bayesian results. This may be due to reduced bias across these models due to lack of introduction of 'false' participants, or to reducing their numbers, or to ensuring that they are divided in a less biased manner across the treatment arms.

7.6 Conclusions

The over-arching clinical conclusion of all the analyses performed is that paroxetine is associated with increased suicidality; however, the evidence from the frequentist analyses cannot reach statistical significance regardless of the type of analysis, or use of continuity corrections. The relative metrics (OR and RR) always exceed 1, and the RD is always positive. However, wide confidence intervals prevent the derivation of any firm conclusions about risk of suicidality.

The studies with zero events should intuitively be included within any analyses, due to the significant number of studies that fall into these categories. However, the frequentist analyses do not lend themselves to inclusion of such studies, as the use of continuity corrections biases the results. The choice of continuity correction also influences the results; in making a decision regarding which continuity correction to use, the concept of proportionately dividing up the 'false' participants between the treatment arms, as discussed by Sweeting *et al*. (2004) should be considered, as should the scale of the continuity correction itself. As the Peto method can be used for studies with zero events in one arm only, without use of a continuity correction, this method is well-placed to be used as a comparison for the methods that do require a continuity correction.

In the frequentist analyses there was little difference found between fixed effect and random effects. However, it has been suggested that for sparse events data in general, the most important element of an analysis is to receive any signal from the analysis, which outweighs the importance of considering between-

studies heterogeneity (Higgins *et al*. eds. 2008). Hence, an FE model can be considered the most acceptable option. From a clinical point also, the adverse/unintended events can be reasonably thought to be invariant across studies (assuming sufficient homogeneity of patients across studies), so an FE approach is justified.

The Bayesian paradigm offers some immediate advantages compared to the frequentist approach, primarily in the ability to incorporate studies with zero events in one or both arms without the use of a continuity correction. Another advantage is the ability to calculate a probability that an outcome metric lies within a certain range (for example, greater than 1).

One disadvantage is in the appropriate choice of prior distributions. Non-informative priors may unduly exert influence over the data where events are sparse. A range of prior distributions allows comparison across different distributions; there may, however, be some difficulty in relating the influence of the prior distribution to the outcome. This choice of prior distribution can also be turned to an advantage in that it allows the incorporation of data from other studies, usually sufficiently similar to be of relevance to the current analysis, but not sufficiently homogeneous to be included within the analysis itself. Examples of such scenarios could include the same drug being used in a different patient group or for a different indication.

The Bayesian analyses, despite having a wide range of outcome metrics, and wide CrIs, all pointed to a high probability of an OR in excess of 1 for the association between paroxetine and suicidality. This evidence was the strongest signal from the dataset that there is genuine cause to believe that paroxetine does indeed increase risk of suicidality, despite the fact that the increase may not be excessive on a relative scale and bearing in mind the low baseline risk, and low values for the RD derived from the frequentist analyses.

Across all analyses, despite the lack of clear evidence to indicate a strongly increased risk of suicidality due to paroxetine use, the overall signal is that there is some increased risk, which is difficult to quantify due to the sparsity of data in many studies and the overall low level of underlying risk. The need for multiple analyses is highlighted, in particular the usefulness of comparison across a difference and relative scale, and the benefit of simple data scrutiny to yield awareness of the baseline risk and absolute numbers involved.

Given the low power associated with both primary studies and meta-analyses of adverse events data, it may be appropriate to adopt a lower level of statistical significance for these outcomes, for example using the 10% level rather than the usual 5% level. For more severe adverse outcomes, such as risk of suicidality, this approach may be acceptable. For less serious outcomes, such an approach may not be justifiable clinically. In this scenario, there is a balance to be achieved between avoiding adverse events, and yet not wanting to deprive patients of the potential benefits of an intervention due to excessive concern regarding adverse outcomes. The issue is effectively one of decision-making and the balance of harms and benefits – this area is discussed in more depth in Chapter 11.

In the light of these analyses, it is reasonable to be concerned regarding the potential effects on suicidality of paroxetine. Although there is no conclusive evidence, it would be worthwhile to highlight concerns to prescribing clinicians so that they can be aware of possible risks and take this into consideration when prescribing. Even if the decision is made to prescribe the drug due to its potential benefits, prior warning regarding suicidality would be helpful to both patients and caregivers, so that they can be aware that any inclinations towards suicidality may be due to the medication and be vigilant for such an effect.

7.7 Selected graphical results

7.7.1 Forest plots from selected meta-analyses

Forest plots associated with selected non-Bayesian meta-analyses, described in Table 7.4 are shown in Figures 7.1–7.9.

7.7.2 Forest plots of pooled analysis results

Forest plots of the results of selected pooled analyses set out in Section 7.4 are included for the OR, RR and RD metrics.

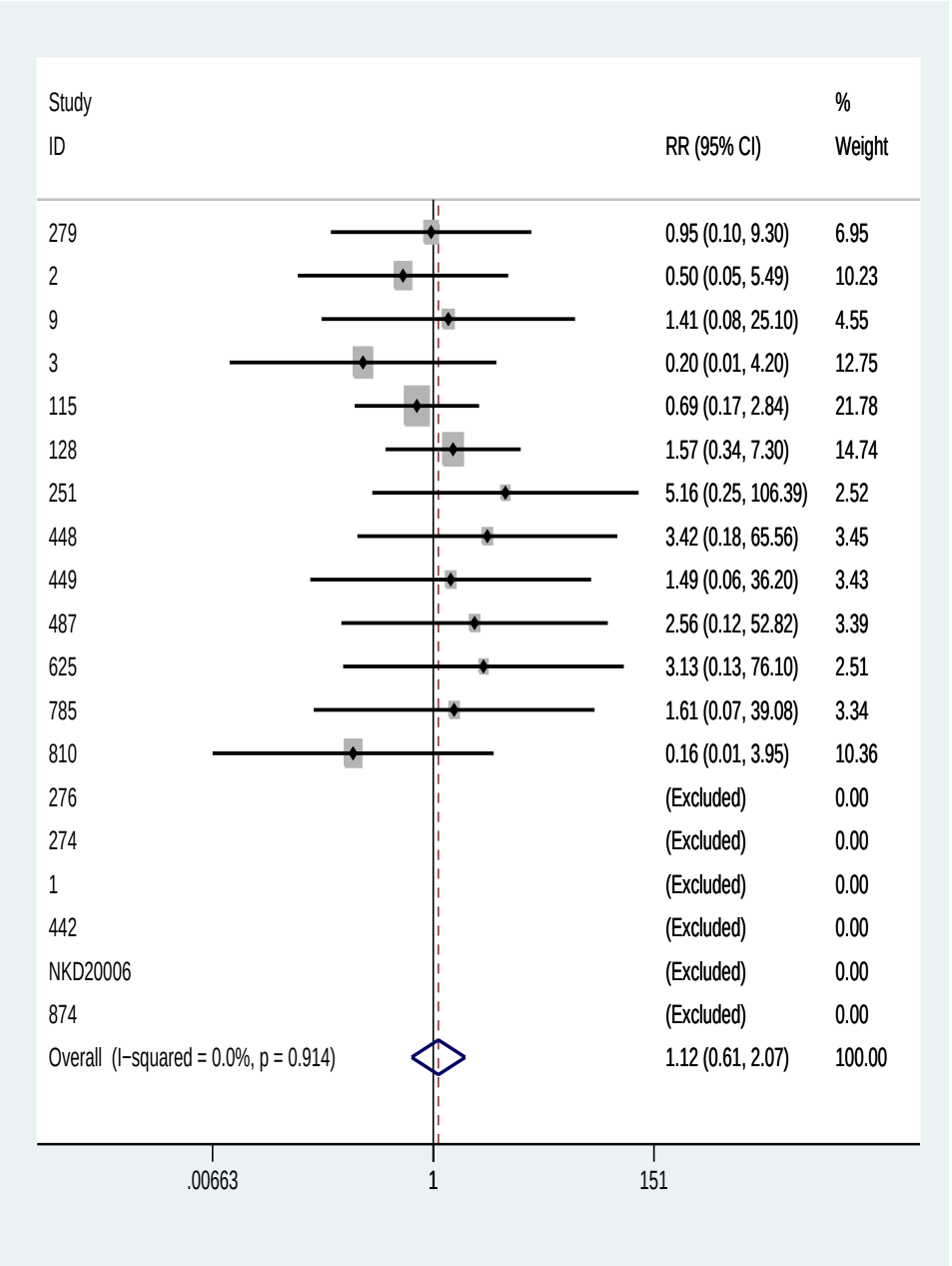


Figure 7.1: Analysis B, Mantel–Haenszel model with relative risk, continuity correction 0.5.

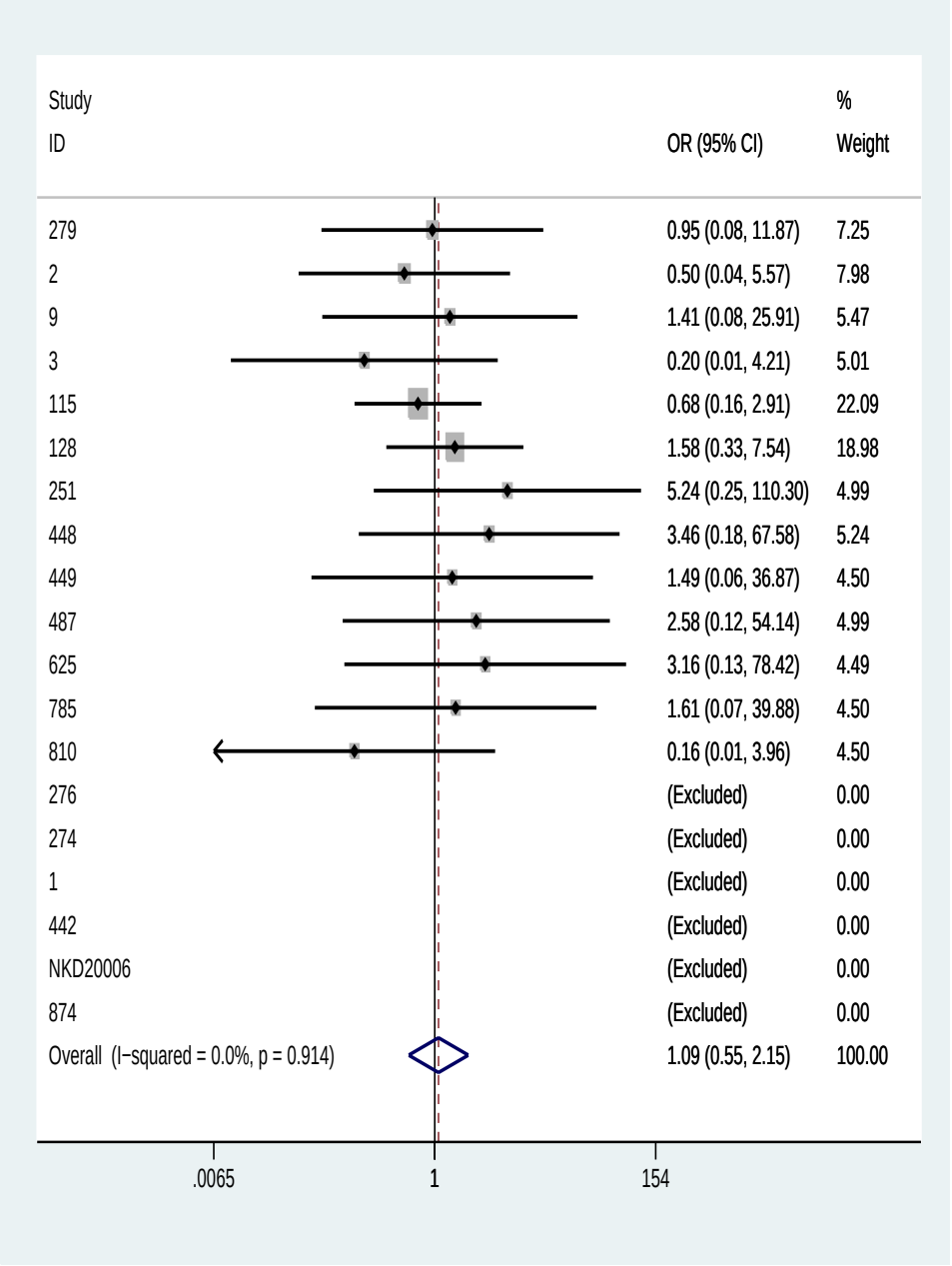


Figure 7.2: Analysis C, inverse variance model with odds ratio, continuity correction 0.5.

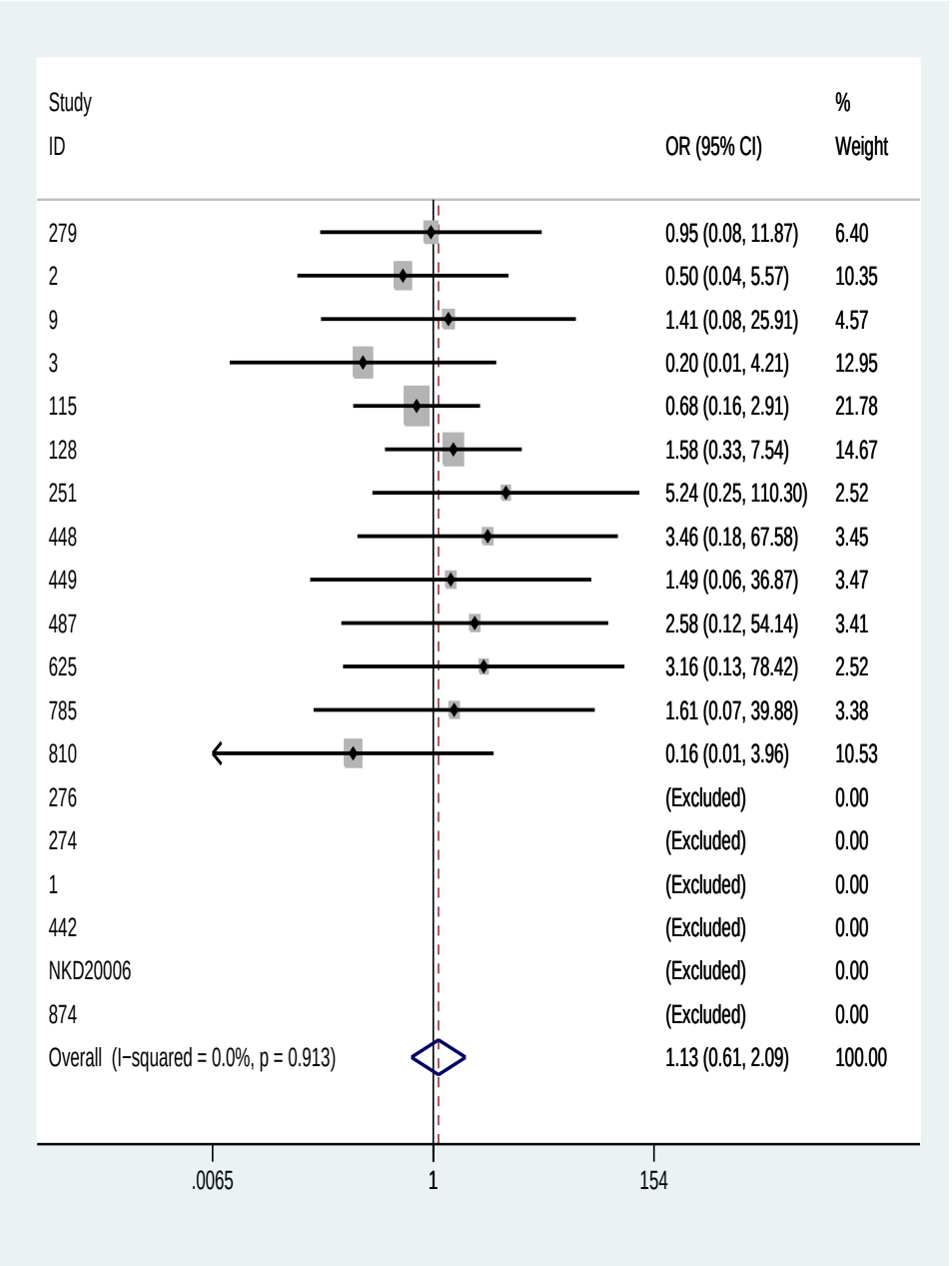


Figure 7.3: Analysis D, Mantel–Haenszel model with odds ratio, continuity correction 0.5.

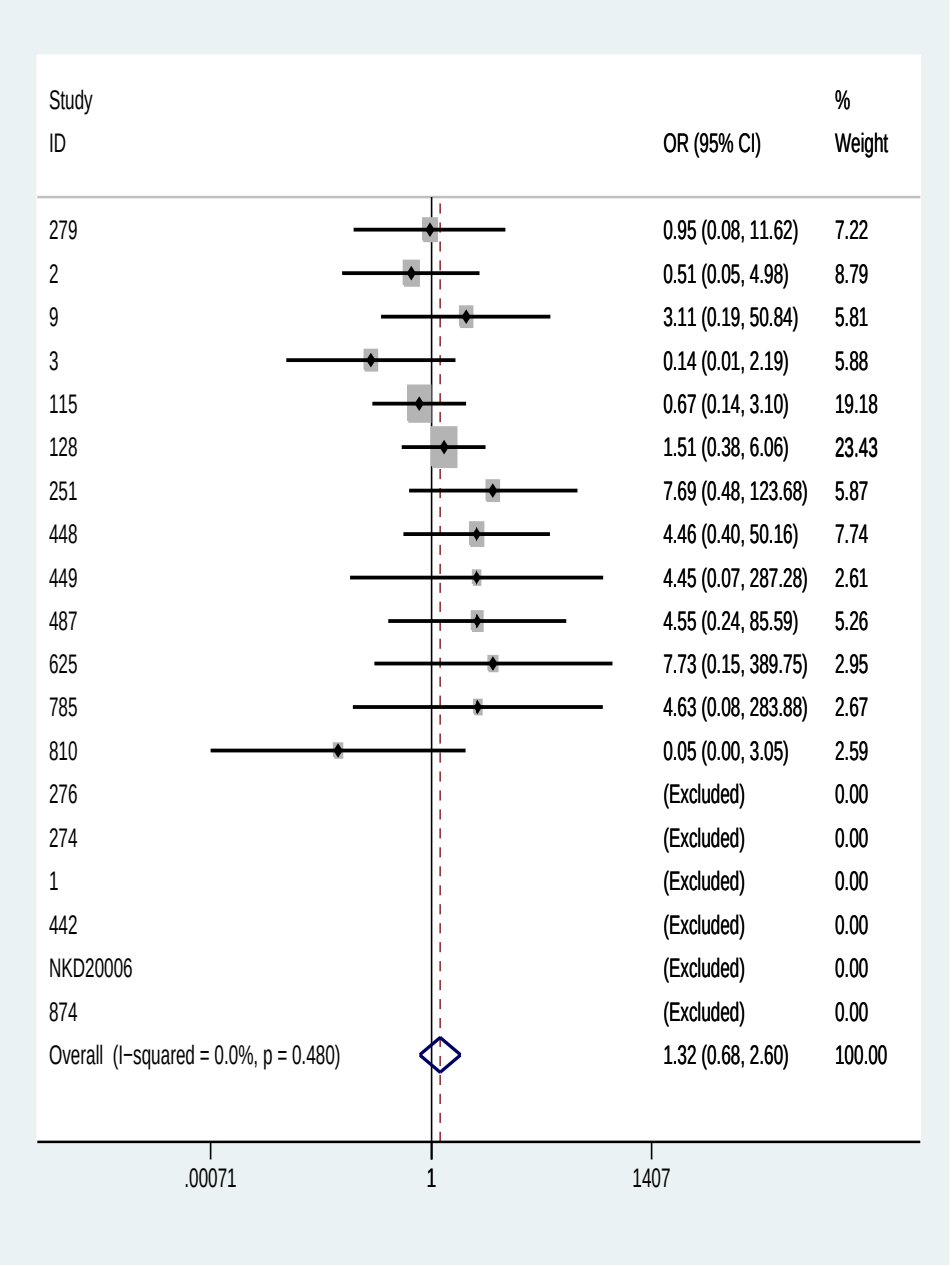


Figure 7.4: Analysis E, Peto model with odds ratio, no continuity correction.

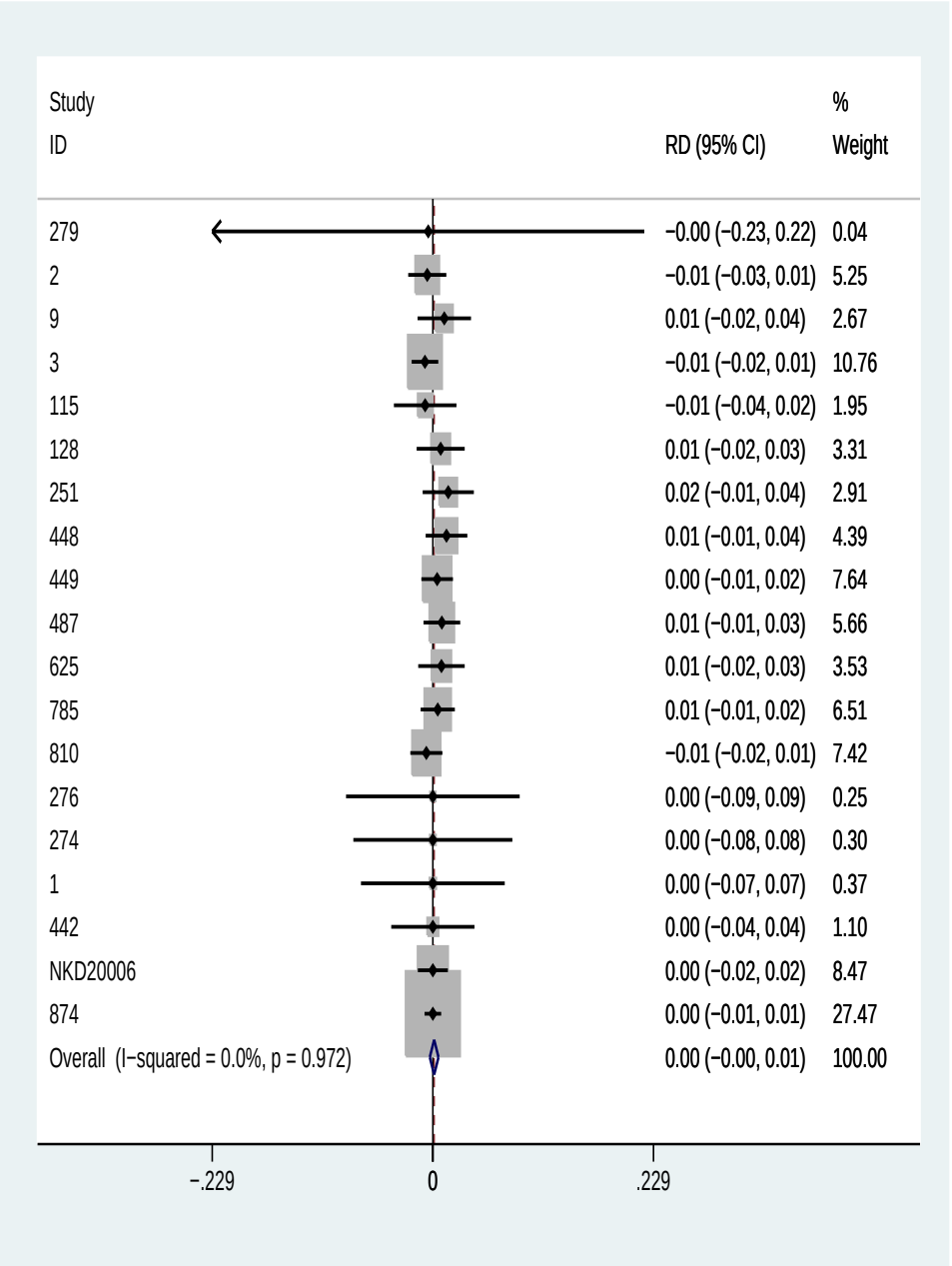


Figure 7.5: Analysis G, Mantel–Haenszel model with risk difference, no continuity correction.

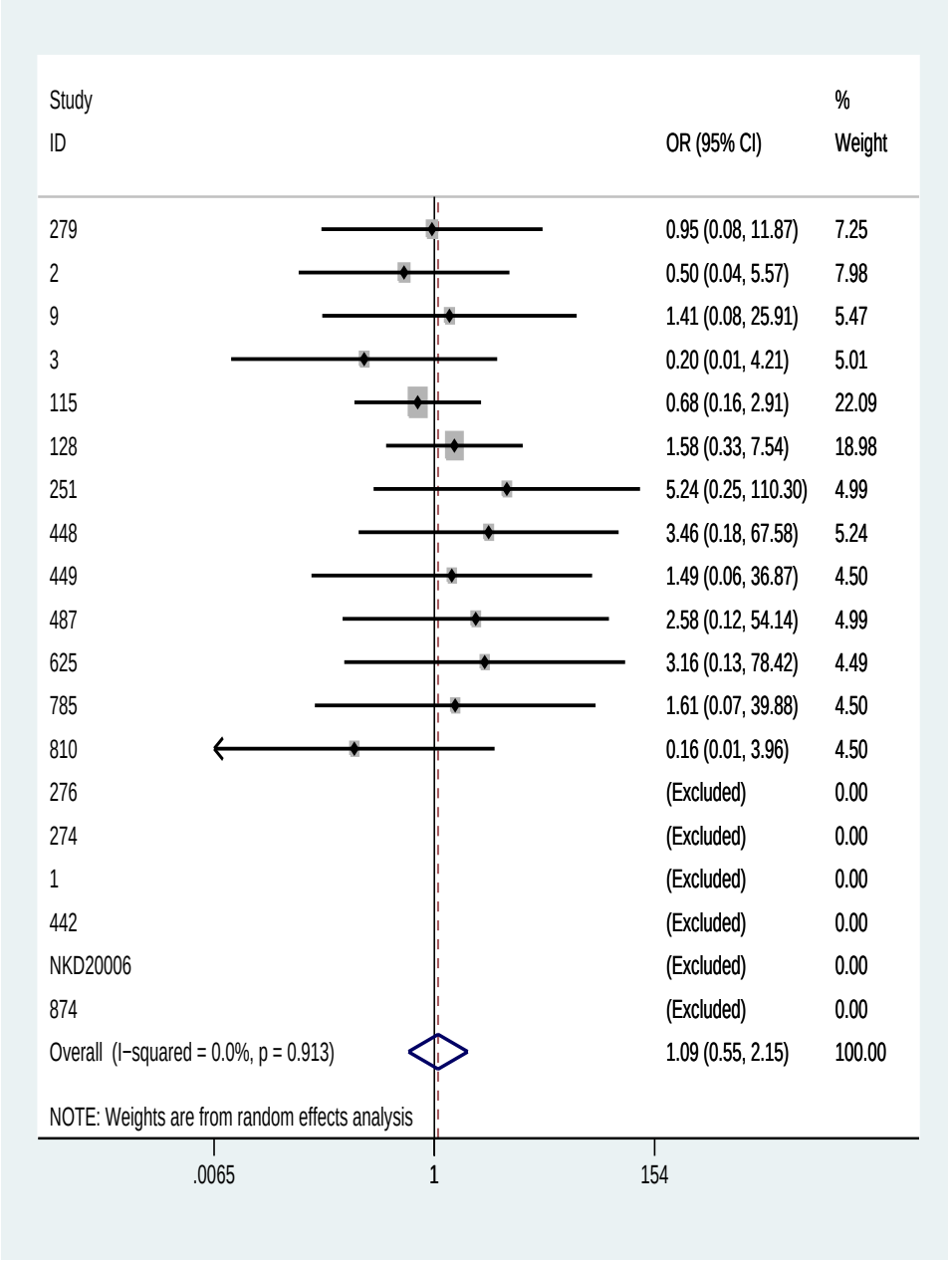


Figure 7.6: Analysis M, DerSimonian & Laird model (Mantel–Haenszel method for calculation of variance) with odds ratio, continuity correction 0.5.

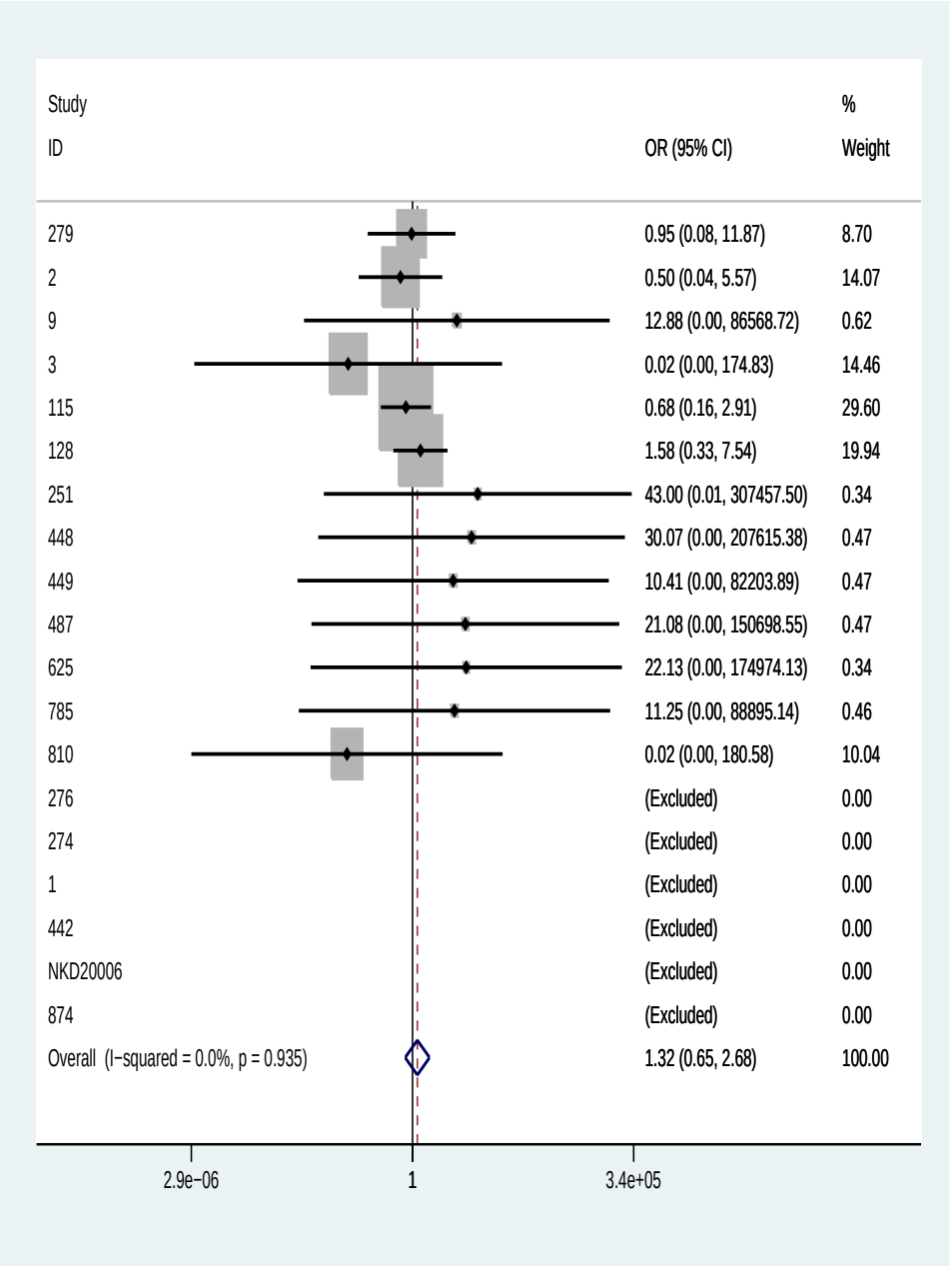


Figure 7.7: Analysis P, Mantel–Haenszel model with odds ratio, continuity correction 0.05.

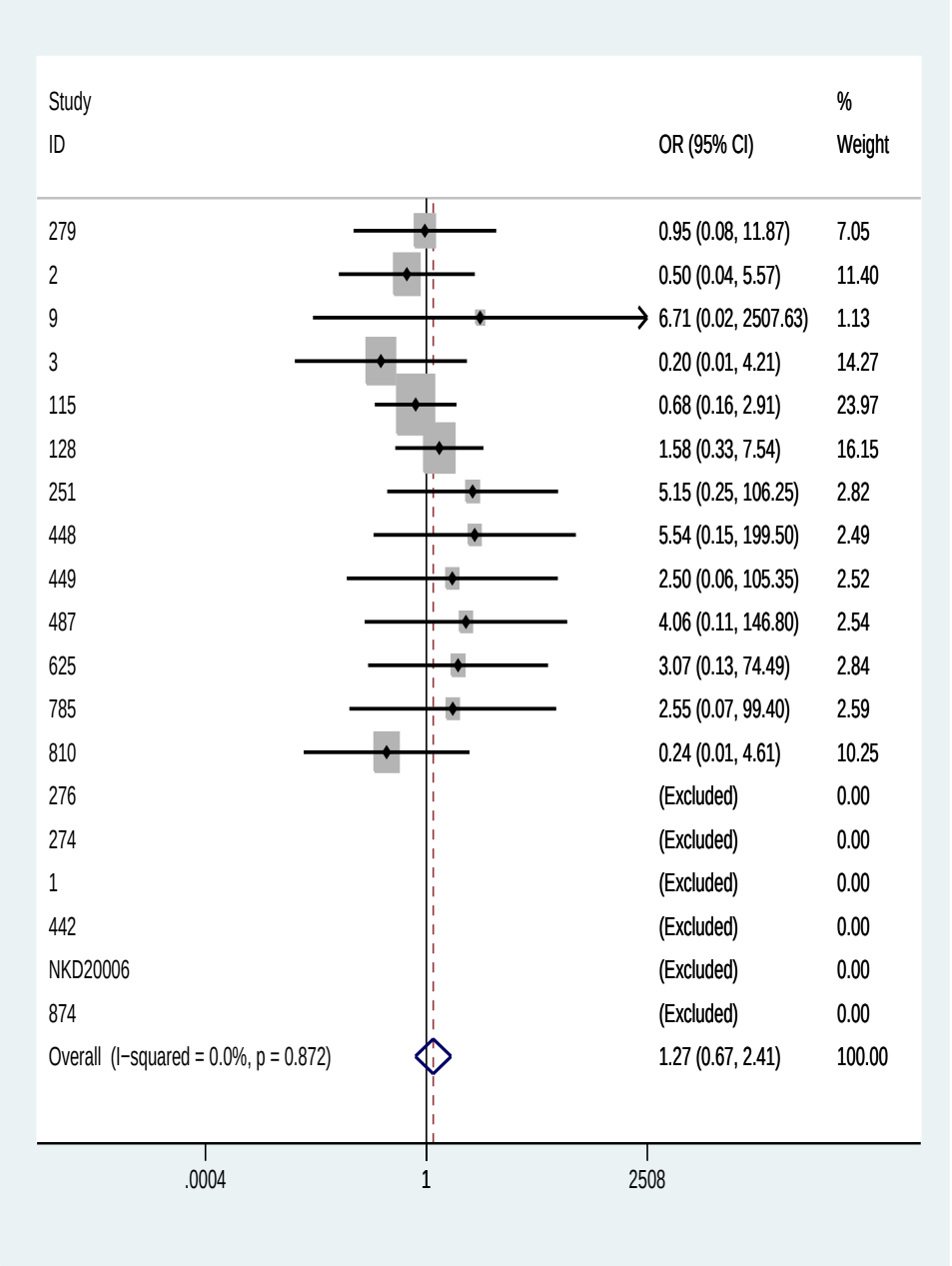


Figure 7.8: Analysis T, Mantel–Haenszel model with odds ratio, continuity correction as described by Sweeting et al. (2004).

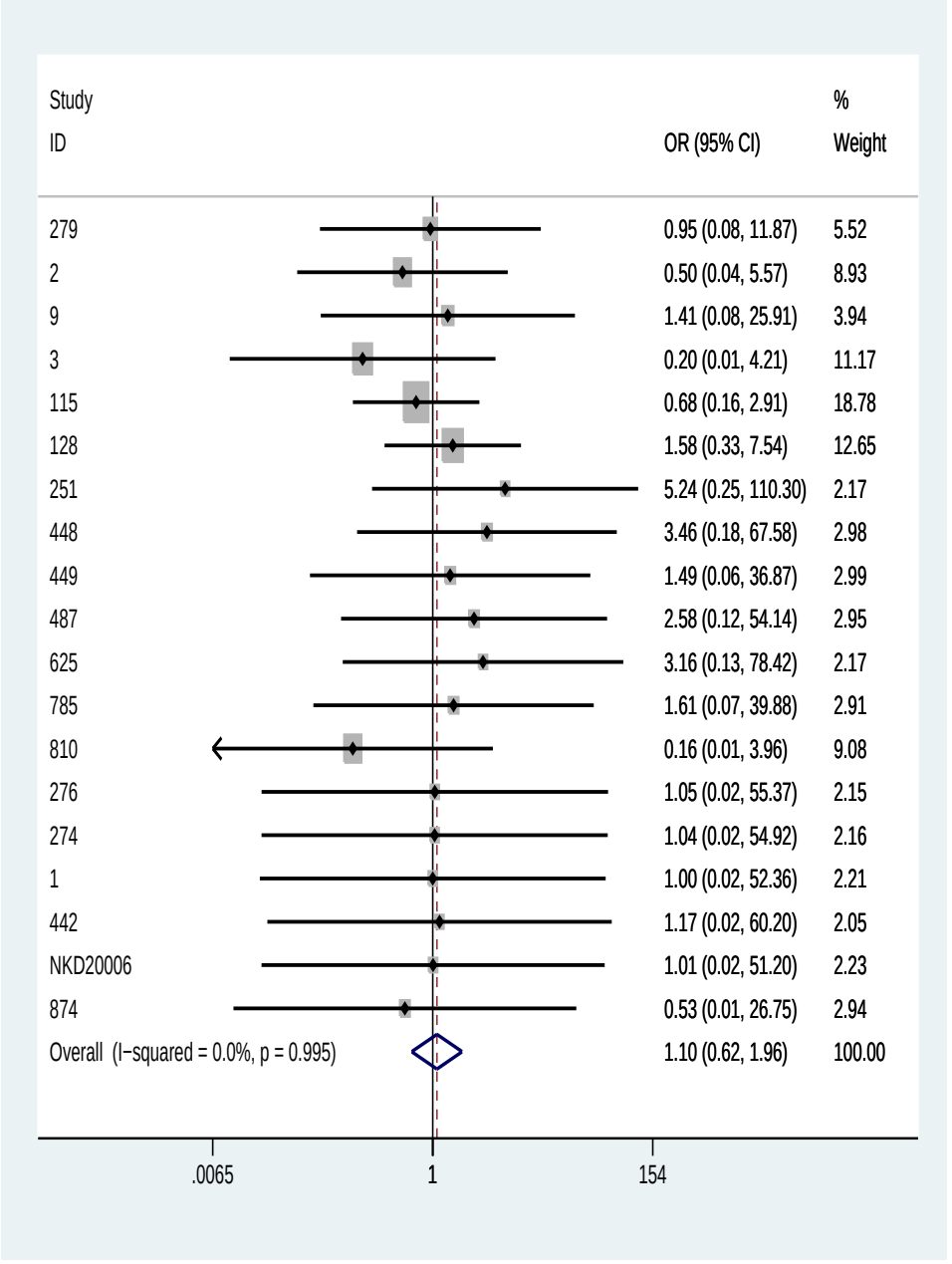


Figure 7.9: Analysis X, Mantel–Haenszel model with odds ratio, continuity correction 0.05 applied to all studies with at least one arm with zero events.

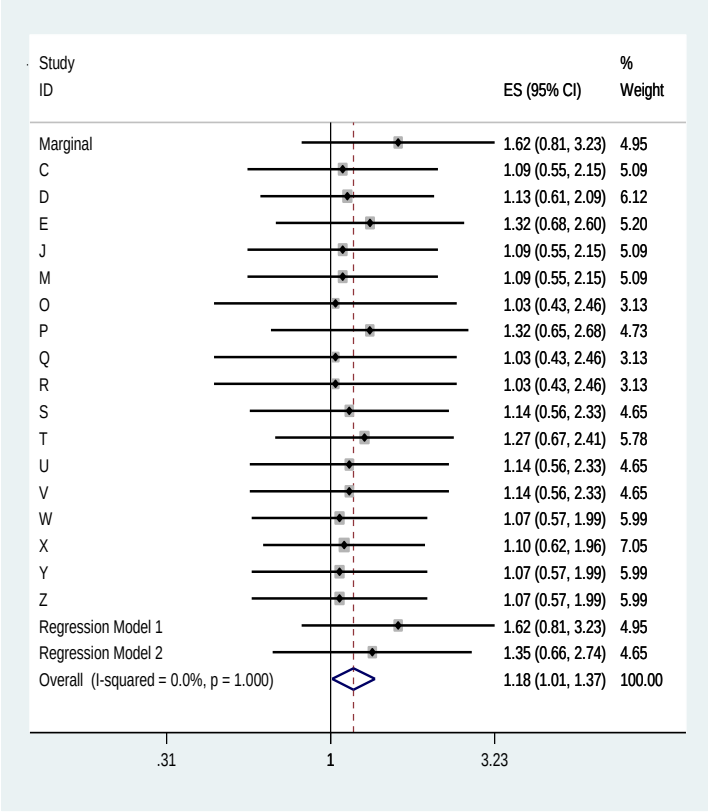


Figure 7.10: Forest plot of odds ratio values for pooled analyses. Study ID refers to marginal analysis, Analysis ID from Table 7.4 (Study ID C–Z) and Regression Models 1 and 2 from Table 7.5.

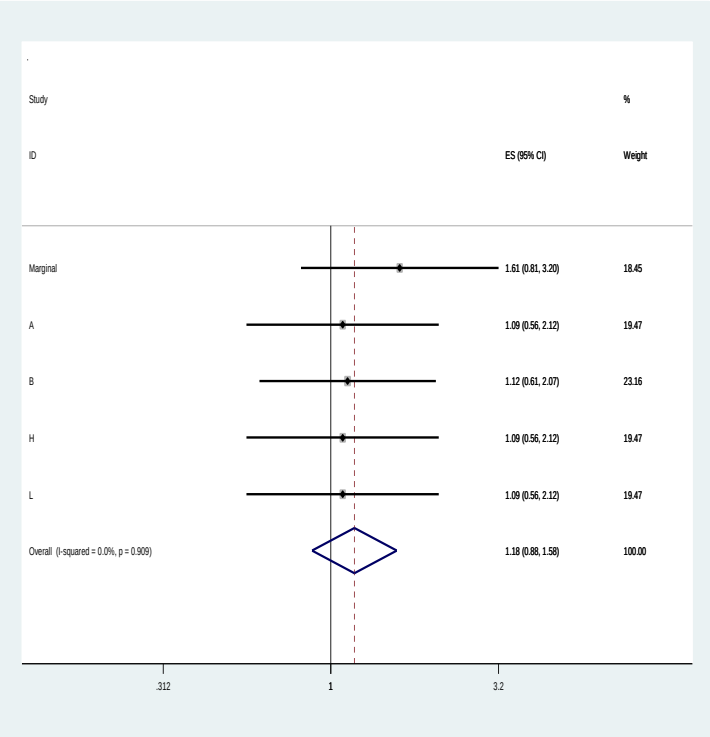


Figure 7.11: Forest plot of relative risk values for pooled analyses. Study ID refers to marginal analysis and Analysis ID from Table 7.4 (Study ID A–L).

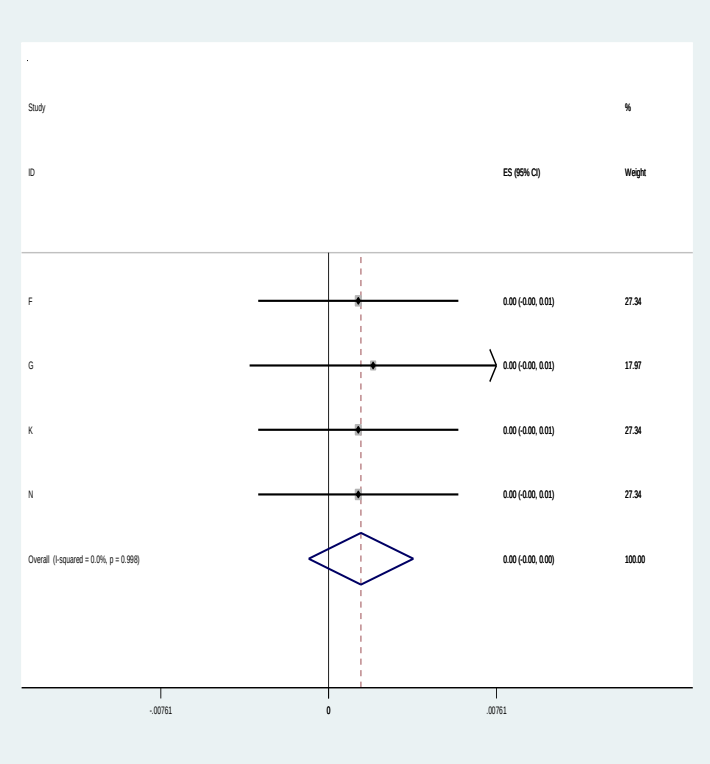


Figure 7.12: Forest plot of risk difference values for pooled analyses. Study ID refers to Analysis ID from Table 7.4 (Study ID F–N).

7.7.3 Densities for selected posterior distributions in Bayesian models

Figures 7.13 and 7.14 show posterior densities for between-studies standard deviation for Bayesian analyses B.1–B.15 as set out in Table 7.6; Figure 7.15 shows posterior densities for μ, the underlying mean log OR for suicidality in paroxetine users compared to controls, in selected Bayesian analyses (B.11 and B.13–B.15).

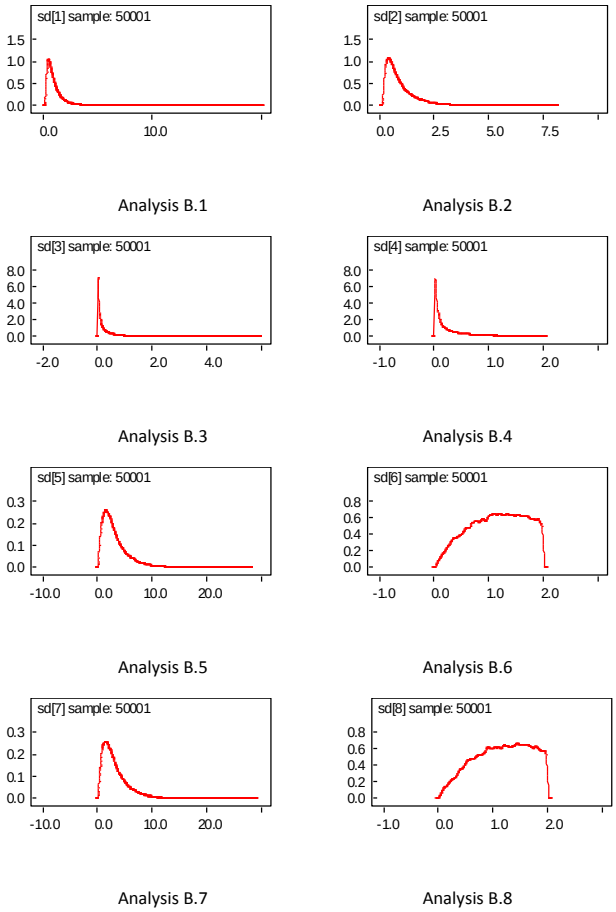


Figure 7.13: Posterior densities on standard deviations for Bayesian analyses B.1–B.8.

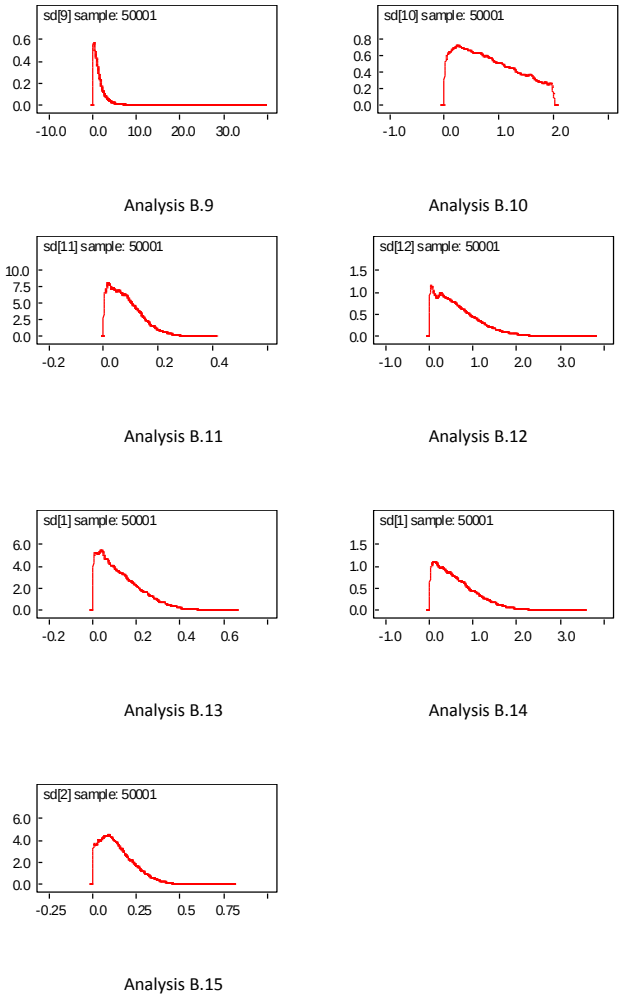


Figure 7.14: Posterior densities on standard deviations for Bayesian analyses B.9–B.15.

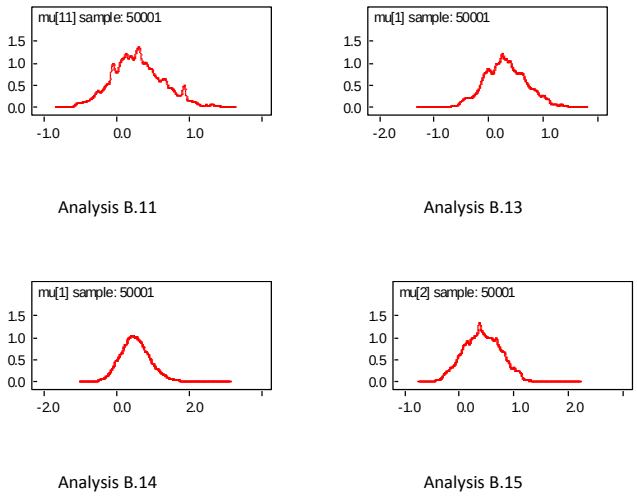


Figure 7.15: Posterior densities on the mean underlying log OR μ for selected Bayesian analyses B.11; B.13–B.15.

7.8 Summary

Multiple analyses were performed on the same dataset, which included data from 19 trials comparing the anti-depressant paroxetine against a placebo. The outcome of these trials was the occurrence of suicidal behaviour and ideation. Frequentist analyses compared both relative and difference scale outcome metrics, and the use of different continuity corrections where required. Also, the inclusion of studies with zero events was considered. Logistic regression methods were also evaluated.

Bayesian analyses focused on the use of different priors, as the inclusion of studies with zero events in one or more studies was not at issue. A range of priors intended to be non-informative was compared, and the derivation of priors using data from studies related but not sufficiently similar to be included in the current analysis was also investigated.

Whilst these analyses did not produce any firm conclusions regarding association of suicidality with paroxetine, the signal from the data, across all analyses, was strong enough to merit concern. The variation of outcomes according to the analysis used highlighted the benefits of multiple analyses for comparison purposes.

8

An individual patient data meta-analysis of randomised controlled trials

8.1 Introduction

The work presented in this chapter was conducted in collaboration with colleagues based at Wyeth Research, USA, and the Mayo Clinic College of Medicine, USA; these colleagues provided the dataset based on original clinical trials involving etanercept. The work was based on a prior protocol, and additional information regarding the background and methodology of the project, as well as the pertinent results, has been published elsewhere (Bongartz *et al*. 2009). The full publication is set out in Appendix D. The contribution to the publication from this thesis was to scrutinise the initial dataset for anomalies in data entry, and when the dataset had been finalised, to perform all statistical analyses and create a technical report providing details of the methodology and results.

Rheumatoid arthritis (RA) is a chronic and disabling condition that has been treated with several different types of drugs, with different mechanisms of action. Some drugs are aimed at simply controlling the symptoms of RA, whilst others aim to influence the progression of the disease itself. Such drugs are known as disease-modifying anti-rheumatic drugs (DMARDS). In this category, there are drugs that interfere with the fundamental inflammatory process that is mediated

by tumour necrosis factor (TNF). These drugs are known as anti-TNF drugs, which include infliximab, adalimumab and etanercept.

Although the positive benefits of anti-TNF drugs have been demonstrated, there are reasons to be concerned regarding possible adverse effects of anti-TNF drugs. There is evidence to indicate that TNF combats infection and has a role in promoting the destruction of tumour cells. TNF may also be involved in tumour promotion, so the issue of whether TNF blockers would also have a beneficial or damaging effect on tumour development is unclear at present.

As well as the concerns from a mechanistic viewpoint, several studies have demonstrated an increased incidence of lymphoma among people with RA (Wolfe & Michaud 2004, citing earlier work). A prospective cohort study of 18 572 participants (Wolfe & Michaud 2004) also found an increased risk of lymphoma among RA patients. When comparing standardized incidence ratios (SIRs) between different treatment groups, the patients using anti-TNF drugs appeared to have a higher incidence of cancer than those patients not using anti-TNF drugs.

Within an RCT, there will be only a few incidences of newly-diagnosed cancers due to the low overall incidence of cancer. Hence, it is very difficult to glean any meaningful information regarding the influence of anti-TNF drugs on cancer promotion from a single trial. A means of addressing this issue is to combine the results of several trials using meta-analysis. A meta-analysis of summary data from trials using infliximab and adalimumab has been conducted by Bongartz *et al*. (2006), investigating the influence of these drugs on infection and malignancy. Both of these outcomes appeared to have a significantly increased risk with use of anti-TNF drugs compared to placebo. For malignancies, the odds ratio (OR) for malignancy was 3.3, with a 95% confidence interval (CI) 1.2; 9.1; for serious infection the OR was 2.0 (95% CI 1.3; 3.1).

Etanercept was deliberately excluded from this earlier meta-analysis due to differences in its molecular structure and mechanism of action within the anti-TNF class. Therefore, a meta-analysis of etanercept trials would be highly desirable to elicit any information regarding its potential influence on malignancy.

The aim of this research was to analyse data from randomised controlled trials (RCTs) that have used etanercept, either alone with another anti-RA drug such as methotrexate, with regard to incidence of cancer.

Data were available from nine separate trials, four sponsored by Wyeth and five from Amgen. For all trials, individual participant (or patient) data (IPD) were provided, allowing for IPD meta-analysis. Data are presented in survival format (time-to-event), allowing for both an IPD survival analysis and traditional meta-analysis of summary results to be performed.

8.2 Methods

8.2.1 Study protocols

Details of the search strategies, trial selection, study quality assessment and data extraction are described in Bongartz *et al.* (2009). This meta-analysis has followed as far as feasible the study protocols as developed by the project collaborators.

8.2.2 Statistical methods

The analysis presented is not a formal intention to treat (ITT) analysis. For an efficacy study, an ITT analysis is regarded as the 'gold standard' procedure. For a study focusing on adverse events however, there is an argument that only those patients who received at least one dose of the drug should be included, to avoid the inclusion of events in patients who may have been randomised to a particular treatment but have not actually been treated due to early withdrawal. As discussed by Higgins *et al.* eds. (2008), some people argue that it would be wrong to attribute an adverse event to a treatment that was not received by the patient.

The risk window has been defined as the date of first dose to the date of final follow-up (as opposed to the date of last dose) or date of first incident cancer, whichever occurs first. For many participants the date of last dose is the same as the date of last follow-up. Rollover into an open-label study has not been considered at this point.

IPD meta-analyses of the survival data using both a fixed effect (Cox's Proportional Hazards model) and random effects, based on a Poisson generalized linear model (GLM) were performed. Such a model is the equivalent of a piecewise

exponential model (Friedman 1982). The data were stratified by trial in order to maintain randomisation and explore any potential heterogeneity in the trial populations.

The fixed effect model allows for an unconstrained baseline (different baseline cancer survival rates between different trials) whilst assuming the same treatment effect among all trials (Tudur Smith 2005a). Whilst this assumption is questionable on philosophical grounds, due to variations in demographic characteristics between participants in different trials, in this case there are also differences between the trials in the treatments given.

Some trials include etanercept alone whilst others include etanercept with another RA drug. There may be some association between cancer and the other drugs in the trial that are included with etanercept. Hence, a random effects model (which allows for the possibility of different treatment effects between the trials) has also been used in the analysis; this approach will allow the detection (by comparison with the fixed effect model) and incorporation of any statistical heterogeneity between the trials (Tudur Smith 2005a).

All analyses compare participants receiving etanercept (alone or in combination) against a comparator, which may be a placebo or a combination of one or more drugs not including etanercept (with the exception of a sensitivity analysis as described in Section 8.6.2).

All variations on the basic model (etanercept against no etanercept) have been performed using the Cox's Proportional Hazards model. Exposure duration has been included in the model by adding a time-dependent covariate for treatment into the model.

The random effects model was conducted using a Poisson GLM, which involved dividing the risk window into time sections to allow for Poisson modelling of the number of events in each time period as a Poisson process (Whitehead 1980; Lindsey 1995; Ma 2003). Ideally, the dataset would be divided into as many time sections as there are events, in order to maximise the accuracy of the Poisson regression model and approximate a Cox model. However, due to computational limitations, this approach was not feasible, and it was decided to divide the risk window into six periods, ensuring that at least one individual in each of the nine trials contributed some person-days to the dataset for each period. This approach enabled the random effects model to be performed.

The basic model was extended in the analyses by including age and gender into the model, separately, together, and with an interaction term. Other factors such as use of concomitant DMARDs, duration of RA and dose of etanercept have not yet been considered, but may be investigated at a later date.

The major outcome is all-cause malignancy. Although it was intended in the original protocol to consider lymphoproliferative disorders separately, the sparsity of incident cancers in this category prevented this (only one case of Hodgkin's lymphoma, and one case of large granular lymphocytic (LGL) leukemia were found). However, a sensitivity analysis excluding basal cell carcinomas (any site) was performed. Additional sensitivity analyses included exclusion of cancers diagnosed within 6 weeks (42 days) of receiving the first dose, and cutting off the trial follow-up at specific time points including 6 months, 1 year, and 2 years.

In addition to the survival analyses, meta-analyses of summary data were performed using the number of events in each trial, regardless of time of occurrence. This approach necessarily loses the time-to-event element of the survival analysis, and therefore cannot reflect issues such as the difference in trial duration, or differences in time to onset of cancer that may be of relevance in addition to overall event frequency between different treatment arms.

Meta-analyses of the odds ratios (ORs) between treatment arms for the different trials were pooled using both fixed and random effect(s) models. The Mantel-Haenszel method was used for the fixed effect model, and the DerSimonian & Laird method for the random effects model (discussed in Sections 3.3.3 and 3.5.1). Heterogeneity is investigated by use of the chi-squared test based on the Q statistic of the appropriate model, as well as estimates using the tau-squared and I-squared statistics (Sections 3.5.1 and 3.9).

With sparsity of data, there are difficulties when using both of these methods as they are unable to compute pooled ORs when there is a zero value in one or both treatment arms. Hence, for studies with one zero value for a treatment arm, a continuity correction has been applied to allow these studies to be included. The traditional continuity correction of 0.5 to all cells of the relevant study has been used, but due to concerns in a sparse dataset of effectively adding an extra case for each study requiring a continuity correction, a smaller continuity correction of 0.05 has also been used. The continuity correction for sparse data devised by Sweeting *et al.* (2004) has also been used (Section 5.2.3). This is based

on dividing the overall continuity correction (for example, 1, in the case where both arms receive 0.5) in proportion to the number of participants in each arm. The studies that did not yield any events are excluded from the meta-analysis.

As well as the OR, the hazard ratio (HR) has been used as a summary statistic for meta-analysis, again using a continuity correction of 0.5 to allow for the inclusion of studies with zero events in one arm and excluding trials with no events.

All analyses were performed using Stata® version 9.2, with the exception of the random effects survival model, which was performed using R version 2.5.0.

8.2.3 Fixed effect and random effects models for hazard ratios

Model 1: Fixed treatment effect with no stratification by trial

$$\lambda(t) = \lambda_0(t)e^{\beta X_i} \tag{8.1}$$

where λ is the hazard function, λ_0 is the baseline hazard function across all individuals, t is time and β is the log hazard ratio for the level of covariate X (in this example treatment) for the ith individual.

Model 2: Fixed treatment effect with stratification by trial (i.e. each trial has a different baseline hazard function derived from a common distribution)

$$\lambda(t) = \lambda_{0j}(t)e^{\beta X_i} \tag{8.2}$$

where λ is the hazard function, λ_{0j} is the baseline hazard function for the jth trial, t is time and β is the log hazard ratio for the level of covariate X for the ith individual.

Model 3: Random treatment effect with stratification by trial (i.e. the treatment effect varies between trials, derived from a common distribution, in addition to the variation of baseline hazard function between trials)

$$\lambda(t) = \lambda_{0j}(t)e^{\beta_j X_i} \tag{8.3}$$

where λ is the hazard function λ_{0j} is the baseline hazard function for the jth trial, t is time, β_j is the log hazard ratio for the level of covariate X for the ith individual in the jth trial and $\beta_j \sim N(\beta, \tau^2)$, where τ^2 is the between-study heterogeneity for the treatment effect.

8.3 Initial data exploration

An 'event' in this analysis is defined as a first cancer that is incident during the study; a recurrence of a previously diagnosed cancer is not included as an event. Time-to-event is therefore time to first incident cancer; any second incident cancers are disregarded as it is only time to first incident cancer that is being considered.

The initial dataset included data from nine trials, with a combined total of 3318 individuals. However, two participants were immediately excluded, one of whom had never received the allocated treatment and one who was not followed up after the first dose. Hence, there were 3316 participants contributing person-days to the dataset. Also, one participant was incorrectly recorded as having cancer (as a result of having abnormal cervical cytology); this reduced the number of recorded cancers to 33 from 34.

There were some other anomalies noted in the dataset that did not impact on the analysis. For example, there were some discrepancies in the Trial Nos. and Universal Patient ID Nos. that indicated that some patients had been involved in more than one trial. Two patients receiving placebo were identified in one trial who had also been allocated to placebo in a previous study also included in the current meta-analysis. There were also six patients who had been allocated to placebo in the earlier trial and then went on to be allocated to etanercept in the subsequent trial. None of these eight patients had an event. Hence, all

of these patients were included as members of both of these trials. Also, there were six participants who had been involved in an earlier trial, not included in the current dataset, four of whom received etanercept in both trials, and two who received etanercept in the first trial and placebo in the second. Again, none of these patients had an incident cancer [whilst in the second trial that is included in the current dataset].

This knowledge of trial transfers provides reassurance that there is no case of any cancer being counted twice within the dataset, or having been caused by a treatment given within an earlier trial. However, this duplication of participants between trials does call into question independence between trials and means that the 3316 observations are not generated by 3316 separate individuals. In actuality, the total dataset comprised 3308 individuals, eight of whom had records from two trials. Due to the small number of transfers between trials however, all participant records have been included in the analysis as if they were from different individuals. Nor has etanercept treatment in a study not being analysed within the current dataset been taken into account.

Table 8.1 shows the distribution of participants across the nine trials, with a breakdown of treatment arms and number of events in each arm. Note that for the analyses, the arms were amalgamated into etanercept (alone or in combination with other drugs) as the treatment group and non-etanercept as the comparator group (this distribution of participants and events is shown in Table 8.2). Of the 33 events, only seven (21.2%) occurred in the non-etanercept comparator groups, while 26 (78.8%) were found in the participants who had received etanercept. The gender breakdown was 2555 (77.1%) females and 761 (22.9%) males. Mean age across the full dataset was 53 years.

When investigating a survival dataset with so few events, graphical methods of displaying the dataset (such as Kaplan-Meier plots, or log-log plots) add little to understanding (so not included).

Table 8.1: Breakdown of trials by treatment arm and numbers of cancer and non-cancer events (% are of total cancers).

	Placebo		**Etanercept**		**Etanercept+MTX**		**Etanercept+SSZ**		**MTX**		**Placebo+MTX**		**SSZ**	
Cancer status	Non-cancer	Cancer	Non-cancer	Cancer	Non-cancer	Cancer	Non-cancer	Cancer	Non-cancer	Cancer	Non-cancer	Cancer	Non-cancer	Cancer
Trial No.														
TNR-00102	50	0(0%)	103	0(0%)										
0881300	105	0 (0%)	451	2 (6.1%)										
0881308			218	5 (15.2%)	226	5 (15.2%)			227	1 (3.0%)				
0881309			102	1 (3.0%)			101	0(0%)					50	0(0%)
160004	44	0(0%)	136	0(0%)										
160009	80	0(0%)	153	1(3.0%)										
160012			405	10 (30.3%)					213	4 (12.1%)				
160014					59	0 (0%)					30	0 (0%)		
160029	266	2(6.1%)	264	2(6.1%)										

MTX: methotrexate; SSZ: sulfalsalazine.

Table 8.2: Breakdown of trials by etanercept and comparator (non-etanercept) arms (% are of total cancers).

	Comparator (non-etanercept)		**Etanercept**	
Cancer Status	Non-cancer	Cancer	Non-Cancer	Cancer
Trial No.				
TNR-00102	50	0(0%)	103	0(0%)
0881300	105	0 (0%)	451	2 (6.1%)
0881308	227	1 (3.0%)	444	10(30.3%)
0881309	50	0(0%)	203	1 (3.0%)
160004	44	0(0%)	136	0(0%)
160009	80	0(0%)	153	1(3.0%)
160012	213	4 (12.1%)	405	10 (30.3%)
160014	30	0 (0%)	59	0 (0%)
160029	266	2(6.1%)	264	2(6.1%)

8.4 Primary meta-analyses

8.4.1 Survival models using individual patient data

Table 8.3 shows the results of the survival models applied to this dataset. All analyses are using a fixed effect model unless stated otherwise. The three primary models of this analysis are set out in Section 8.2.3.

In an initial model (Model 1, Section 8.2.3), ignoring the effect of the individual trial (not an appropriate method for a multi-trial analysis but undertaken for comparison), the pooled HR was 1.56 (95% CI 0.68; 3.59). When including the effect of the different trials within the model (Model 2, Section 8.2.3), the pooled HR increased slightly to 1.84 (95% CI 0.79; 4.28). Although the HR is raised in the etanercept recipients compared to the non-etanercept groups, the wide CI suggests that this result may be simply due to chance.

Using the random effects model as described above (also set out as Model 3, Section 8.2.3), the results were very similar, with an HR of 1.81 (95% CI 0.78; 4.22). Using this model, the estimate of between-studies heterogeneity was 1.20. Hence, due to the similarity of results between the fixed and random effect(s) models, as a function of lack of heterogeneity, no additional random effects models, (or models including treatment by covariate interactions attempting to explain heterogeneity in treatment effects) were considered.

Table 8.3: Results of individual patient data survival meta-analyses, stratified by trial except where stated otherwise.

Dataset	Model	Covariate	HR	LCI	UCI	p-value
Full	Treatment (Model 1)*†	Treatment	1.56	0.68	3.59	0.30
Full	Treatment (Model 2)†	Treatment	1.84	0.79	4.28	0.16
Full	Treatment (RE) (Model 3)†	Treatment	1.82	0.78	4.22	0.17
Females	Treatment	Treatment	1.67	0.60	4.64	0.32
Males	Treatment	Treatment	2.12	0.46	9.83	0.34
BCC Excluded	Treatment	Treatment	1.37	0.54	3.51	0.51
Excluding cancers diagnosed within 42 days	Treatment	Treatment	1.87	0.75	4.62	0.18
Full (6 month cut-off)	Treatment	Treatment	1.50	0.38	5.90	0.56
Full (1 year cut-off)	Treatment	Treatment	2.59	0.87	7.70	0.086
Full (2 year cut-off)	Treatment	Treatment	1.78	0.76	4.16	0.18
Full	Treatment, treatment varying with ln(time)	Treatment	2.17	0.05	102.12	0.69
Full	Treatment, treatment varying with ln(time)	Treatment varying with ln(time)	0.97	0.47	2.01	0.93
Etanercept only v. placebo	Treatment	Treatment	1.45	0.26	8.17	0.67

*Unstratified; † see Section 8.2.3 for models; HR: hazard ratio; LCI: lower confidence interval bound; UCI: upper confidence interval bound.

In order to determine if etanercept treatment increased risk of cancer as time progressed, treatment was added into the model as a time-varying covariate, varying by log(time). The treatment effect did not appear to vary with log(time), with an HR of 0.97 (95% CI 0.47; 2.01).

8.5 Meta-analyses of summary data

Meta-analyses of summary data cannot encompass the 'time-to-event' element of the IPD analyses, but do provide a useful alternative approach against which to compare the survival analyses. If the issue of importance is simply the occurrence or non-occurrence of an event, and the timing of the event is not considered of importance, then use of summary data is appropriate.

Using summary data in the form of ORs between the etanercept and comparator groups, meta-analyses were performed, allowing for the inclusion of single-zero trials by means of continuity corrections. This approach immediately excluded the three double-zero studies (studies with no total events). (It has been demonstrated by Sweeting *et al*. (2004) that such studies do not add to the overall analysis.) These were studies TNR-00102, 160004 and 160014. The continuity corrections were applied to studies where there was only one arm with zero events: studies 0881300, 0881309 and 160009.

Using a continuity correction of 0.5, the OR using a Mantel-Haenszel fixed effect model was 1.68 (95% CI 0.77; 3.69). When using a smaller magnitude continuity correction (0.005), the OR increased to 2.04 (95% CI 0.87; 4.83). This second result may be arguably the more valid as fewer 'false' participants are added with the smaller continuity correction (0.02 per study rather than one per study). Interestingly, the continuity correction of Sweeting *et al*. (2004) produced a greater OR than the 0.5 continuity correction (but less than that of the 0.005 continuity correction), despite adding the same number of 'false' participants to the dataset, due to the overall continuity correction per study summing to 2 (OR 1.93, 95% CI 0.85; 4.38). The full results of the OR meta-analyses are set out in Table 8.4. Figure 8.1 shows the forest plot for the Mantel-Haenszel meta-analysis with a continuity correction of 0.5.

Table 8.4: Results of meta-analyses using odds ratios as summary measure.

Meta-analysis method	Exclusions	Continuity correction	Continuity correction applied to	OR	LCI	UCI	I-squared %
M–H	Double-zero studies	0.5	Single-zero studies	1.68	0.77	3.69	0.0
M–H	Double-zero studies	0.005	Single-zero studies	2.04	0.87	4.83	0.0
D&L	Double-zero studies	0.5	Single-zero studies	1.50	0.67	3.37	0.0
D&L	Double-zero studies	0.005	Single-zero studies	1.63	0.66	4.03	0.0
M–H	Double-zero studies	SS sum to 1	Single-zero studies	1.93	0.85	4.38	0.0
D&L	Double-zero studies	SS sum to 1	Single-zero studies	1.71	0.73	4.01	0.0

D&L: DerSimonian & Laird; LCI: lower confidence interval bound; M–H: Mantel–Haenszel; OR: odds ratio; SS: Sweeting & Sutton continuity correction; UCI: upper confidence interval bound.

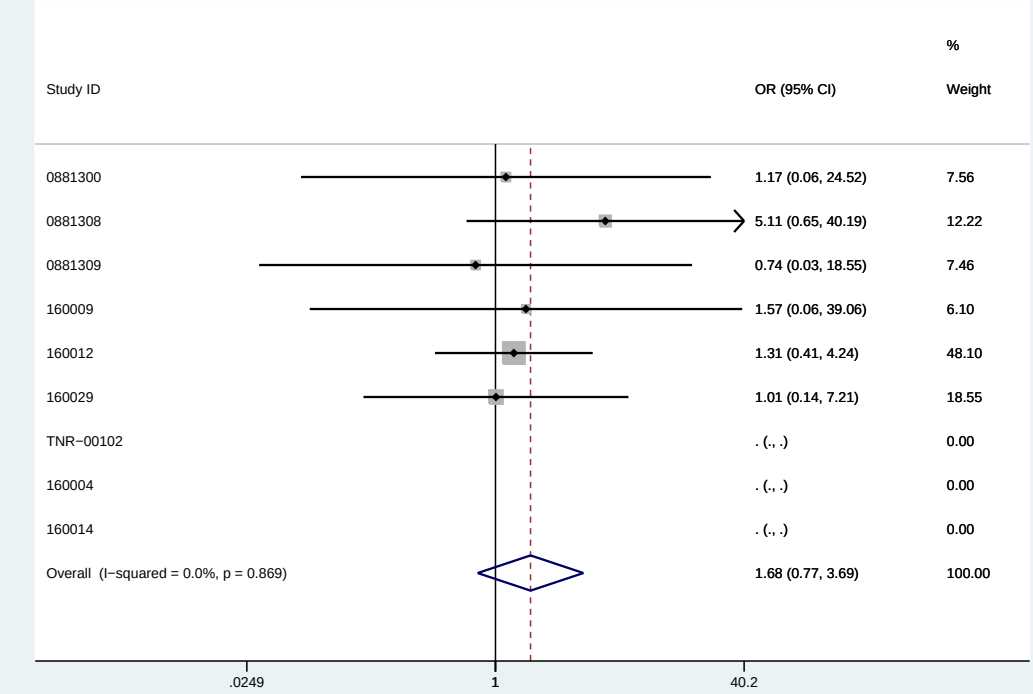


Figure 8.1: Forest plot for Mantel–Haenszel meta-analysis of summary odds ratios using a continuity correction of 0.5 where appropriate.

The results using a random effects DerSimonian & Laird model were very similar to those of the Mantel–Haenszel model, reflecting the fact that estimates of heterogeneity using I-squared were all very small (0.0% in all cases). Furthermore, the *p*-values for the chi-squared tests of heterogeneity were all non-significant (taking 0.1 as the significance level). When using the random effects model, the tau-squared estimates were also 0.0, indicating no heterogeneity between studies. However, a different pattern was seen regarding the use of the continuity corrections with the 0.5 continuity correction yielding the lowest OR and that of Sweeting *et al*. (2004) the highest OR. Regardless of the continuity correction applied, heterogeneity was still not an issue.

A second method for a two-stage IPD analysis was to use the summary HRs from each study. As it is impossible to calculate a meaningful HR if there is an arm with zero events, a continuity correction of 0.5 was again applied to studies with no events in one arm only, whilst studies with no events overall

were excluded. This approach resulted in an HR of 1.34 (95% CI 0.60; 2.99). See Figure 8.2 for the associated forest plot.

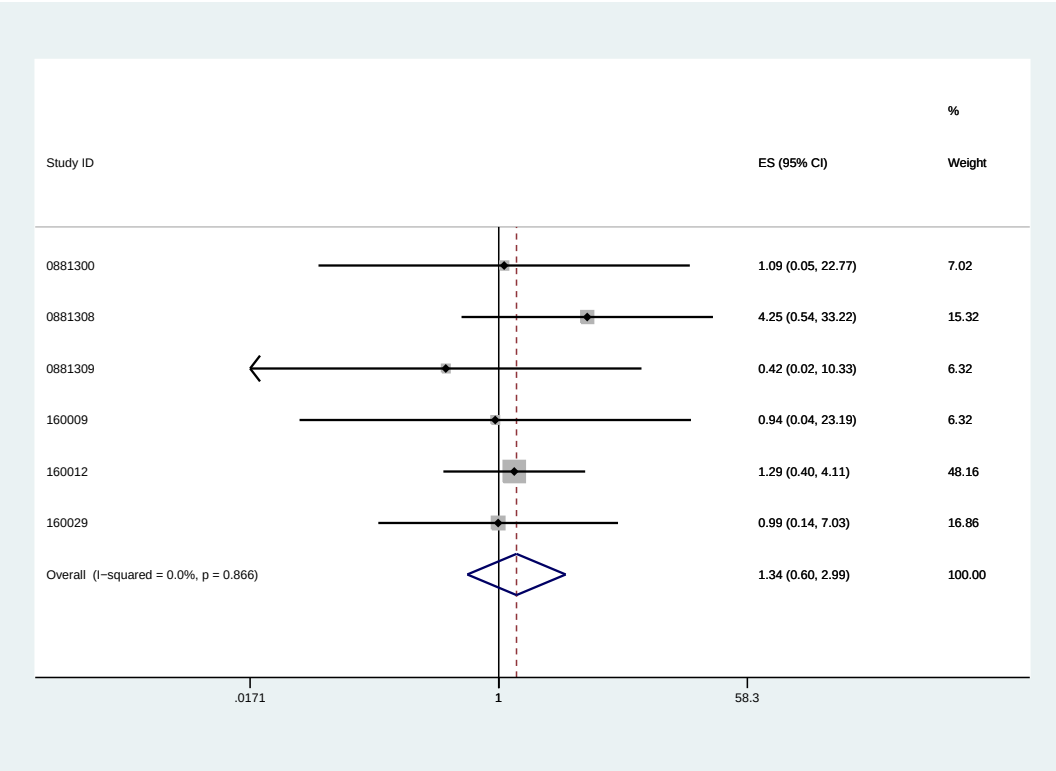


Figure 8.2: Forest plot for meta-analysis of summary hazard ratios using a continuity correction of 0.5 where appropriate.

8.6 Additional analyses

8.6.1 Subgroups

When analysing males and females separately in the dataset, the HR for males was 2.12 (95% CI 0.46; 9.83), while for females the HR was 1.67 (95% CI 0.60; 4.64).

8.6.2 Sensitivity analyses

Exclusion of basal cell carcinomas

It was decided to exclude basal cell carcinomas (BCCs) of all sites as part of a sensitivity analysis. There were nine BCCs in total, seven skin, one lip and one nose. Of these, three had a second cancer, two of which were also BCCs, with one skin cancer of unknown cell type. These second cancers were not included as events, therefore survival time for these individuals was re-calculated as time to last follow-up. With this revised dataset with only 24 events, the HR in a model with treatment as the only covariate was now 1.37 (95% CI 0.54; 3.51). Again, this does not indicate any statistically significant increased risk of cancer with etanercept treatment.

Exclusion of cancers diagnosed within 6 weeks of first dose

For this analysis, all participants with a first incident cancer diagnosed at less than 42 days from the first treatment dose were excluded from the dataset, reducing the number of individual records to 3312. There were only four such events in total, thus leaving 29 events in the dataset. It was decided to exclude these participants because it was almost certain that their malignancy was already in development when commencing the trial, also their follow-up after diagnosis may be curtailed due to their illness; hence, they may be fundamentally different from the other participants. With these exclusions, the HR for cancer incidence in the etanercept group compared to the non-etanercept group was 1.87 (95% CI 0.75; 4.62).

Censoring follow-up at specific times

In order to avoid the lengthy follow-up times for certain individuals adding undue weight to the dataset and to investigate whether there are any particular time periods where etanercept treatment is associated with increased incidence in cancer, the dataset was censored at three different time points: 6 months, 1 year and 2 years. With a follow-up period of only 6 months, there were only 11 events. The HR for the etanercept group was 1.50 (95% CI 0.38; 5.90). When censoring the dataset at 1 year, there were 24 events in the dataset. At this cut-off point the HR for the etanercept group was 2.59 (95% CI 0.87; 7.70), with a p-value of 0.086. Only one incident cancer occurred after 2 years, hence cutting off the follow-up at 2 years added little to the analysis.

Restricting the analysis to etanercept alone compared to placebo

An analysis was performed with the exclusion of participants who had received etanercept in combination with other drugs, and comparator group participants who had received drugs other than etanercept (which may also have the potential to influence cancer incidence). This analysis included data from only five trials, all of which had two arms, one of etanercept alone and the other with placebo. There were only seven incident cancers, hence analysis is limited. The HR of etanercept only compared to placebo was 1.45 (95% CI 0.26; 8.17).

8.7 Discussion

In view of the sparse number of events in this dataset, the analyses have limited power and any results should be regarded with caution. The main feature of interest in this dataset is that the risk of cancer is always higher for the treatment group than the comparator group, regardless of the method of analysis. Whilst the 95% CIs always include 1, this outcome is sufficient to cause concern regarding the use of etanercept. Only one result came close to being statistically significant at the 5% level; this was the HR for the IPD meta-analysis of cancers censored at 1 year, which was 2.59 (95% CI 0.87; 7.70). There are, however, many caveats against taking this result at face value. It must always be remembered that the participants in this study (in both etanercept and control arms) are at a higher risk of developing malignancy than the general population, and that some of the participants in the etanercept arms were receiving other drugs, while those in the comparator group may have been receiving other drugs or a placebo. Such discrepancies in treatment may have influenced the results. However, the comparison of etanercept alone compared to placebo alone was also non-significant.

There are also many other factors in the dataset that may have confounded the results. No account was taken of duration of RA, or of concomitant (or even previous) DMARD therapy. As many cancers have a very long latency period of development prior to detection through symptoms or screening, it is possible that many of the cancers in this dataset had been initiated well before trial commencement. Due to lack of heterogeneity within the dataset, also the overall sparsity of events, investigation of demographic characteristics such as

age, ethnic group and smoking status was not considered worthwhile. From a methodological perspective, one feature of interest was the different pattern in the ORs when the different continuity corrections were applied to both fixed effect and random effects model. With the fixed effect model the 0.005 continuity correction produced the highest HR, whilst in the random effects model the continuity correction of Sweeting *et al*. (2004) resulted in the highest HR. This indicates that when using a random effects model the Sweeting *et al*. (2004) continuity correction may yield the highest HR. When looking for adverse events, in the interests of caution, an HR which may be spuriously inflated is preferable to one that is artificially low. Hence, the continuity correction of Sweeting *et al*. (2004) may be the most appropriate.

8.8 Conclusions and potential for further analysis

Whilst no results of statistical significance have been found, there are signals from the dataset that leave cause for concern regarding the association between etanercept and cancer incidence. As an outcome, cancer incidence has a long lead-in time (it can be in existence for a long period of time prior to detection), is often difficult to detect and diagnose, and has many different factors including genetic and environmental that influence its occurrence. When using RCTs that are relatively short in duration as the data source, it is very difficult to generate adequate data of high quality to investigate this question.

There are many ways in which the investigation of this issue could be extended, for example:

1. inclusion of data from trials looking at other anti-TNF drugs;
2. inclusion of data from observational studies of anti-TNF drugs and cancer incidence, if available;
3. analysis of the current dataset using Bayesian methods to incorporate external data in the form of a prior distribution; or by modelling it explicitly within an hierarchical model, possibly allowing for potential biases (Spiegelhalter & Best 2003);
4. incorporation of data regarding anti-TNFs used for other indications;

5. inclusion of other aspects of the dataset as covariates, such as concomitant DMARD therapy or duration of RA;
6. analysis of different drug combinations within the treatment and comparator groups; and
7. comparison of results with those of summary data meta-analyses using data extracted from published literature.

Data on this subject are difficult to obtain, therefore it is very important to use data from all available sources so that any signal from the data can be magnified. When data are sparse, it is difficult to reach statistical significance, but this does not indicate that there is nothing of clinical concern in the dataset. Additional data and analysis could go some way to strengthening evidence, either for or against an association between cancer incidence and etanercept (or anti-TNF drugs as a class), which could then impact on clinical decision-making.

The association between anti-TNFs as a therapy for RA and the risk of malignancy is discussed further in Chapters 9 and 10, using the summary data as published in Bongartz *et al*. (2009), and additional summary data from trials of adalimumab and infliximab. Given the multiple issues of different individual anti-TNFs within the class of drug, dose effects, and use of additional anti-rheumatic drugs, a mixed treatment comparison (MTC) approach was used to investigate these elements of therapy within the framework of RA and anti-TNF treatment. This allows more detailed investigation of the clinical aspects of the problem, as well as extending the statistical methodology within the broader context of adverse events.

8.9 Summary

Meta-analysis methods have been applied to a dataset of 3316 participant records in nine RCTs of etanercept in people with rheumatoid arthritis. The outcome of interest was a first incident cancer during the follow-up period of the trial following the first treatment dose. The main comparison was between participants who received etanercept, alone or in combination with any other RA drugs, against those who did not receive etanercept. This comparator group may have received a placebo or one or more RA drugs not including etanercept (or any other anti-TNF drug).

A variety of meta-analysis methods have been applied to this dataset. IPD were available for all trials, presented as survival (time-to-event) data. Therefore, it was possible to analyse the data using survival methods whilst taking into account the fact that the data were from different trials (stratifying the data by trial). A fixed effect model using Cox's Proportional Hazards modelling was used, and additionally a random effects model based on a Poisson generalized linear model.

Separate analyses for male and female participants were performed, as were sensitivity analyses excluding basal cell carcinomas, excluding participants with cancer diagnosed at less than 6 weeks following first treatment, and by censoring follow-up at specific times.

Also, in addition to this IPD meta-analysis, for comparison purposes, 2-stage IPD meta-analyses have been performed, using both the odds ratio (OR) and hazard ratio (HR) as summary measures for each trial. A disadvantage of such measures is that studies with no events are excluded. A range of continuity corrections was applied to the OR summary data to enable inclusion of studies where there were no events in one of the treatment arms, and to assess robustness of results to the inclusion of such factors.

9

Use of mixed treatment comparisons in the context of adverse events data

9.1 Introduction

Traditional meta-analysis methods (as described in Chapter 3) are useful in scenarios where there were only two treatments that required comparison. When the situation arises where multiple treatments for a specific condition are being tested, against inactive or standard treatments, such that a network of comparisons is created, traditional meta-analysis methods are not adequate for this situation. Systematic pairwise comparisons for each pair of treatments lack validity, as the fact that different treatments are compared against each other in different trials is not accounted for in pairwise meta-analysis. Furthermore, only direct comparisons are possible using pairwise meta-analysis, with the result that if two treatments are not used in the same trial their relative treatment effects cannot be directly compared.

The concept of mixed treatment comparisons (MTCs) was developed to address such situations. Such MTC methods have been used effectively to investigate efficacy of multiple treatments for a specific condition and to evaluate different treatments that are not directly compared within a single trial. An example of

this is the MTC performed by Cooper *et al.* (2006), which evaluated nine anti-thrombotic treatments (eight active treatments plus placebo/no treatment) in an MTC for efficacy in prevention of strokes in patients with atrial fibrillation. This MTC analysis also considered adverse events, by investigating the incidence rate for fatal or major bleeding episodes associated with the different treatments. This study highlighted the use of MTC analysis where many treatments are not directly compared against each other; in this example two of the treatments were compared directly against only one other treatment, while the maximum number of treatments any individual treatment was compared against was five.

Bearing in mind this precedent for using MTCs with adverse events data, there is potential to extend the methods further in this area, which is the aim of this chapter. The individual patient data (IPD) meta-analysis using etanercept as a therapy for rheumatoid arthritis (RA), with malignancy as the outcome (Chapter 8), is extended in this chapter, by using an MTC analysis to incorporate data on other anti-TNFs, and to include data regarding other aspects of treatment.

Due to the length of this chapter, an overview is provided here. Section 9.2 sets out the background to MTC methodology, whilst Section 9.3 discusses the potential for use of MTC methodology in the context of the scenario discussed in Chapter 8, that of anti-TNF drugs being used in rheumatoid arthritis, with arising concerns regarding increased risk of malignancy. This is followed by a specific description of the use of MTC models in this chapter, in Sections 9.4 and 9.5. The methods used to derive the dataset used for the analyses are described in Section 9.6. Finally, the results are set out and discussed in Sections 9.7 and 9.8. There is a summary of the chapter in Section 9.10. Network diagrams for all the MTC models used are included at the end of the chapter. The MTC methodology employed by this chapter is extended in ways that are novel to adverse events meta-analysis in Chapter 10.

9.2 Mixed treatment comparisons methodology

9.2.1 Baseline methods for mixed treatment comparisons

A detailed description of the use of MTCs has been provided by Lu & Ades (2004). These authors point out three situations that often create an MTC

scenario. The first of these is when there is no direct evidence relating an intervention to an outcome (for example, when an intervention is associated with an outcome in a chain of events, but not with subsequent outcomes in the chain). The second is when there are direct comparisons for a particular treatment comparison and a specific outcome, but this evidence is not substantial enough to provide a robust statistical analysis. In this situation, it is desirable to 'borrow strength' from indirect comparisons relating the two treatments, as described by Higgins & Whitehead (1996). The third situation occurs when there are multiple treatments and all need to be simultaneously compared or ranked against each other.

Similarly to direct meta-analyses, an MTC can assume either that all trials have the same underlying treatment effect (fixed effect (FE) meta-analysis) or that the true underlying treatment effects for all studies are derived from a common distribution (random effects (RE)). An alternative way to describe an RE analysis would be to argue that the underlying treatment effects for all primary studies are exchangeable, in that none of the primary studies 'stands out' from the rest in any way (discussed further in Section 3.5.1).

Furthermore, an MTC makes the assumption that indirect comparisons of treatments are the same, or from the same distribution, as direct comparisons of the same treatments. This assumption holds even where no direct head-to-head comparisons have been conducted between two treatments. This requirement is in practice difficult to validate. Another assumption relating to the MTC models is that the heterogeneity parameter for all relative treatment effects is the same, but this assumption may not be valid if there are certain treatments with more variability between studies. However, the assumption can be relaxed, if there are clinical reasons for doing so (e.g. a specific treatment may be more variable in its effects across different populations). It would be possible to investigate this assumption for each dataset, but Higgins & Whitehead (1996) have stated that they believe it unlikely that investigation would yield sufficient evidence to reject this assumption.

The basic premise of an MTC is as follows (Lu & Ades 2004): suppose there are three treatments, A, B and C, and θ_{AB} is the true underlying treatment effect such as a log odds ratio (OR) between two treatments, A relative to B, and $\hat{\theta}_{AB}$ is the estimated log OR for Treatment A compared to Treatment B. This treatment effect $\hat{\theta}_{AB}$ is directly estimated from the trials that include both

Treatment A and Treatment B. However, if there are also in existence trials that compare A against C, and B against C, the value of $\hat{\theta}_{AB}$ can be indirectly estimated (denoted as $\tilde{\theta}_{AB}$) by using the result:

$$\tilde{\theta}_{AB} = \hat{\theta}_{AC} - \hat{\theta}_{BC}. \tag{9.1}$$

However, the indirect estimate of Treatment A against treatment B results in higher degree of uncertainty about the estimated parameter compared to a direct estimate. We have (subject to certain caveats such as lack of correlation between trial groups and the estimates of treatment effect having the same precision):

$$\text{Var}(\tilde{\theta}_{AB}) = \text{Var}(\hat{\theta}_{AC}) + \text{Var}(\hat{\theta}_{BC}) > \text{Var}(\hat{\theta}_{AB}). \tag{9.2}$$

This implies that if there is positive correlation between the direct estimates of treatment effect, for example between $\hat{\theta}_{AC}$ and $\hat{\theta}_{BC}$ in Equation 9.2, this can be used in an MTC model to reduce the uncertainty in the indirect estimate, $\tilde{\theta}_{AB}$.

If there are two treatments to be compared, a full Bayesian two-level hierarchical model can be written, where the estimated parameter is on the log OR scale (Lu & Ades 2004, citing Smith *et al.* 1995). We have:

$$r_i^T \sim \text{Bin}(p_i^T, n_i^T), \tag{9.3}$$

and

$$r_i^C \sim \text{Bin}(p_i^C, n_i^C), \tag{9.4}$$

where r_i^T is the number of events in the treatment group, r_i^C is the number of events in the control group, n_i^T is the number of participants in the treatment group, n_i^C is the number of participants in the control group and i indexes the trials. Then:

$$\text{logit}(p_i^T) = \mu_i + \delta_i/2, \tag{9.5}$$

$$\text{logit}(p_i^C) = \mu_i - \delta_i/2, \tag{9.6}$$

and

$$\theta_i \equiv \delta_i \sim \text{Normal}(d, \tau^2), \tag{9.7}$$

where μ_i is the average log odds (on the logit scale) for an event in trial i, δ_i is the estimated log OR for the treatment group relative to the control group for study i (and is equivalent to $\hat{\theta}$ in Equations 9.1–9.2) and d is the pooled mean treatment effect for treatment relative to control, pooled across all studies, and τ^2 refers to the between-studies variance for log OR. All that is now required to apply this model in Bayesian software such as WinBUGS is the setting of a prior distribution for the stochastic parameters μ_i, d and τ^2.

9.2.2 Extension of mixed treatment comparison model

This model can be extended to cases where there are K treatments. If Treatment 1 is used as the baseline, then:

$$r_{ik} \sim \text{Bin}(p_{ik}, n_{ik}), \tag{9.8}$$

and:

$$\text{logit}(p_{i1}) = \mu_i - \delta_{i2}/K - \delta_{i3}/K - ... - \delta_{iK}/K, \tag{9.9}$$

and:

$$\text{logit}(p_{i2}) = \mu_i + (K-1)\delta_{i2}/K - \delta_{i3}/K - ... - \delta_{iK}/K, \tag{9.10}$$

and so forth until:

$$\text{logit}(p_{iK}) = \mu_i - \delta_{i2}/K - \delta_{i3}/K - ... + (K-1)\delta_{iK}/K. \tag{9.11}$$

The values for the δ_{iK} values for each treatment $i2$ to iK are all distributed

normally, with a generic mean value for δ_K, d_k, $d_1 = 0$ and a variance Σ:

$$(\delta_{i2}, ..., \delta_{iK})^T \sim \text{Normal}(d, \Sigma). \tag{9.12}$$

In this case, the values of d_2–d_K are the population mean treatment effects relative to the baseline and Σ is the $(K-1) \times (K-1)$ variance–covariance matrix.

An additional advantage when using an RE model is that the correlation between trial arms in trials with three or more arms, and hence three or more distinct treatments, can be accounted for (Higgins & Whitehead 1996). In such a trial, the estimates involving common arms will not be independent. If it is assumed that heterogeneity parameters for each relative treatment effect are equal, then due to the non-independence of trial arms, the marginal treatment effects will be represented by a bivariate normal distribution. Using an example where there are three treatments, A, B and C, the relative treatment effects can be designated θ_{AB}, θ_{AC} and θ_{BC}. For each study, each relative treatment effect is distributed normally with a mean value specific to that treatment and with a common heterogeneity parameter, τ^2. For example, for treatment θ_{BC},

$$\theta_{BC_i} \sim \text{Normal}(\mu_{BC}, \tau^2), \tag{9.13}$$

where i indexes an individual study.

To connect Equation 9.13 with Equations 9.9–9.11:

$$\theta_{BC_i} \equiv \delta_{i2} - \delta_{i3}, \tag{9.14}$$

where Treatment B is equivalent to Treatment 2 and Treatment C is equivalent to Treatment 3.

Also

$$\mu_{BC} = \mu_{AB} - \mu_{AC}, \tag{9.15}$$

hence the covariance between any two θ_i values will be $\tau^2/2$, by using a standard result for covariance. The resulting bivariate normal distribution for two of the three relative treatments, θ_{AB} and θ_{AC} is:

$$\begin{pmatrix} \theta_{AB_i} \\ \theta_{AC_i} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{AB} \\ \mu_{AC} \end{pmatrix}, \begin{pmatrix} \tau^2 & \tau^2/2 \\ \tau^2/2 & \tau^2 \end{pmatrix}\right). \quad (9.16)$$

With this result, the remaining distribution of treatment θ_{BC} is defined. These results can then be used to account for correlation between arms in the same trial. The methodology of MTCs is extended further in Section 9.4, where the specific methods used within the analyses conducted within this chapter are discussed.

It is important to note that studies with only one treatment group should be excluded from an MTC (Lu & Ades 2004). As in the case of a meta-analysis comparing only two treatment arms, such studies do not include within themselves an estimate of comparison between groups and cannot fulfil the criteria required for fixed or random effect(s) among studies with more than one group.

9.2.3 Discussion of graphical networks for mixed treatment comparison analyses

The graphical networks that are formed for the purpose of MTC analysis have been discussed in detail by Salanti *et al*. (2008). These networks facilitate the use of an MTC meta-analysis. They are constructed by creating a graph, with each treatment forming a node of the graph. Treatments are connected when there is a primary study directly comparing the two treatments. If there are any treatments not directly compared with any other treatment within the network, then this treatment is not connected to the rest of the network and cannot be included in an MTC. A connected MTC network diagram ensures that the randomisation within the primary studies is maintained, whilst allowing inclusion of all available comparisons between treatments (Cooper *et al*. 2006).

The number of primary studies directly comparing any two treatments can also be noted in the network (Salanti *et al*. 2008). Treatments may also be combined to 'collapse' the number of nodes, where this is clinically reasonable, for example by combining drugs of the same pharmacological class.

These authors point out two extreme styles of network layout, one where all treatments are compared against a common comparator, but not against each other, resulting in a star pattern. At the other extreme the graph is fully

connected, with all treatments compared against all others. Networks may be very complex, usually based on the number of treatments. There may also be asymmetries in the network, due to the fact that different treatments may be more heavily represented than others, and certain comparisons may occur more frequently within trials than others.

Of interest in the field of both efficacy and adverse events, direct comparisons between active treatments may be avoided, as pharmaceutical companies may have concerns regarding direct comparisons with a competitor treatment. Lack of direct comparisons results in increased difficulty in determining which treatment is preferable. The authors conclude that awareness of the geometry of a treatment network can inform on whether the correct data are available for decision-making in each situation, and it is particularly important to be aware when there is no direct evidence for clinically important comparisons.

The MTC networks used in this chapter are set out in Section 9.11.

9.2.4 Further discussion of mixed treatment comparison analyses

Additional discussion of MTC analyses in the context of making evidence-based decisions has been provided by Caldwell *et al*. (2005). These authors point out that a situation may arise where direct evidence is inconclusive but indirect evidence is not [resulting in a situation where it is unclear whether to base a decision on the direct or indirect evidence]. Also, the numbers of direct and indirect comparisons increases as the number of individual treatments rises. In the light of these situations, it is sensible to combine both direct and indirect comparisons in one model.

A potential way of conceptualising the assumptions of an MTC is to view all trials as having included all treatments, but for each trial the majority of treatments were lost at random (Caldwell *et al*. 2005). For a fixed effect model it can then be assumed that for all treatment effects comparing one trial to another, the effect is the same for all studies. The analogous assumption for a random effects model is that each relative treatment effect is derived from a common distribution for all trials. These assumptions are similar to those for a standard pairwise meta-analysis, with the subtle distinction that the assumption is considered to hold for all trials and all treatments, regardless of which treatments that trial actually did include.

In a decision-making context, it would be helpful to consider whether all trials should be used, for example if a decision is to be applied across a wide range of patient groups, or if a subgroup of trials, related to a specific patient group, should be included in an MTC. When judging which trials are to be included in an MTC, it is unlikely that assumptions relating to an MTC could be assessed statistically, hence clinical judgement may be required when deciding which studies to combine. If studies are included inappropriately, this may add heterogeneity to the MTC. It is also assumed for an MTC that the outcome measures are on an additive scale such as log OR or risk difference.

This section has provided an overview of MTC models. Each model used in the example discussed below is set out in detail in Section 9.5.

9.3 Mixed treatment comparisons used to investigate anti-TNF treatments in rheumatoid arthritis and their association with malignancy

The IPD analysis of patients with RA patients using etanercept (see Chapter 8) generated further questions about how the use of anti-TNFs may impact on risk of malignancy in users with RA. Interest centres on the potential effect of each individual drug, and on the possible effects of anti-TNFs in general.

Besides etanercept, the two main anti-TNFs used in RA are adalimumab and infliximab. Although trials have been performed using both of these drugs, the data are not available on an individual patient basis, with details of time to malignancy for each patient. Rather, the data are only available in aggregate form, with numbers of malignancies for each arm of a trial. Furthermore, the anti-TNFs are trialled against only placebo or active disease modifying anti-rheumatic drugs (DMARDs) as a comparison group; there are no trials that directly compare two or more different anti-TNFs.

There are several areas of interest in investigating the potential association of anti-TNFs with the adverse event of malignancy, listed below.

1. Does use of anti-TNF drugs as a class by RA patients have an association with malignancy?

2. Of the three anti-TNFs under consideration, etanercept, adalimumab and infliximab, are there any with a lower or higher risk than the others?
3. Does the use of anti-TNFs in conjunction with DMARDs influence malignancy risk?
4. Is there any relationship between dose of anti-TNF and risk of malignancy?

These queries require the data to be analysed in several ways to elicit the maximum information from the dataset. One method by which these questions can be addressed is through the use of MTCs, described above. The use of MTCs is particularly appropriate in this situation as we wish to make comparisons between different anti-TNFs that are not compared directly within a trial. However, there are certain methodological issues, relevant to this dataset, that can also be investigated by means of this example. These issues, set out below, are particularly pertinent to adverse events, but could be applied to any situation with sparse events.

1. Do MTC analyses with fewer treatment nodes (i.e. the same data but with treatments combined into fewer nodes) result in narrower confidence intervals?
2. Is there a level of separation of treatments at which the confidence intervals become excessively wide, indicating a very high degree of uncertainty?
3. What degree of treatment combination facilitates the 'best' use of data, in terms of deriving clinically useful outcomes counterbalanced against statistical uncertainty?
4. Is there a level of treatment separation (creating a model with increasingly narrow definitions of treatments, and hence increasing numbers of nodes, as opposed to combining treatments into fewer nodes) at which the MTC becomes impossible to fit?

These issues are addressed with regard to the motivating example of anti-TNF drugs and malignancy in Sections 9.7 and 9.8. The use of hierarchical modelling, for example with reference to issues regarding dose, is included in Chapter 10.

9.4 Statistical methods 1: meta-analysis models including mixed treatment comparisons

9.4.1 Outline of models

Given that there are several aspects of data analysis, as set out in Section 9.3, requiring both direct (for example within one study with different dose regimes for the same anti-TNF) and indirect (for example between studies using different anti-TNFs) comparisons, it was decided to use MTC methods, as described in Section 9.2.

The models used in this study are described by Ades *et al.* (2007)[1], and include:

1. A. Fixed effect model;
2. B. Random effects model; and
3. C. Random effects model with correlation for arms within same trial.

These models are discussed below. Extensions to these models are the subject of Chapter 10.

9.4.2 A. Fixed effect model

In this model, no account is taken for correlation between multiple arms in the same trial, i.e. all trial arms are considered independent.

The basic FE model is as follows:

$$r_i \sim \text{Binomial}(p_i, n_i), \tag{9.17}$$

where r_i refers to the number of events for each trial arm (i) denoting the datapoint, in this case based on a trial arm, p_i is the probability of a malignancy in the ith trial arm, and n_i is the number of participants in the ith trial arm.

[1] Ades *et al.* (2007). *Introduction to mixed treatment comparisons*. Available online [February 2010] at: `https://www.bris.ac.uk/cobm/research/mpes/mtc.html`

The logit of p_i is given by:

$$\text{logit}(p_i) = \mu_{s,i} + d_t - d_b, \tag{9.18}$$

where $\mu_{s,i}$ refers to the study level log odds of an event, for the study s to which arm i belongs, while d_t is the treatment effect (log OR) for the treatment t, and d_b is the treatment effect for the baseline treatment (e.g. a placebo or standard treatment) in that study (both compared to the baseline treatment). When considering the study arms receiving the baseline treatment for a particular study, $\mu_{s,i}$ is the log odds of an event for the baseline group.

The next stage of the analysis is to calculate the log ORs for each two-treatment comparison, and then to determine the probability for each treatment of being the 'best' or 'worst'. Note that where the outcome of interest is an adverse event, the 'best' treatment has the smallest treatment effect, and the 'worst' treatment has the largest treatment effect.

This is achieved by first determining the absolute log odds of malignancy for the baseline treatment, taking the mean value for μ_1, the log odds of an event for the baseline treatment, denoted Treatment 1. This can then be averaged across the total number of studies where Treatment 1 is used. For each treatment of interest (non-baseline treatment), the logit of the treatment effect T_k is given by:

$$\text{logit}(T_k) = \text{average treatment effect for baseline treatment} \quad + d_k, \tag{9.19}$$

where the treatments are indexed by k.

Having determined the absolute treatment effects for each treatment, these can be ranked according to magnitude, and the probabilities of being 'best' or 'worst' calculated across all iterations of the WinBUGS model.

Using the values for the log ORs (d_k) for each treatment k compared to the baseline treatment, it is then possible to calculate the log ORs for each pairwise combination of treatments, simply by subtracting the relevant log ORs for each treatment compared to the baseline.

To implement this model in WinBUGS, a prior distribution is required for the μ_j parameters, which refers to the study-level effect (odds of an event in the base-

line arm), where the study is indexed by j; a vague distribution is appropriate, for example:

$$\mu_j \sim \text{Normal}(0, 10\ 000). \tag{9.20}$$

A prior distribution is also needed for the log ORs for each treatment effect, d_k; for d_1, the treatment effect for the baseline treatment, the log OR is set to 0, as this is the reference treatment by which all the others are evaluated. For example:

$$d_k \sim \text{Normal}(0, 1000), \tag{9.21}$$

where k indexes the treatments.

Another feature of this model is the ability to calculate the deviance for each datapoint, which can then be summed and compared to the total number of datapoints to provide an evaluation of the goodness of fit of the model. The sum of the deviance complements the deviance information criterion (DIC) when assessing model fit, the DIC being useful in model comparison, whilst the sum of the deviance can point to an absolute goodness of fit. (The DIC is discussed further in Section 4.4.1.)

In this case, the deviance residuals for each datapoint i are given by:

$$\hat{r_i} = \hat{p_i} \times n_i. \tag{9.22}$$

These can then be used to calculate the deviance for each datapoint, on the binomial distribution (based on McCullagh & Nelder (1989):

$$\text{deviance}_i = 2 \times (r_i \times (\log(r_i) - \log(\hat{r_i})) + (n_i - r_i) \times (\log(n_i - r_i) - \log(n_i - \hat{r_i}))). \tag{9.23}$$

These deviances can then be summed to give the overall sum of deviance.

9.4.3 B. Random effects model

The FE model makes the assumption that the underlying treatment effect for each individual treatment remains the same across all studies. Where this assumption is not considered appropriate, a random effects model is often used

instead. The assumption for an RE model is that all the treatment effects are derived from a common distribution and hence are not automatically the same.

The RE model can be derived very easily from the FE model.

As before, the number of events r_i are distributed binomially for each datapoint, i:

$$r_i \sim \text{Binomial}(p_i, n_i). \tag{9.24}$$

The logit of p_i is as follows for the baseline treatment:

$$\text{logit}(p_i) = \mu_{s,i}, \tag{9.25}$$

and for the non-baseline treatments:

$$\text{logit}(p_i) = \mu_{s,i} + \delta_i. \tag{9.26}$$

Hence, as for the FE model, $\mu_{s,i}$ refers to the study level log odds of malignancy for the baseline group. For the non-baseline treatment, a random quantity, δ_i is added to $\mu_{s,i}$. This quantity δ_i is distributed as follows:

$$\delta_i \sim \text{Normal}(md_i, \tau^2), \tag{9.27}$$

and:

$$md_i = d_{t,i} - d_{b,i}, \tag{9.28}$$

where $d_{t,i}$ refers to the log odds of an event in the treatment arm for study i, and $d_{b,i}$ refers to the log odds of an event in the baseline arm for study i.

Prior distributions are required for the μ_j parameters, where j indexes the study. As previously a vague prior is appropriate:

$$\mu_j \sim \text{Normal}(0, 10\ 000). \tag{9.29}$$

A prior distribution is also required for the value of τ pertaining to the normal distribution for δ_i:

$$\tau \sim \text{Uniform}(0, 2). \tag{9.30}$$

Prior distributions on the pooled (across treatments) log ORs (d_k values) are as for the FE Model described above. The other elements of the RE model reflect the FE model, including the calculation of absolute treatment effects, rankings and deviance calculations.

9.4.4 C. Random effects model with correlation for arms within same trial

The RE model described above does not make any provision for the fact that in trials with three or more arms, the arms themselves will not be independent from one another. Therefore, this non-independence should be accounted for within the MTC model. This method is described by Ades *et al*. (2007); (Footnote 1, page 217).

The following method was used to adjust for non-independence in multi-arm trials. The binomial likelihood is described as follows, for each study and trial arm:

$$r_{i,k} \sim \text{Binomial}(p_{i,t_{i,k}}, n_{i,k}), \tag{9.31}$$

where number of events r is indexed by study i and study arm k, probability of an event p is indexed by study and treatment t corresponding to trial i arm k, and number of participants is indexed by study i and study arm k.

For all trials, the adjustment for the baseline treatment arm with itself $w_{i,1}$, was set to equal 0. The log OR for the baseline treatment (arm 1) in all trials was also set to equal 0.

$$\delta_{i,t_{i,1}} = 0. \tag{9.32}$$

The values of the log ORs for each trial can then be set for the non-baseline trial arms across all studies as follows:

$$\delta_{i,t_{i,k}} \sim \text{Normal}(md_{i,t_{i,k}}, \tau^2_{i,t_{i,k}}), \tag{9.33}$$

$$md_{i,t_{i,k}} = d_{t_{i,k}} - d_{t_{i,1}} + sw_{i,k}, \tag{9.34}$$

$$\tau^2_{i,t_{i,k}} = \tau^2 \times 2 \times (K-1)/K. \tag{9.35}$$

In the above equations, K is the total number of arms, d is the mean underlying log OR for each treatment d_t (indexed by study i and arm k), where $d_0 = 0$, and md is the mean underlying log OR for each study i, and the specific treatment arm $t_{i,k}$ within that study. In a similar vein, $\tau^2_{i,t_{i,k}}$ refers to the variance for each study-level underlying log OR (md).

For each non-baseline treatment arm, the adjustment w is calculated as follows:

$$w_{i,k} = (\delta_{i,t_{i,k}} - d_{t_{i,k}} + d_{t_{i,1}}), \tag{9.36}$$

with $w_{1,1} = 0$, and then the cumulative adjustment sw across trial arms within a trial is calculated as follows:

$$sw_{i,k} = \frac{\sum_{k=1}^{k} w_i}{(K-1)}. \tag{9.37}$$

Priors are now required for the distributions of baseline parameters, again, vague distributions are appropriate, for example:

$$\mu_i \sim \text{Normal}(0, 10\ 000), \tag{9.38}$$

$$d_1 = 0, \tag{9.39}$$

$$d_k \sim \text{Normal}(0, 10\ 000), \tag{9.40}$$

$$\tau \sim \text{Uniform}(0, 2), \tag{9.41}$$

where d_k is the treatment log OR compared to the baseline treatment d_1.

As in the previous models, deviance residuals can be calculated across all data-points, but in this model the deviances need to be summed across the individual trial arms, then summed across trials, to provide the final overall sum of deviances.

Once the model is completed, the log ORs for each pairwise comparison of treatments and the 'best' and 'worst' rankings can be carried out as in the previous models.

9.4.5 Summary of mixed treatment comparison models

The MTC models described above are summarised below.

1. Fixed effect model: all treatment effects for individual treatments across studies are the same; no account of correlation between multiple arms in the same trial is made.
2. Random effects model: all treatment effects for individual treatments across studies are derived from the same underlying distribution, hence any observed differences are due to random error.
3. Random effects model with correlation across multiple treatment arms: all treatment effects for individual treatments across studies are derived from the same underlying distribution, hence any observed differences are due to random error; correlation between multiple arms is accounted for.

In the list above, a 'treatment' is considered to be a specific combination of variables that create an overall treatment regime, but the included variables may change across MTC models. For example, in Models 2a and 2b below (Section 9.5.1, Table 9.1) a 'treatment' refers to the specific anti-TNF received. In Models 5a and 5b, the 'treatment' refers to the specific anti-TNF, whether or not there is concomitant disease-modifying anti-rheumatic drug (DMARD) therapy, and the dose of the specific anti-TNF (at three levels). Hence, the definition of a 'treatment' changes across models and is more detailed in Models 5a and 5b compared to Models 2a and 2b.

Extended models including hierarchical modelling, with and without constraints, are considered in Chapter 10.

9.4.6 Implementation using WinBUGS

All models were implemented using WinBUGS 1.4. Three chains with different initial values were used for each model (unless otherwise stated), with assessment for convergence using the Brooks-Gelman-Rubin method (Brooks & Gelman 1998), as well as visual inspection of the trace. Convergence was confirmed prior to the selection of an adequate burn-in period which was always at least 10 000 iterations. Following burn-in at least 50 000 iterations were performed to provide the results for each model.

Prior distributions are as described in Sections 9.4.2–9.4.4. For the prior distribution on the d_k parameter, a normal distribution centred on 0, with a variance of 1000 was used for the models with no correlation by arm, or correlation across two arms only, whilst the variance was 10 000 for models with correlation across three or more arms.

9.5 Statistical methods 2: construction of mixed treatment comparison networks

9.5.1 Initial definition of required networks

Multiple MTC meta-analyses would be required due to the nature of the questions being asked; different methods of combining data within trials would create different treatment nodes in an MTC network and hence allow different treatments to be compared. Table 9.1 sets out the different MTC models, including information on whether the anti-TNFs would be considered individually or combined, whether the presence of an additional DMARD was included in the model, whether dose level was considered and whether the control group (non-anti-TNF group) was taken as placebo only with placebo plus additional DMARD as a separate treatment in the MTC (resulting in two non-anti-TNF treatment nodes), or placebo with or without DMARD was considered as one treatment in the MTC.

Dose level was recorded according to the classification set out by Leombruno *et al*. (2008). Dosage regimes varied between studies, and it was decided to classify dose according to total weekly dose. For example, adalimumab administered at 40mg once weekly would be considered as recommended dose, as would 20mg administered twice weekly. This approach was used to avoid 'overstretching' already 'thin' data into a larger number of treatment nodes in the MTC model.

To further clarify the terminology used in Table 9.1, in the column headed 'Anti-TNFs' for Model 1a/b only, the anti-TNFs are combined across the studies to give a basic meta-analysis of anti-TNF against non-anti-TNF controls. In all other models, the three anti-TNFs are treated as separate nodes across the MTC.

Table 9.1: Mixed treatment comparison models (for 13 studies with a total of 44 treatment arms when uncombined).

MTC model number	Anti-TNFs	Other DMARD	Dose	Control	No. arms	No. treatment nodes	Potential no. treatment nodes
1a	Comb	No	No	Comb	26	2	2
1b	Comb	No	No	P/D	26	3	3
2a	Indiv	No	No	Comb	26	4	4
2b	Indiv	No	No	P/D	26	4	5
3b	Indiv	P/D	No	P/D	29	7	8
4a	Indiv	No	Yes	Comb	35	8	10
4b	Indiv	No	Yes	P/D	35	9	11
5a	Indiv	P/D	Yes	Comb	38	12	19
5b	Indiv	P/D	Yes	P/D	38	13	20

Comb: combined; Indiv: individual; P/D: Placebo/DMARD.

For the column headed 'Other DMARD', a 'No' in this column indicates that it is not considered within the MTC nodes whether an anti-TNF is administered with or without concomitant DMARD therapy. For example, adalimumab alone would be combined with adalimumab plus methotrexate within the same node. Alternatively P/D indicates that an anti-TNF administered with placebo (or no other active treatment) would be considered as a different treatment node compared to an anti-TNF administered with a DMARD.

The Dose column indicates whether dose is included when determining the MTC nodes; a Yes indicates that nodes are defined by dose as well as drug(s). The Control column indicates whether the non-anti-TNF treatments are combined across studies into one node, regardless of whether the non-anti-TNF group includes a DMARD or is placebo only (note that for all studies there is only one non-anti-TNF node). A 'Combined' entry in this column indicates that DMARD and non-active controls are combined into one, while a 'P/D' entry indicates that placebo and DMARD are separate nodes.

The number of arms is the total number of treatment arms (and hence datapoints) across the entire dataset of 13 trials, while the number of treatment nodes refers to the number of nodes within each MTC model; clearly this quantity increases as the complexity of the MTC increases. The potential number of treatment nodes refers to the maximum number of nodes that would exist if each *potential* treatment node was in existence. The maximum number of treatment nodes possible would only occur if each anti-TNF was present with

and without concomitant DMARD therapy, and present at all three dose levels for both of these treatment types. In practice however, the maximum number of treatment nodes are not present when DMARD therapy and/or dose are taken into account in the MTC.

All of the models set out in Table 9.1 can be analysed using the FE and RE models discussed above. The hierarchical models used in these analyses (see Chapter 10) also use the baseline Models 1a–5b as described in Table 9.1.

The 'goodness of fit' of these models is assesed by using the sum of residual deviance (Section 9.4.2), as well as the DIC and pD values, discussed in Section 4.4.1.

9.5.2 Detailed model descriptions

From Table 9.1, it is evident that the extracted data can be analysed at several levels of combination of the data across studies. Increasingly complex models allow the investigation of increasingly specific treatments and the comparisons between them. However, with sparse data such as in this case, creating a complex model with increasingly specific treatments means that the sparse data are spread more 'thinly' across the treatment nodes, with associated reduced power, increased uncertainty in point estimates, and possibility of the model failing to fit. The separate models are discussed below.

Model 1a: All anti-TNFs combined against all non-anti-TNF controls

By combining all three anti-TNFs, regardless of dose or any additional DMARD and comparing this treatment arm against any non-anti-TNF arm (either a placebo or DMARD control), each trial is reduced to two arms, resulting in 26 arms in total. (All MTC diagrams for all models presented in this chapter are shown in Figures 9.2 to 9.9, presented at the end of the chapter. See Figure 9.2 for Models 1a and 1b.)

There are potentially two treatment nodes, which is the minimum number of treatments required for an MTC in order to provide at least one treatment comparison. In this MTC, all studies have at least one anti-TNF arm and at least one non-anti-TNF arm, therefore all studies can theoretically be included within all MTC networks.

Studies with zero events have been excluded from the MTC, as these are thought not to contribute to the overall estimates, considering the work of Sweeting *et al*. (2004), discussed in Section 5.2.4. There were six studies with zero events in one of the two arms; interestingly, the arm with zero events was the control arm across all six studies.

Model 1b: All anti-TNFs combined against placebo only or DMARD controls

As for Level 1a, there were in this case 26 arms, potentially with three treatment nodes and in actuality there were three treatment nodes.

Model 2a: All individual anti-TNFs against all non-anti-TNF controls

At this level there were potentially four treatment arms, and inevitably four treatment arms existed as there were three anti-TNFs and each was compared against non-anti-TNF controls (Figure 9.3). As none of the studies included more than one anti-TNF it was impossible to have 'cycles' within the MTC network (or alternatively to have more than two arms per trial). Hence, the MTC network resembles a 'star' in the classification developed by Salanti *et al*. (2008).

Model 2b: All individual anti-TNFs against all placebo only or DMARD controls

In this model (Figure 9.4) there were potentially five arms and in total five arms. In this case, the network was more widely spread and less connected, as the infliximab trials were connected only to DMARD controls and not to placebo-only controls.

Model 3b: All individual anti-TNFs with or without additional DMARD against placebo only or DMARD controls

At this level the models become potentially more complex (Figure 9.5). For the first time there are several studies with potentially three study arms. This would include a control arm, a treatment arm without additional DMARD and a treatment arm with additional DMARD. There were no studies that had more than one non-anti-TNF arm, and, as the anti-TNF arms were being considered alone or with an additional DMARD, it was considered appropriate to separate the placebo controls from the DMARD controls for this model (hence there is no Model 3a). The MTC network now becomes more complex with two cycles

of three treatment nodes that are all interconnected. Due to the fact that there are three study arms for two trials in this model, for the first time there is the potential for a comparison with 0 events but this does not happen. There is, however, one comparison deriving data from two trials where one trial has zero events within this comparison. The potential number of treatment nodes is not achieved due to the fact that one anti-TNF (infliximab) is only administered in the presence of another DMARD (methotrexate in both cases).

For all models where there are three or more treatment arms within the same study, the correlation between multiple arms in the same trial is included in the model.

Model 4a: All individual anti-TNFs divided by dose level against all controls

Each anti-TNF is potentially divided into three dose levels, recommended, low and high (Figure 9.6). In fact, only adalimumab was present at all three doses, with etanercept present at recommended and low doses, and infliximab present at recommended and high doses. Hence, there were in actuality eight treatment nodes, with a potential for ten. Despite there being three trials with three arms, there were no comparisons with zero events for this model, although there were six distinct comparisons where one of the trials with data for this comparison included zero events.

Model 4b: All individual anti-TNFs divided by dose level against placebo only or DMARD controls

For this model there is an additional treatment node compared to Model 4a. This has the effect of making the network more complex (Figure 9.7), and for the first time the model breaks the studies down into a sufficiently dispersed format to allow the presence of a comparison with only one study and zero events within that comparison. This comparison is adalimumab at low dose against DMARD. There are also five other comparisons where one trial has zero events within that comparison. It is at this level, where dose is included within the dataset and the controls are separated into active and inactive that the sparsity of the events becomes a point of concern in the analysis.

Model 5a: All individual anti-TNFs with or without individual DMARD and divided by dose level against all controls

At this level, the treatment nodes are specified by both dose and additional DMARD, resulting in a model with the highest degree of complexity, the greatest number of treatment nodes and potentially the greatest sparsity of data as the available events are divided between a larger number of treatment nodes (Figure 9.8). In this model there are 12 treatments with a potential of 19 treatments. Despite the control in this model being either placebo only or DMARD only, there are two treatment comparisons with zero events, where data are available from one trial only. Also, in five additional comparisons, there is one study with zero events.

Model 5b: All individual anti-TNFs with or without individual DMARD divided by dose level against placebo only or DMARD controls

For this model (Figure 9.9) the control groups are divided into placebo only and DMARD only. This division results in 13 actual treatments with a total of 20 potential treatments. The number of comparisons with zero events is two, as for Model 5a. There are again five comparisons where one trial of the total number has zero events.

9.6 Dataset creation

9.6.1 Data sources

The aggregate data from anti-TNF trials with malignancy information were gathered from earlier reviews.

The study by Bongartz *et al*. (2009) provided data from nine etanercept trials, eight of which were published (either as a journal paper or an abstract) and one of which was unpublished. This publication is the subject of the IPD analysis discussed in Chapter 8.

The more recent meta-analysis on anti-TNFs by Leombruno *et al*. (2008) also provided data on seven studies on etanercept, all except one (Keystone *et al*. 2004a) of which is included in Bongartz *et al*. (2009). Hence, there were potentially ten primary trials regarding etanercept identified.

The other meta-analysis by Bongartz *et al*. (2006) included data on five published studies on adalimumab and four on infliximab. The meta-analysis by Leombruno *et al*. (2008) also provided data on six adalimumab studies, one of which (Breedveld *et al*. 2006) was not included in the previous meta-analysis by Bongartz *et al*. (2006). Also in Leombruno *et al*. (2008) were data from five studies using infliximab, only one of which (Abe *et al*. 2006; cited by Leombruno *et al*. 2008) was not included by Bongartz *et al*. (2006), either as a cited reference or as a later publication.

Hence, in total there were potentially six studies with data on adalimumab, and five for infliximab.

9.6.2 Data selection

Two studies included a design where participants' treatment was allowed to be altered at a certain point in the trial. These studies were Keystone *et al*. (2004a) which concerned etanercept and Westhovens *et al*. (2006) which concerned infliximab. From the available data (derived from the original reference), it was not always possible to determine at which point during the trial or what treatment a participant was receiving when a malignancy occurred, making it impossible to assign a treatment to the event. It was therefore decided to exclude these studies to avoid incorrect classification of malignancies.

It was also decided to exclude from the meta-analyses all studies that did not include a single malignancy case across all treatment arms. These studies would not contribute meaningful information to the analysis, (due to lack of a point estimate for the study, as the analyses were being conducted on the OR scale) and their exclusion avoided difficulties in calculating confidence intervals for these studies. These studies with zero overall events were, for etanercept, study TNR 00102 (unpublished; cited in Bongartz *et al*. 2009), Moreland *et al*. (1997) and Weinblatt *et al*. (1999); for adalimumab, Van de Putte *et al*. (2003); and for infliximab, Maini *et al*. (1998) and Abe *et al*. (2006). (The data regarding the lack of events was derived from the reviews, Bongartz *et al*. 2006; 2009; and Leombruno *et al*. 2008). All other studies included at least one event in at least one arm.

A risk difference scale would allow the calculation of a point estimate for studies with zero events, but due to low baseline risk, a relative scale would be better

placed to demonstrate any potential signal from the data. Hence, it was decided to use the OR scale, with exclusion of studies with zero events. This is in accordance with the work of Sweeting *et al.* (2004) which indicated that studies with zero events do not contribute to the overall dataset for a meta-analysis (discussed in Section 5.2.4).

This resulted in an overall set of primary studies that included data from six studies involving etanercept, five involving adalimumab and two involving infliximab. All studies were conducted using a placebo or a DMARD that was not an anti-TNF for the comparison group. Some studies also used different doses of anti-TNFs, and this aspect of treatment was included in some of the MTC models.

9.6.3 Data extraction

Throughout the data extraction process, to promote simplicity and ease of data extraction and to reduce the number of sources used, data were extracted from a review, with a more detailed examination of cases where there were discrepancies. If there were insufficient data in any review to allow accurate extraction for a specific primary study, it was then necessary to derive the data from the original paper.

Regarding the etanercept studies, all data were extracted from the meta-analysis by Bongartz *et al.* (2009), because in this study the authors had used IPD and had followed up individual cases of malignancy in order to verify each event. An exception to this was the reference by van der Heijde *et al.* (2006), for which neither the Bongartz *et al.* (2009) review nor the Leombruno *et al.* (2008). review reported the malignancy outcome by trial arm in sufficient detail; fortunately, for this primary study, the number of malignancies reported in the primary reference coincided with the total reported in the two reviews. Another exception was the reference by Combe *et al.* (2006), for which the Bongartz *et al.* (2009) review reported only one malignancy but did not report whether it was in the etanercept only or etanercept plus DMARD arm; this issue had to be resolved by referring to the original paper.

Data on adalimumab studies were extracted from the review by Bongartz *et al.* (2006). This was because in some cases the authors had discovered additional cases of malignancy not reported in the original publications and hence not

Table 9.2: Sources of data for analysis of anti-TNF use in rheumatoid arthritis and malignancy.

Original reference	Data source
Etanercept	
Ericson *et al.* 1999	Bongartz *et al.* 2009
Moreland *et al.* 1999	Bongartz *et al.* 2009
Genovese *et al.* 2002	Bongartz *et al.* 2009
Combe *et al.* 2006	Primary study & Bongartz *et al.* 2009
van der Heijde *et al* 2006.	Primary study & Bongartz *et al.* 2009
Weisman *et al.* 2007	Bongartz *et al.* 2009
Adalimumab	
Furst *et al.* 2003	Bongartz *et al.* 2006
Weinblatt *et al.* 2003	Bongartz *et al.* 2006
Keystone *et al.* 2004b	Bongartz *et al.* 2006
Van de Putte *et al.* 2004	Bongartz *et al.* 2006
Breedveld *et al.* 2006	Primary study
Infliximab	
Maini *et al.* 2004	Primary study
St Clair *et al.* 2004	Bongartz *et al.* 2006

included by Leombruno *et al.* (2008). In other cases, there were discrepancies in reporting of primary studies between the two reviews. Regarding the references by Furst *et al.* (2003) and Keystone *et al.* (2004b), only the malignancies included by Bongartz *et al.* (2006) were included for this meta-analysis, for the same reasons as cited by Bongartz *et al.* (2006). The data from the study by Breedveld *et al.* (2006) had to be extracted from the original reference due to insufficient detail in reporting by Leombruno *et al.* (2008) for the purposes of this meta-analysis.

The infliximab data were extracted directly from the reference for the original study by Maini *et al.* (2004); this was because the study had been reported previously (Lipsky *et al.* 2000; cited by Bongartz *et al.* 2006) and therefore the more recent publication was preferred. Insufficient details were available in Leombruno *et al.* (2008), plus there were some inconsistencies in reporting compared to the original reference. Data from St Clair *et al.* (2004) were extracted from Bongartz *et al.* (2006).

Table 9.2 sets out the sources of data, by primary reference, the source of data used, and anti-TNF. (Note that studies not directly used to extract data are not included in the bibliography.)

Although all studies used a non-anti-TNF control, this control was in some studies placebo alone (no active treatment) and in other cases involved active

treatment with a DMARD, usually stated to be methotrexate or sulfasalazine, but in one study (Furst *et al.* 2003; cited by Bongartz *et al.* 2006) was simply referred to as DMARD, implying that the DMARD drug could vary between participants. Furthermore, there were studies that provided a direct comparison between two or more anti-TNFs.

The final dataset used in the following meta-analyses is set out below in Table 9.3. This dataset represents the dataset used in the most 'deconstructed' models, Models 5a and 5b. In these models, treatments including anti-TNFs were also defined by dose and whether an additional DMARD was being used. Some studies had multiple different regimes that would be classified as the same dose according to the scheme used in these meta-analyses. As the highest resolution in the models was by dose and not regime, the figures in Table 9.3 relate to dose, although regimes have been described for interest, but without the numbers of patients allocated to each regime.

The difficulties of deriving a dataset with coherency when there are multiple reviews (sometimes with access to different source data compared to primary references), as well as the original references, are highlighted by this example of deriving a dataset for malignancies occurring in conjunction with anti-TNF use in RA patients. In this case also, there were often multiple primary references related to a single clinical trial, further complicating the situation.

9.7 Results and initial discussion

9.7.1 Initial inspection of data

The full dataset for all 13 studies included in the analyses for this chapter is set out in Table 9.2.

The results for each model are presented individually, followed by a general discussion relating the results of all models. An initial analysis of the data shows that over 13 studies there were 76 malignancy events, with 7233 participants. This breaks down to 14 in the control treatments out of 2275 participants (0.006%), and 62 in the anti-TNF treatments out of 4958 participants (0.013%).

Table 9.3: Primary data used in mixed treatment comparison meta-analyses.

First author (Year)	Control (no anti-TNF)	DMARD	anti-TNF dose	anti-TNF regime(s)	Number	Cases
Etanercept						
Ericson (1999)	Y	N	NA	NA	105	0
"	N	N	Rec	25mg biw	111	0
"	N	N	Low	10mg qw or 25mg qw or 10mg biw	343	2
Moreland (1999)	Y	N	NA	NA	80	0
"	N	N	Rec	25mg biw	78	1
"	N	N	Low	10mg biw	76	0
Genovese (2002)	Y	Y	NA	NA	217	4
"	N	N	Rec	25mg biw	207	5
"	N	N	Low	10mg biw	208	5
Combe (2006)	Y	Y	NA	NA	50	0
"	N	N	Rec	25mg biw	103	1
"	N	Y	Rec	25mg biw	101	1
Van der Heijde (2006)	Y	Y	NA	NA	228	1
"	N	N	Rec	25mg biw	223	5
"	N	Y	Rec	25mg biw	231	5
Weisman (2007)	Y	N	NA	NA	269	2
"	N	N	Rec	25mg biw	266	2
Adalimumab						
Furst (2003)	Y	Y	NA	NA	318	0
"	N	Y	Rec	40mg eow	318	4
Weinblatt (2003)	Y	Y	NA	NA	62	0
"	N	Y	Rec	40mg eow	67	0
"	N	Y	Low	20mg eow	69	0
"	N	Y	High	80mg eow	73	1
Keystone (2004b)	Y	Y	NA	NA	200	1
"	N	Y	Rec	20mg qw or 40mg eow	419	8
Van de Putte (2004)	Y	N	NA	NA	110	1
"	N	N	Rec	20mg qw or 40mg eow	225	2
"	N	N	Low	20mg eow	106	1
"	N	N	High	40mg qw	103	1
Breedveld (2006)	Y	Y	NA	NA	257	4
"	N	N	Rec	40mg eow	274	2
"	N	Y	Rec	40mg eow	268	2
Infliximab						
Maini (2004)	Y	Y	NA	NA	88	1
"	N	Y	Rec	3mg/kg q8w	86	1
"	N	Y	High	3mg/kg q4w or 10mg/kg q8w or 10mg/kg q4w	254	8
St Clair (2004)	Y	Y	NA	NA	291	0
"	N	Y	Rec	3mg/kg q8w	372	0
"	N	Y	High	6mg/q8w	377	4

biw: twice weekly; eow: every other week; NA: not applicable; Rec: recommended; q4w: 4-weekly; q8w: 8-weekly; qw: weekly.

9.7.2 Model 1a

This is the simplest model with only two treatment arms. Using the fixed effect (FE) model the mean (of the posterior distribution) probability that the control was the 'best', producing the lowest risk of malignancy, was 0.997, while probability of being 'worst', with the greatest risk of malignancy, was 0.003. The 'best' and 'worst' probabilities were reversed for the anti-TNF group. Hence, the anti-TNF group appeared to be emphatically the treatment with the highest malignancy risk. The median (of the posterior distribution of the mean) log OR (LOR) for malignancy in the anti-TNF group compared to controls was 0.759, with a 95% credible interval (CrI) of 0.198 to 1.392. The mean (of the posterior distribution) sum of the deviance was 25.88, indicating that the model was a good fit (close to the number of datapoints, 26).

Using a random effects (RE) model the results were similar, though slightly less extreme. For the anti-TNF group the probability of being 'best' was 0.008, with a probability of being 'worst' of 0.992. Hence, the RE model tended to 'even out' the probabilities and provide less extreme results. The median LOR for malignancy for anti-TNFs compared to controls was 0.908 (95% CrI 0.176; 1.995), higher than in the FE model.

The mean standard deviation was 0.662, suggesting that the range of ORs would lie between 10.51 (corresponding to a standard deviation of 0.6) and 15.55 (corresponding to a standard deviation of 0.7). These ranges are derived from Spiegelhalter *et al*. (2004), and correspond to the ratio of the 97.5% point to the 2.5% point of the distribution of the OR. The sum of the deviance was 24.14, indicating a better-fitting model compared to the equivalent FE model.

9.7.3 Model 1b

This model is interesting as it allows a possible differentiation in cancer malignancy between placebo controls and those receiving an active DMARD. Using an FE model, the probability that the placebo group was the 'best' was 0.295, with a probability of being 'worst' of 0.279. The active DMARD group had a probability of being 'best' of 0.704 and a probability of being 'worst' of 0.003. This gives the impression that an active DMARD has some protective influence in reducing the risk of malignancy in RA sufferers, compared to those receiving

no active treatment. The probability of being 'best' for the anti-TNF group was 0.0007, with a probability of being 'worst' of 0.718. Hence, the ordering of the treatments by probability for the 'best' treatments was in reverse order compared to the 'worst' categories. For the 'best' category, from highest to lowest, the treatments were DMARD, placebo and finally anti-TNF. Comparing the anti-TNF group to the DMARD group, the median LOR was 0.847, (95% CrI 0.224; 1.562). The sum of the deviance was 26.55, indicating a slightly poorer fit compared to the model where the control group was combined to include both placebo only and DMARD groups.

The RE model produced similar probabilities. The 'best' and 'worst' probabilities for the placebo group were 0.312 and 0.250, for the DMARD group 0.686 and 0.007, and for the anti-TNF group 0.002 and 0.743. The median LOR for the anti-TNF group compared to the DMARD group was 1.046 (95% CrI 0.182; 2.381). The sum of deviance was 24.58, indicating that the RE model is again a better fit than the (equivalent) FE. The mean standard deviation was 0.739, indicating higher between-studies variability when a distinction was made between placebo and DMARD controls. This is an interesting point, as intuitively separating the two different types of non-anti-TNF control should account for some of the between-study variability, rather than increase it.

9.7.4 Model 2a

At this level, the main focus of interest is to determine which one of three anti-TNFs appears to be 'safest' with regard to malignancy, by comparison to any non-anti-TNF control.

The results for the FE and RE models are shown in Table 9.4.

For the FE model the sum of the deviance was 27.6. For the RE model the sum of the deviance was 25.06, again indicating that the RE model is a better fit across the 26 datapoints. The mean standard deviation for the RE model was 0.851. The most interesting aspect of these results is that the three anti-TNFs all have roughly equivalent probabilities of being 'best', but the probabilities for the 'worst' anti-TNF were more varied, with infliximab being clearly the 'worst' with a probability of 0.742 (FE) and 0.679 (RE), whereas the probabilities for adalimumab and etanercept were much lower. This high probability of being worst is reflected in the high median OR for infliximab.

Table 9.4: Model 2a: Results for fixed and random effect(s) models.

	P(best)		P(worst)		FE		RE	
Treatment	FE	RE	FE	RE	Median† LOR*	95% CrI	Median LOR	95 % CrI
Control	0.866	0.810	0.000	0.000	NA	NA	NA	NA
Etanercept	0.036	0.063	0.136	0.161	0.728	-0.075; 1.664	0.910	-0.314; 2.598
Adalimumab	0.066	0.076	0.121	0.160	0.668	-0.195; 1.685	0.875	-0.378; 2.599
Infliximab	0.033	0.052	0.742	0.679	1.707	-0.113; 5.063	1.928	-0.410; 5.655

†Median of posterior mean distribution; *baseline for LOR is Control; CrI: credible interval; LOR: log odds ratio; FE: fixed effect; RE; random effects.

Table 9.5: Model 2b: Results for fixed and random effect(s) models.

	P(best)		P(worst)		FE		RE	
Treatment	FE	RE	FE	RE	Median† LOR*	95% CrI	Median LOR	95% CrI
Placebo	0.310	0.322	0.065	0.064	NA	NA	NA	NA
DMARD	0.621	0.572	0.000	0.000	-0.3863	-1.926; 1.338	-0.449	-1.184; 1.987
Etanercept	0.012	0.023	0.138	0.179	0.843	-0.052; 8.246	1.109	0.603; 3.137
Adalimumab	0.031	0.039	0.107	0.142	0.739	-0.164; 1.785	0.984	0.509; 2.906
Infliximab	0.026	0.043	0.690	0.614	1.718	-0.108; 5.033	1.957	1.046; 6.016

†Median of posterior mean distribution; *baseline for LOR is DMARD for anti-TNFs and placebo for DMARD; CrI: credible interval; LOR: log odds ratio; FE: fixed effect; RE; random effects.

9.7.5 Model 2b

From Model 2a, it appears that infliximab is the anti-TNF associated with highest risk of malignancy. By dividing the controls into placebo only and DMARD, this may influence the results. The results for Model 2b are shown in Table 9.5.

Additionally, using etanercept as the baseline and the RE model, the median LOR for adalimumab was -0.123 (95% CrI -0.778; 2.047), with the median OR for infliximab being 0.843 (95% CrI -0.199; 5.016). This further supports the argument that etanercept and adalimumab are similar in their malignancy risk,

Table 9.6: Model 3b: Results for fixed and random effect(s) models.

	P(best)		P(worst)		FE		RE	
Treatment	FE	RE	FE	RE	Median† LOR*	95% CrI	Median LOR	95% CrI
Placebo	0.276	0.300	0.043	0.037	NA	NA	NA	NA
DMARD	0.488	0.374	0.000	0.000	-0.375	-1.945; 1.458	-0.291	-2.775; 2.366
Etanercept	0.010	0.012	0.069	0.104	0.476	-0.935; 2.18	0.839	-1.160; 3.409
Etanercept + DMARD	0.076	0.122	0.121	0.125	0.827	-0.585; 2.242	0.917	-1.565; 3.460
Adalimumab	0.095	0.120	0.078	0.081	0.283	-1.314; 2.277	0.424	-2.009; 3.287
Adalimumab + DMARD	0.031	0.037	0.074	0.114	0.761	-0.206; 1.885	1.036	-0.523; 3.099
Infliximab + DMARD	0.023	0.035	0.615	0.540	1.711	-0.117; 5.172	2.025	-0.513; 5.848

†Median of posterior mean distribution; *baseline for LOR is placebo for DMARD and anti-TNFs only, DMARD for anti-TNFs plus DMARD; CrI: credible interval; LOR: log odds ratio; FE: fixed effect; RE; random effects.

while infliximab has a higher risk (although not significantly greater than that for etanercept).

For the FE model the mean sum of deviance was 28.38, and for the RE model 25.49. The mean standard deviation for the RE model was 0.934.

9.7.6 Model 3b

At this level, the aim of the MTC models is to discern if there is any difference in malignancy risk between the anti-TNFs with and without additional DMARD treatment.

The results for this MTC model are set out in Table 9.6

For the FE model the sum of the deviance was 32.47, while for the RE model the sum of the deviance was 28.74, indicating that the RE model is a better fit, with a sum of deviance closer to the number of datapoints (29). The standard deviation for the RE model was 1.003.

Using the RE model, it is useful to compare the anti-TNF treatments against each other. Taking etanercept alone as the baseline, the LOR for adalimumab was -0.386 (95% CrI -3.140; 2.079). Taking etanercept plus DMARD as the

Table 9.7: Model 4a: Results for fixed and random effect(s) models.

	P(best)		P(worst)		FE		RE	
Treatment	FE	RE	FE	RE	Median† LOR*	95% CrI	Median LOR	95% CrI
Control	0.193	0.182	0.000	0.000	NA	NA	NA	NA
Etanercept (Rec)	0.009	0.022	0.034	0.044	0.7538	-0.102; 1.727	0.885	-0.467; 2.536
Etanercept (Low)	0.042	0.060	0.042	0.065	0.668	-0.514; 1.879	0.824	-0.972; 2.966
Adalimumab (Rec)	0.013	0.022	0.020	0.029	0.652	-0.229; 1.689	0.812	-0.519; 2.561
Adalimumab (Low)	0.289	0.270	0.054	0.055	0.203	-3.248; 2.594	0.290	-3.366; 3.431
Adalimumab (High)	0.054	0.047	0.194	0.223	1.103	-1.161; 3.169	1.413	-1.292; 4.452
Infliximab (Rec)	0.401	0.395	0.005	0.011	-0.092	-3.748; 3.687	-0.127	-4.121; 3.865
Infliximab (High)	0.000	0.002	0.652	0.572	2.057	0.229; 5.488	2.303	-0.053; 5.974

†Median of posterior mean distribution; *baseline for LOR is Control; CrI: credible interval; LOR: log odds ratio; FE: fixed effect; RE: random effects; Rec: recommended.

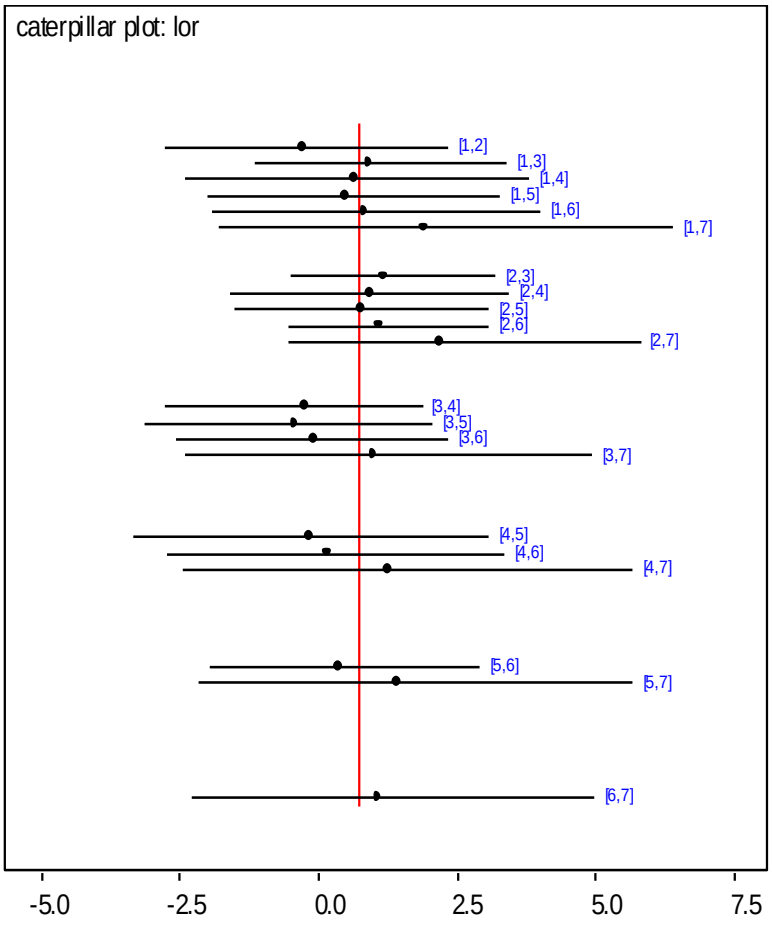
baseline, the LOR for adalimumab plus DMARD was 0.132 (95% CrI -2.700; 3.372), and for infliximab plus DMARD was 1.140 (95% CrI -2.443; 5.722).

Another interesting comparison is to compare etanercept and adalimumab alone against the same anti-TNF plus DMARD. Such a comparison is not possible for infliximab due to the lack of data for infliximab alone. Taking etanercept alone as the baseline, the LOR for etanercept plus DMARD was -0.185 (95% CrI -2.767; 1.886). Taking adalimumab alone as the baseline, the LOR for adalimumab plus DMARD was 0.317 (95% CrI -1.959; 2.922). From the caterpillar plot of LORs (Figure 9.1) there were no treatments for which the 95% CrI of the LOR did not include 0, or the overall mean log OR.

9.7.7 Model 4a

The results for Model 4a are set out in Table 9.7. One of the most important things to focus on in this MTC model is the comparison between doses of the same anti-TNF. The sum of deviance for the FE model was 38.14. For the RE model the total sum of the sum of deviances across individual studies was 34.28, indicating a better fit for the RE model. The standard deviation for the RE model was 0.8998.

Figure 9.1: Caterpillar plot of log odds ratios for Model 3b, random effects. (Numbers in square brackets refer to treatments.)



LOR; log odds ratio; Treatment 1: placebo; Treatment 2: Disease-modifying antirheumatic drug (DMARD); Treatment 3: etanercept; Treatment 4: etanercept + DMARD; Treatment 5: adalimumab; Treatment 6: adalimumab + DMARD; Treatment 7: infliximab.

Table 9.8: Model 4b: Results for fixed and random effect(s) models.

	P(best)		P(worst)		FE		RE	
Treatment	FE	RE	FE	RE	Median† LOR*	95% CrI	Median LOR	95% CrI
Placebo	0.109	0.130	0.019	0.018	NA	NA	NA	NA
DMARD	0.177	0.171	0.000	0.000	-0.464	-2.024; 1.317	-0.516	-2.914; 1.951
Etanercept (Rec)	0.004	0.012	0.036	0.046	0.879	-0.076; 1.978	1.064	-0.511; 3.080
Etanercept (Low)	0.025	0.039	0.043	0.068	0.789	-0.458; 2.092	1.009	-0.975; 3.489
Adalimumab (Rec)	0.009	0.014	0.016	0.025	0.714	-0.194; 1.779	0.951	-0.482; 2.849
Adalimumab (Low)	0.239	0.209	0.061	0.069	0.385	-3.118; 2.833	0.587	-3.435; 4.010
Adalimumab (High)	0.040	0.030	0.220	0.258	1.296	-1.037; 3.418	1.696	-1.163; 5.080
Infliximab (Rec)	0.398	0.392	0.004	0.011	-0.101	-3.744; 3.714	-0.128	-4.196; 3.955
Infliximab (High)	0.000	0.002	0.602	0.505	2.048	0.235; 5.601	2.320	-0.127; 6.095

†Median of posterior mean distribution; *baseline for LOR is placebo for DMARD, DMARD for anti-TNFs; CrI: credible interval; LOR: log odds ratio; FE: fixed effect; RE: random effects; Rec: recommended.

Using the RE model, comparing the low dose of etanercept with the recommended dose, the LOR was -0.059 (95% CrI -1.922; 1.829). Using the recommended dose of adalimumab as the baseline, the LOR for the low dose was -0.559 (95% CrI -4.232; 2.32), whilst for the high dose the LOR was 0.572 (95% CrI -2.059; 3.287). Finally, for infliximab, the LOR for the high dose compared to the recommended dose was 2.416 (95% CrI 0.034; 6.152).

9.7.8 Model 4b

The results for Model 4b are shown in Table 9.8. The sum of the deviance for FE model was 38.83 (with 35 datapoints), while the total sum of the sum of deviances for the RE model was 34.56. Again the RE model was superior in fit. The standard deviation for the RE model was 0.964. Choosing the RE model, the LOR for etanercept at low dose compared to recommended dose was -0.051 (95% CrI -1.975; 1.953). For adalimumab, the LOR for low compared to recommended dose was -0.377 (95% CrI -4.339; 2.622); for high compared to recommended dose the LOR was 0.726 (95% CrI -1.984; 3.656). The LOR for high dose infliximab compared to recommended dose was 2.443 (95% CrI

-0.025; 6.193). For all etanercept and adalimumab the median LOR for the low dose was less than 0, and for adalimumab and infliximab the LOR was greater than 0 for the high dose compared to recommended, thus indicating the possibility of a dose–response relationship.

9.7.9 Model 5a

For this model, it was impossible to derive results for ORs on the natural scale, possibly due to the extreme values generated by this model (hence all models have been reported in terms of LORs for consistency). The results are set out in Table 9.9.

The FE model using a control group amalgamated from placebo and DMARD treatment converged after a burn-in period of 10 000 iterations, and the standard 50 000 iterations was used to generate the posterior densities. There were also some concerns regarding convergence of the model, even after a longer burn-in period than previously used. For the RE model, convergence was achieved after 100 000 iterations, and again 50 000 iterations were used to generate the posterior densities.

The sum of deviance for the FE model was 40.29, compared to total sum of sum of deviances for the RE model of 35.96. With 38 datapoints in the dataset this implies the RE model is again the better fit. The RE model also had a mean standard deviation of 0.931. Choosing the RE model for additional reporting of results, comparing etanercept at recommended dose with additional DMARD to that without additional DMARD, the LOR was -0.114 (95% CrI -2.51; 1.863). Comparing etanercept at low dose to that at recommended dose, the LOR was -0.072 (95% CrI -1.969; 1.870).

Considering the adalimumab treatments, using the adalimumab recommended dose without additional DMARD as the baseline, the LORs are set out in Table 9.10.

Finally, for infliximab, this anti-TNF is only tested in conjunction with DMARD therapy, but taking the recommended dose as the baseline, the LOR for high dose was 2.426 with a 95% CrI of -0.007: 6.156.

Table 9.9: Model 5a: Results for fixed and random effect(s) models.

	P(best)		P(worst)		FE		RE	
Treatment	FE	RE	FE	RE	Median† LOR*	95% CrI	Median LOR	95% CrI
Control	0.034	0.021	0.000	0.000	NA	NA	NA	NA
Etanercept (Rec)	0.002	0.002	0.001	0.001	0.737	-0.162; 1.725	0.894	-0.449; 2.670
Etanercept (Low)	0.009	0.008	0.001	0.002	0.652	-0.535; 1.87	0.812	-0.990; 3.115
Etanercept (Rec) + D	0.010	0.015	0.001	0.002	0.756	-0.618; 2.131	0.774	-1.562; 3.152
Adalimumab (Rec)	0.010	0.010	0.000	0.000	0.524	-0.799; 1.881	0.569	-1.415; 2.671
Adalimumab (Low)	0.081	0.052	0.002	0.002	0.202	-3.310; 2.730	0.219	-3.804; 3.575
Adalimumab (High)	0.079	0.047	0.002	0.002	0.228	-3.283; 2.744	0.268	-3.835; 3.587
Adalimumab (Rec) + D	0.004	0.003	0.000	0.001	0.681	-0.282; 1.767	0.893	-0.659; 2.885
Adalimumab (Low) + D	0.649	0.770	0.012	0.028	-10.45	-65.56; 32.83	-42.08	-166.5; 51.43
Adalimumab (High) + D	0.000	0.000	0.957	0.976	24.20	1.942; 70.09	51.58	3.274; 193.6
Infliximab (Rec) + D	0.122	0.071	0.000	0.000	-0.093	-3.789; 3.558	-0.097	-4.092; 3.904
Infliximab (High) + D	0.000	0.000	0.026	0.013	2.056	0.239; 5.383	2.299	-0.1059; 6.062

†Median of posterior mean distribution; *baseline for LOR is Control; CrI: credible interval; D: DMARD; FE: fixed effect; LOR: log odds ratio; RE: random effects; Rec: recommended.

Table 9.10: Model 5a: Results for adalimumab using recommended dose without DMARD as baseline, RE model.

Treatment	Median† LOR	95% CrI
Adalimumab (Low)	-0.336	-4.248; 2.775
Adalimumab (High)	-0.296	-4.345; 2.775
Adalimumab (Rec) + DMARD	0.325	-1.806; 2.757
Adalimumab (Low) + DMARD	-42.6	-167.1; 51.02
Adalimumab (High) + DMARD	51.03	2.612; 193.3

†Median of posterior mean distribution; CrI: credible interval; LOR: log odds ratio; Rec: recommended.

Table 9.11: Model 5b: Results for fixed and random effect(s) models for probabilities of being 'best' and 'worst'.

	P(best)		P(worst)	
Treatment	FE	RE	FE	RE
Placebo	0.029	0.025	0.000	0.000
DMARD	0.041	0.018	0.000	0.000
Etanercept (Rec)	0.001	0.001	0.000	0.000
Etanercept (Low)	0.007	0.007	0.001	0.001
Etanercept (Rec) + D	0.008	0.011	0.002	0.001
Adalimumab (Rec)	0.007	0.007	0.000	0.000
Adalimumab (Low)	0.068	0.041	0.003	0.002
Adalimumab (High)	0.068	0.040	0.003	0.001
Adalimumab (Rec) + D	0.003	0.003	0.000	0.000
Adalimumab (Low) + D	0.641	0.418	0.023	0.028
Adalimumab (High) + D	0.000	0.000	0.957	0.986
Infliximab (Rec) + D	0.127	0.071	0.000	0.000
Infliximab (High) + D	0.000	0.000	0.026	0.005

CrI: credible interval; D: DMARD; FE: fixed effect; RE: random effects; Rec: recommended.

Table 9.12: Model 5b: Results for fixed and random effect(s) models for log odds ratios for anti-TNFs only compared to placebo.

	FE		RE	
Treatment	Median† LOR	95% CrI	Median LOR	95% CrI
Etanercept (Rec)	0.477	-1.011; 2.241	0.761	-1.276; 3.456
Etanercept (Low)	0.385	-1.305; 2.315	0.709	-1.734; 3.861
Adalimumab (Rec)	0.235	-1.522; 2.240	0.355	-2.241; 3.325
Adalimumab (Low)	0.031	-3.509; 2.792	0.071	-4.018; 3.832
Adalimumab (High)	0.056	-3.501; 2.801	0.141	-3.952; 3.884

†Median of posterior mean distribution; CrI: credible interval; D: DMARD; FE: fixed effect; LOR: log odds ratio; RE: random effects; Rec: recommended.

9.7.10 Model 5b

The FE model for this network converged in 10 000 iterations and a sample of 50 000 iterations was used to generate the posterior densities. For the RE model, 50 000 iterations were required prior to convergence, followed by a sample of 50 000 iterations to acquire the posterior densities.

Table 9.11 shows the values for the probabilities of being 'best' and 'worst' for all 13 treatments included in this MTC network.

Considering the LORs, it is useful to present these with the anti-TNFs alone being compared against placebo only, and the anti-TNFs used in conjunction with DMARDs being presented against DMARD therapy. Results are presented in Tables 9.12 and 9.13.

Table 9.13: Model 5b: Results for fixed and random effect(s) models for log odds ratios for anti-TNF plus DMARD compared to DMARD only.

	FE		RE	
Treatment	Median† LOR	95% CrI	Median LOR	95% CrI
Etanercept (Rec) + D	0.851	-0.579; 2.302	0.845	-1.708; 3.470
Adalimumab (Rec) + D	0.720	-0.270; 1.82	0.924	-0.715; 3.086
Adalimumab (Low) + D	-9.836	-65.40; 35.04	-20.46	-262.0; 89.66
Adalimumab (High) + D	24.74	1.992; 75.33	67.08	5.898; 208.1
Infliximab (Rec) + D	-0.100	-3.753; 3.618	-0.101	-4.275; 3.912
Infliximab (High) + D	2.046	0.227; 5.449	2.348	-0.168; 6.179

†Median of posterior mean distribution; CrI: credible interval; D: DMARD; FE: fixed effect; LOR: log odds ratio; RE: random effects; Rec: recommended.

The standard deviation for the RE model was 1.041, whilst the sum of deviance was 36.16 with 38 datapoints for the FE model; the same value also occurred for the RE model.

9.7.11 Model comparison using DIC and sum of deviance

Table 9.14 sets out the deviance information criterion (DIC) and pD value (the effective number of parameters in the model), as well as the total sum of deviance for each model and total number of datapoints in the model, for comparison.

9.7.12 Additional analyses

It may be possible that a signal regarding the influence of dose on malignancy would be more easily discernible with all three anti-TNFs combined at each of the three dose levels, low, recommended and high. This model was used, with correlation for multi-arm trials and vague priors as previously, with a random effects model. The results are shown in Table 9.15.

The standard deviation for this model was 0.622, with a total sum of deviances of 32.44 with 35 datapoints.

As it had been noted that there were issues with convergence regarding the most deconstructed model, without any combination of treatments, it was determined that this model was worthy of further investigation. One particular trial (Weinblatt *et al.* 2003; cited by Bongartz *et al.* 2006) had only one event

Table 9.14: DIC and sum of deviance results for all standard models.

Model	FE/RE	Total DIC	pD	Sum of deviance	No. data-points
1a	FE	97.151	13.371	25.88	26
1a	RE	98.135	16.093	24.14	26
1b	FE	98.671	14.218	26.55	26
1b	RE	99.508	17.027	24.58	26
2a	FE	100.643	15.138	27.6	26
2a	RE	100.712	17.748	25.06	26
2b	FE	102.270	15.985	28.38	26
2b	RE	101.910	18.518	25.49	26
3b	FE	113.245	17.889	32.47	29
3b	RE	112.275	20.643	28.74	29
4a	FE	122.575	18.580	38.14	35
4a	RE	121.974	21.837	34.28	35
4b	FE	124.080	19.389	38.83	35
4b	RE	122.973	22.559	34.56	35
5a	FE	131.707	20.576	40.29	38
5a	RE	130.716	23.907	35.96	38
5b	FE	133.470	21.462	36.16	38
5b	RE	131.754	24.752	36.16	38

DIC: deviance information criterion; FE: fixed effect; pD: effective number of parameters in model; RE: random effects.

Table 9.15: Dose only. Results for random effects models for log odds ratios for dose levels of all anti-TNFs combined, compared to non-anti-TNF controls.

Treatment	P(best)	P(worst)	Median† LOR	95% CrI
Control	0.837	0.000	NA	NA
Recommended	0.022	0.018	0.732	-0.045; 1.684
Low	0.140	0.053	0.710	-0.637; 2.138
High	0.001	0.929	1.973	0.684; 3.705

†Median of posterior mean distribution; CrI: credible interval; OR: odds ratio; RE: random effects.

Table 9.16: Weinblatt (2003) removed from dataset. Results for Model 5b using random effects.

Treatment	P(best)	P(worst)	Median† LOR*	95% CrI
Placebo	0.102	0.015	NA	NA
DMARD	0.091	0.000	-0.303	-2.767; 2.471
Etanercept (Rec)	0.005	0.039	0.771	-1.294; 3.446
Etanercept (Low)	0.026	0.072	0.706	-1.800; 3.891
Etanercept (Rec) + D	0.052	0.073	0.842	-1.678; 3.440
Adalimumab (Rec)	0.032	0.032	0.387	-2.202; 3.361
Adalimumab (Low)	0.177	0.088	0.117	-4.027; 3.833
Adalimumab (High)	0.182	0.096	0.147	-3.933; 3.900
Adalimumab (Rec) + D	0.017	0.053	0.931	-0.729; 3.065
Infliximab (Rec) + D	0.315	0.011	-0.117	-4.258; 4.022
Infliximab (High) + D	0.001	0.520	2.335	-0.212; 6.335

†Median of posterior mean distribution; *baseline for LOR is placebo for DMARD and anti-TNF only, DMARD for anti-TNF plus DMARD; CrI: credible interval; D: DMARD; LOR: log odds ratio.

across four arms at this level of deconstruction of the treatment arms. It was decided to remove this one trial, with the result that two treatments (low dose adalimumab plus DMARD and high dose adalimumab plus DMARD) were removed from the treatment network as they did not appear in any other trial. This also eliminated two direct comparisons between two treatments that had zero events across the comparison. On fitting this model using random effects with a reduced dataset, convergence occurred after 10 000 iterations, which was a shorter burn-in period than that used for the same model but with all studies, where issues with convergence necessitated a burn-in of 50 000 iterations. The results are set out in Table 9.16.

This model had a standard deviation of 1.025 (compared to 1.041 for the equivalent model including all studies) and a total sum of deviances of 35.25 (with 34 datapoints) compared to 36.16 with 38 datapoints when all studies were included in Model 5b.

9.8 Further discussion

9.8.1 Initial inspection of dataset

On first appearances, the data support the conclusion that malignancies occur more commonly in the anti-TNF patients compared to non-anti-TNF controls

(based on data in Section 9.6.3). However, the possibility of differential follow-up, with anti-TNF groups receiving longer follow-up than control groups, may be distorting these results, with the anti-TNF groups having more opportunity to develop a malignancy over the longer time period. This problem highlights the need for individual patient data (IPD) for outcomes that are very associated with long time periods for development. The following MTC analyses do not address this problem, but nevertheless may identify broad areas of concern warranting further analysis.

9.8.2 Baseline mixed treatment comparison models

Selected additional results are introduced in this section, for the purpose of comparison with results set out in Section 9.7.

Nine MTC network models have been fitted to investigate the issue surrounding malignancy and anti-TNF treatment for RA. Considering the most basic model, with all anti-TNFs and all non-anti-TNFs considered together, (Models 1a and 1b; see Sections 9.7.2 & 9.7.3), it was evident that the highest risk of malignancy occurred in the anti-TNF group, with a probability of being the 'worst' treatment of 0.992 (RE Model 1a). The OR for malignancy of 2.480 (based on the median LOR of 0.908) for anti-TNFs compared to controls supports this conclusion. In Model 1b, the control arms were divided into two treatments, placebo only and DMARDs. The purpose of this model is to shed light on whether a DMARD may be capable of offering protection against malignancy in those with RA, bearing in mind that RA is associated with a higher level of malignancy. In the RE version of this model, the placebo had a probability of being 'worst' of 0.250, while the active DMARD group had a probability of 0.007.

For the probability of being 'best', the values were 0.686 for the DMARD group and 0.312 for the placebo group. This indicates that the DMARD group was at lower risk of malignancy and hence indicates that this should be taken into account in the interpretation of more complex models. The anti-TNF group had both the lowest probability of being 'best' and the highest probability of being 'worst' (0.002 and 0.743 respectively). Also, the median OR for the anti-TNF group compared to the DMARD group was significantly higher. Hence, even in this relatively simple model, it appears that those on anti-TNF treatments are at higher risk of malignancy than those not. However, at this stage there is no

possibility of discerning whether any particular anti-TNF has a higher or lower risk than any others, or if the risk is roughly equal across all three anti-TNFs.

One point of interest worth mentioning here is that for the sum of deviance (and total sum of deviances for multi-arm models), the FE model consistently produces a higher value compared to the RE model, indicating that the RE model is a better fit. The (total) sum of deviance(s) is close to the total number of datapoints for all models, indicating a reasonably good fit for all models considered. The sum of deviance can influence model selection, in that it provides an indication of how well the model fits the available data (discussed further below).

The DIC (Table 9.14) also provides a means to compare models without indicating absolute goodness-of-fit. For the most basic Models, 1a and 1b, the FE model appears to be the better fit. For Models 2a and 2b, there is little to choose between the FE and RE models for Model 2a, but it appears that for Model 2b, the RE model is an improved fit. For Models 3b, 4a, 4b, 5a and 5b, this pattern, with the RE model being the better fit is perpetuated.

In contrast to the DIC, however, the pD values appear to consistently favour the FE models, by comparison to their RE counterparts. Overall then, based on the three methods of model assessment, there is no clear evidence to strongly support either an FE or an RE model in these analyses.

From a clinical perspective, all studies are in similar patients (all USA-based) but there are differences among them, for example, different treatment regimes, and possibly varying severities of rheumatoid arthritis. Therefore, for the sake of conservativeness and to avoid over-generalising between studies, especially in view of the sparsity of events data, the RE model may be preferable, on the basis of the DIC.

This contrasts with the work of Sweeting *et al.* (2004) and Bradburn *et al.* (2007). These authors, using traditional frequentist methods, found that the RE methods were not of value for analyses with low baseline rates, in terms of producing low estimates for heterogeneity, which therefore impacts on the results of the RE model, yielding results similar to an FE model. Similar results were found in the multiple analyses of the GSK dataset in Chapter 7, inasmuch as the use of an FE model and an RE model often yielded effectively identical results.

It is interesting to contrast how Bayesian and frequentist models approach between-studies heterogeneity, and it is perfectly reasonable that the frequentist approach favours FE methods for sparse data, possibly because at low event rates, heterogeneity on an absolute scale is low, even if significant on a relative scale, and is less easy to detect using an RE model. With the Bayesian models however, the goodness-of-fit often appears to improve using the RE models (even if the results produced by the two methods are not greatly different).

It is also relevant to note that the DIC increases in value for each increasingly complex network, although pD also increases. This implies that the more complex models are a poorer fit relative to the simpler models with fewer treatments. Such a phenomenon could be due to sparsity of both trials and events. Performance of a similar MTC, using a larger number of trials and events, would be able to determine whether or not sparsity of data is a major factor in the relative goodness of fit of the different models.

Model 1b promotes the conclusion that the DMARD controls have a lower risk of malignancy compared to placebo controls. There is a clinically plausible explanation for this, in that patients with RA are at greater risk of malignancy, due to the inflammation resulting from their condition, and using a DMARD helps to reduce the inflammation, and hence the risk of malignancy.

Moving on to the network with the next level of complexity, Models 2a and 2b, there is now the possibility of discovering whether any of the anti-TNFs has a higher or lower risk of malignancy compared to the others (Tables 9.4 & 9.5). In Model 2a, the control arm is both placebo and DMARD combined. The control arm is clearly the safest with regard to malignancy, having the lowest probability of being 'worst' and the highest probability of being 'best'. At this point there is a divergence in the evidence between the concepts of 'best' and 'worst' with regard to malignancy risk. The probability of being 'best' for each of the three anti-TNFs is roughly the same for all three anti-TNFs (0.063 for etanercept, 0.076 for adalimumab and 0.052 for infliximab in the RE model).

Based on these data alone, there is little evidence to prefer one anti-TNF over another for malignancy safety reasons. For the probability of being 'worst' however, there is more variation between the three drugs. The probability of 'worst' for infliximab is 0.679 (in the RE model), much higher than etanercept and adalimumab. It is also worth noting that infliximab in all trials is administered

along with a DMARD (methotrexate) which may be exerting a protective effect, while the other two drugs are administered sometimes with a DMARD and sometimes alone. The median LOR for malignancy in infliximab compared to the control arms is 1.928, but with a very wide CrI, so it is impossible to conclude that there is any significant increased risk of malignancy.

Moving on to Model 2b, where the control arms are divided into DMARD and placebo, the results are very similar, with the three anti-TNFs showing very little difference in the probability of being 'best' but again infliximab is the anti-TNF with the highest probability of being 'worst'.

So at this stage, there is reasonable evidence to indicate that the anti-TNFs increase risk of malignancy compared to placebo or DMARD only, but it may be the case that the majority of this increased risk is due to infliximab, whereas there may be less increased risk with etanercept and adalimumab.

Model 3b (see Table 9.6) takes into account whether the anti-TNF is administered alone or with a DMARD. Again, the DMARD alone appears to be associated with a lower risk of malignancy than the placebo alone. One point of interest is to determine whether there is a clear pattern in malignancy risk for the anti-TNFs administered both alone and with an anti-TNF. For etanercept, the picture is not clear-cut. For both FE and RE models the probability of being 'best' is greater for the arm with DMARD than for etanercept alone. However, in contradiction to this result, etanercept plus DMARD also has a higher probability of being 'worst'. Considering the LORs, etanercept plus DMARD has a lower median LOR for malignancy using placebo as the baseline compared to etanercept only (0.633 compared to 0.839 in the RE model).

For adalimumab, the situation is different. Adalimumab alone has a higher probability of being 'best' compared to adalimumab plus DMARD, in both FE and RE models. For the probability of being 'worst' there is little to choose between the two treatments for the FE model, whereas the RE model shows that adalimumab plus DMARD has a slightly higher probability than adalimumab alone. When considering the LORs, adalimumab plus DMARD has a higher LOR compared to placebo than adalimumab alone (0.767 compared to 0.424). Hence, there is a conflict of results when trying to determine any difference in malignancy between anti-TNF alone and with DMARD.

Infliximab is administered only in the presence of DMARD in these trials, and it is immediately striking that infliximab plus DMARD has the highest probability of being 'worst', at 0.540 for the RE model. Also, the LOR for malignancy (compared to placebo) is 1.794, by far the highest LOR of any of the anti-TNF treatment options. It appears that this more detailed model is still supporting the view that infliximab is the anti-TNF with the highest risk of malignancy whereas there is no clear evidence to indicate that the presence of a DMARD with the anti-TNF either reduces or increases any increased risk with an anti-TNF.

At this point it is useful to consider one of the additional models, which investigated a possible dose–response relationship across all three anti-TNFs combined. The RE model applied to dose alone showed that the high dose was associated with the highest probability of being 'worst' and lowest probability of being 'best'. These results also showed little substantive difference between the low and recommended doses, with them both having similar probabilities of being the 'worst' anti-TNF, while the low dose had the highest probability of being 'best' out of the three anti-TNFs. Consistently with the other models, the non-anti-TNF control was associated with the highest probability of being the 'best' treatment. These results were borne out by the LORs, with the high dose anti-TNF having the highest LOR and a 95% CrI with a lower bound greater than 0. Amalgamating the three anti-TNFs by dose has supported the argument that dose is the defining factor in relating anti-TNF use to malignancy, rather than the specific anti-TNF.

The next models (Models 4a and 4b, FE and RE models, with results shown in Tables 9.7 & 9.8) to be considered included dose level for each anti-TNF individually, but did not take into account the presence of additional DMARDs. These models are useful in determining whether a dose–response relationship alluded to by the dose-only model would be seen in any of the individual anti-TNFs, and if so, which. Model 4a (using placebo/DMARD as control, RE model) shows that etanercept has similar median LORs for both recommended (0.885) and low (0.824) doses. Adalimumab however, does indicate a dose–response relationship from low to recommended and then to high dose, with median LORs being 0.290, 0.812 and 1.413 respectively. Infliximab shows the strongest element of dose–response, with median LOR for the recommended dose being -0.127, and for the high dose 2.303, this result almost reaching

statistical significance with a lower bound for the 95% CrI of -0.053. Whilst in all cases the 95% CrIs are very wide and overlap one another, the dose trend seems clear with both adalimumab and infliximab.

The most complex models (Models 5a and 5b, FE and RE models) are also the most difficult to interpret in a useful way, because once both additional DMARD and dose are included, these two dimensions of treatment are difficult to disentangle, and complexity of the model is inextricably linked with increasing sparsity of data, the events available being spread over the largest possible number of treatments.

Using Model 5a (with placebo/DMARD as control, RE model), a comparison can be made of DMARD status across different dose levels for individual anti-TNFs, as well as different dose levels for each anti-TNF combined with DMARD status. Considering first etanercept at recommended dose, the median LOR for etancercept alone was 0.894 (95% CrI -0.449; 2.670), whilst for etanercept plus DMARD, the median LOR was 0.774 (95% CrI -1.562; 3.152). Results for Model 5a are shown in Table 9.9.

For adalimumab at recommended dose, the median LOR was 0.569 (95% CrI -1.415; 2.671), while for adalimumab at the same dose plus DMARD, the equivalent value was 0.893 (95% CrI -0.659; 2.885). For adalimumab alone at low dose the median LOR was 0.219 (95% CrI -3.804; 3.575), whilst with DMARD the median LOR was -42.08 (95% CrI -166.5; 51.43). At high dose adalimumab, the median LOR was 0.268 (95% CrI -3.835; 3.587), whilst for high dose with DMARD the median LOR was 51.58 (3.274; 193.6).

The extreme values for the adalimumab at high and low doses plus DMARD, as well as very wide CrIs, make interpretation difficult and reduce any validity of conclusions. The median LOR values could be lacking in validity as a point estimate for the treatment effect, due to the sparsity of data to support investigation of these treatments. Only one event occurred in the treatment group receiving adalimumab at high dose in conjunction with DMARD, whilst there were no events at all in the treatment group receiving low dose adalimumab plus DMARD; furthermore, both of these treatments occurred in only one primary study. However, the width of the CrIs indicates the lack of certainty around the estimated LOR, and therefore reduces any confidence placed in the point estimate.

Technically, this may be due to the fact that the Gibbs sampler, although convergence has occurred, cannot narrow down the area of the profile very precisely due to the lack of data, hence the risk of an extreme point estimate and wide CrI. This highlights the issue that sparse data may produce artefactual results, even if those results appear plausible. A comparison of the results is made in Section 9.8.4, comparing results of analyses including all studies and the results from equivalent models excluding the only study that investigated adalimumab at high and low dose with DMARD (Weinblatt *et al*. 2003, as cited by Bongartz *et al*. 2006).

Returning to the clinical aspect of the influence of DMARD at different anti-TNF doses, there is no clear picture across etanercept and adalimumab that additional DMARD increases or reduces risk of malignancy, even at recommended dose levels where the amount of data is greatest.

When making a distinction between treatment groups receiving DMARD and those not, the dose–response relationship among etanercept patients becomes less clear, as the median LOR for etanercept patients on low dose (without DMARD) is less than those on recommended dose without DMARD, but greater than those on recommended dose with DMARD.

For adalimumab patients without DMARD, the results are again anomalous with regard to a dose–response relationship, with the recommended dose having the highest risk of malignancy. When adalimumab with DMARD is considered, there is a dose–response relationship with increased risk of malignancy at increasing doses of adalimumab, although the caveats regarding sparsity of data (both studies and events) mentioned above apply in this case.

Infliximab was only used in these studies in conjunction with DMARD, but in this model also, the dose–response relationship is clearly seen based on the median LORs for recommended and high doses. Unfortunately, there is no low dose for infliximab that may help to confirm or refute a dose–response relationship.

The results and potential conclusions derived from different MTC models with varying ways of defining treatments can be very different; this indicates the importance of using multiple models and making a consideration of the evidence from all models before arriving at any clinical conclusions.

9.8.3 Discrepancies between probabilities for 'best' and 'worst'

There were often discrepancies in the results for a particular model, whereby the treatment with the highest probability of being the 'best' treatment, with regard to having the lowest risk of malignancy, is not always the treatment with the lowest probability of being the 'worst' treatment, in terms of having the highest risk of malignancy.

These probabilities were based on the rankings of the treatments relative to the baseline at each iteration. According to the relative densities of the mean treatment effects, the treatment with the greatest probability of being 'best' may not also be the treatment with the greatest probability of being 'worst' where there are multiple treatments. Only the 'best' and 'worst' treatment in the rankings are being considered (and not the ranking positions that lie between best and worst).

9.8.4 Sensitivity analysis: removal of the primary study by Weinblatt *et al*. 2003

The sparsity of events in some of the treatments defined in the most complex model described above was a cause for concern. It was decided to perform a sensitivity analysis to determine any potential influence of the inclusion of one primary study that introduced treatments with zero events into the MTC network. This primary study was Weinblatt *et al*. (2003), cited by Bongartz *et al*. (2006). As shown in Table 9.3, this study contributed only one event, but could be broken down into four separate treatments when including dose and additional DMARD as part of the treatment designation. (All results comparing the inclusion and exclusion of Weinblatt *et al*. (2003) are based on the RE model for Model 5b.)

On removing this study, two treatments were excluded completely; adalimumab at low and high doses plus DMARD, which appeared exclusively in this study. This reduced the number of treatments in the MTC model from 13 to 11 with the loss of just one event. As these treatments were associated with extreme values for the LORs (median values of -42.08 and 51.58 respectively in Model 5a, Table 9.9, and -20.46 and 67.08 respectively in Model 5b, Table 9.13, where the

comparator is DMARD only), their exclusion does not result in any significant loss of information, as these values are very difficult to interpret.

The removal of the Weinblatt *et al*. (2003) study did have consequential effects on the probabilities of individual treatments being the 'best' and 'worst' (shown in Table 9.16). As an example, in the original Model 5b with all 13 studies included, the treatment with the highest probability of being 'worst' was adalimumab at high dose plus DMARD, with a probability of 0.986 (Table 9.11). Compared to this treatment all the probabilities of being 'worst' for other treatments were (inevitably) far lower. It is to be noted however, that this treatment occurred only in the Weinblatt *et al*. (2003) study, and had only one event in this treatment group.

This extreme sparsity of data therefore heavily influenced the probabilities used as a means of directly comparing risk of malignancy across multiple treatments. A similar situation occurred with adalimumab at low dose plus DMARD which was associated with a probability of being 'best' of 0.418 in the model including all studies, despite having appeared in only one primary study, and having zero events in this study. In the absence of the Weinblatt *et al*. (2003) study the treatment with the highest probability of being 'best' transferred to infliximab at recommended dose plus DMARD (results for Model 5b using RE, with the exclusion of Weinblatt *et al*. (2003) are shown in Table 9.16). This example illustrates the sensitivity of the probability of 'best' and 'worst' treatment to sparsity of both primary studies and events within treatment arms.

When comparing the median LORs between the models including and excluding Weinblatt *et al*. (2003), there were slight differences in the median LOR values. In five cases the median LOR for an individual treatment (including an anti-TNF) in the model excluding Weinblatt *et al*. (2003) was greater than in the model including this study. Despite wide CrIs in all cases, changes in LORs and ORs would give different impressions of malignancy risk for different treatments, highlighting the potential for changes in results when including or excluding any particular study, especially where events are sparse in that study.

It is also relevant to note that changes in probability for 'best' or 'worst ' status may change dramatically, despite the overall LOR not changing greatly, on exclusion of one primary study. For example, on exclusion of Weinblatt *et al*. (2003), the treatment with the highest probability of being 'best' was

infliximab at recommended dose plus DMARD (with a probability of 0.315), associated with a median LOR of -0.117. When all studies are included, the probability of this treatment being best falls dramatically to 0.071, whereas the median LOR of malignancy becomes only slightly less negative (-0.106). Another example of this phenomenon is illustrated by infliximab at high does plus DMARD, which has a probability of being 'worst' of 0.005 when all studies are included, which jumps to 0.520 when Weinblatt *et al*. (2003) is excluded. This corresponds with a change in median LOR of malignancy of 2.348 with all studies included, to 2.335 when Weinblatt *et al*. (2003) is excluded. Effectively, a lower probability of being 'worst' is associated with a greater median LOR for malignancy risk, which appears to be a contradictory result.

This re-analysis of the dataset to exclude one study shows that the probabilities of being the 'best' and/or 'worst' treatment are more susceptible to being altered than the LORs, possibly indicating that the LORs should be given greater weight when determining which treatment is safer (or more effective depending on the context).

When the data are very sparse (in terms of both events and studies) at certain nodes across the MTC network, the results become very sensitive to the included data, especially where the exclusion of one study can change the network, by removing one or more nodes. Hence, if the specific MTC network is of interest (in this case, including both dose and additional DMARD within the model), it is advisable to perform sensitivity analyses to ensure that any conclusions are robust to which particular studies are included. If conclusions are not robust, then caveats regarding the strength of conclusions can be considered.

9.8.5 Alternative parameterisations

Noting that the WinBUGS code being used for the models in these analyses differed in structure from the algebraic models in the original papers setting out the MTC models, it was decided to reparameterise the model for certain model examples, such that the WinBUGS code reflected the original algebra. In theory, the two models would produce the same results being algebraically identical. To illustrate, the original papers set out their models as:

logit $(p_i^T) = \mu_i + \delta_i/2$,

logit $(p_i^C) = \mu_i - \delta_i/2$,
whilst the WinBUGS code expresses the model slightly differently:

logit $(p_i^T) = \mu_i + \delta_i$,
logit $(p_i^C) = \mu_i$.

Although it is evident that the difference between the control and treatment groups for each study is δ_i in both models (where δ_i represents the log OR for the treatment group compared to the control group), and therefore should be estimated by the same posterior densities in both models, in practice this was not the case. There were in fact some discrepancies between the two parameterisations, which were thought to be resulting from correlation between the values of μ and δ in each study, hence altering the value of δ as μ is referring to different 'nuisance' parameters in each model. In the light of the sparse data, this was thought to yield differences in the value of δ, but this avenue of thought was not pursued further, due to acceptance of the WinBUGS code as an accurate representation of the desired model.

9.8.6 Discussion of previous research

Considering previous work, the study by Bongartz *et al*. (2006) compared all anti-TNFs against 'placebo' (in fact all non-anti-TNF arms including those with active DMARDs). These authors used a Mantel–Haenszel method with and without a continuity correction, as well as Bayesian models with fixed and random effect(s) and with inclusion and exclusion of studies of trials with zero events in total, as well as a conditional maximum likelihood model. In all of their models, there was a significant increase in risk of malignancy in those receiving the anti-TNF. Further analyses comparing low-dose against high-dose anti-TNFs demonstrated a higher risk of malignancy with the higher doses. Indeed, low-dose therapy did not appear to be significantly associated with increased risk of malignancy. The authors duly point out that the sparsity of events may influence the results, with low precision and wide confidence intervals. The issue of unequal follow-up times across studies was also not addressed. However, this study, although including etanercept, adalimumab and infliximab studies was unable to make any distinction between anti-TNFs.

The study by Bongartz *et al*. (2006) also included serious infections as well as malignancy. Picking up on the problems of underpowered trials, another

group of authors (Leombruno *et al*. 2008) attempted to further investigate the relationship between anti-TNFs and malignancy as well as other adverse events. These authors attempted to address the problem of unequal follow-up times by using two outcome metrics, firstly an OR, and secondly a rate ratio adjusted for follow-up time. They also restricted the analysis to fixed effect models only, as they argued that this method often produced narrower CIs with sparse data. The authors state that they used this approach to maximise the chances of finding a significantly increased risk of adverse events. However, the justification of using fixed or random effect(s) models should be based on the nature of the studies themselves (regarding between-studies heterogeneity) and whether the fixed effect model can be justified, and if not, then a random effects model can be used for conservativeness.

Although it is clearly highly desirable to detect a risk of adverse events where such a risk exists, it is equally undesirable to promote concerns about a spurious risk for adverse events, which may result in patients not receiving a potentially beneficial treatment. The authors did, however, test for heterogeneity using the Q statistic and the I^2 statistic, and used a random effects model if heterogeneity was present.

The chosen meta-analysis method was the Mantel–Haenszel, due to imbalanced group sizes (where the Peto method is not recommended) and because other methods, such as inverse variance, are not recommended for rare event data. The preferred continuity correction was that of Sweeting *et al*. (2004), based on using the reciprocal of the opposite treatment arm size. Sensitivity analyses for both continuity correction and meta-analysis method were also performed. A broader range of outcomes was also investigated, including death, serious adverse events, and cancer broken down according to type.

All comparisons were made against 'placebo' which appears to comprise all non-anti-TNF arms, including those with DMARD treatment. At recommended dose levels, the OR for non-cutaneous cancers and melanomas was 1.31 (95% CI 0.69; 2.48). For each individual anti-TNF the OR was greater than 1, with infliximab having a higher OR than either etanercept or adalimumab individually, although none were statistically significant. Considering the exposure-adjusted meta-analysis, the risk ratios for all three anti-TNFs individually and for all anti-TNFs combined were lower than the corresponding OR, with the risk ratio for all anti-TNFs being 1.21 (95% CrI 0.63; 2.32).

At higher doses, the OR for non-cutaneous cancers and melanomas was 2.91 for all anti-TNFs combined; the corresponding risk ratio was 3.04 (0.95; 9.68). Some results were contradictory; for example, for lymphomas only the higher dose ORs and risk ratios appeared to be lower than for the recommended dose, while for non-cutaneous cancers and melanomas, the ORs and risk ratios were consistently higher at higher doses, leading to difficulty in determining the presence of a dose–response relationship.

Overall, no statistically significant relationships were found between anti-TNF use and any form of malignancy (including the primary analyses and various sensitivity analyses). The authors point out the drawbacks of indirect comparisons and the lack of validity when compared to randomised controlled trials, and the fact that the overall sample was underpowered to detect any increase in such rare events.

It should also be pointed out that this model used no Bayesian analyses. Studies with zero events in total appear to have been included in the primary analyses (presumably by inappropriate use of continuity corrections) and excluded as part of the sensitivity analyses (with no alteration of results).

Considering the recent work by Bongartz *et al*. (2009), which is based on the analyses detailed in the previous chapter, it is now possible to compare the MTC models to the non-MTC meta-analyses previously performed. The major difference between this publication and the earlier work by Bongartz *et al*. (2006) and Leombruno *et al*. (2008) is that this study investigated only etanercept, as there are pharmacological differences between etanercept and the other anti-TNFs.

In the published reference, the primary method of analysis was a survival method based on the Cox Proportional Hazards model, using IPD for each trial. Hence, this study had an advantage over the previous meta-analysis in terms of having IPD and being able to use survival methods. Time-to-event methods enable the issue of unequal follow-up times to be appropriately addressed. Fixed and random effect(s) methods were also used, along with a sensitivity analysis using the Mantel–Haenszel model with the continuity correction developed by Sweeting *et al*. (2004).

Another advantage of this study was the fact that the authors had direct access to trial data, allowing the verification of all cases of malignancy. Using the

hazard ratio (HR), the fixed and random effect(s) models yielded similar non-significant results (an HR of 1.82, 95% CrI 0.78, 4.22 for the random effects model). The OR sensitivity analyses also yielded non-significant results but with an increased risk for malignancy in those using etanercept compared to non-anti-TNF groups.

9.9 Conclusions

Ultimately, the overall conclusion of the MTC is that the signal from the data indicates that there is a higher risk of malignancy in those using anti-TNFs compared to those who are not, whether they are receiving no medication for RA or whether they are receiving other DMARDs. A clear pattern has emerged from all studies in this field (Section 9.8.6).

However, CrIs are tending to be wide, largely due to the small number of events, hence it is very difficult to draw any firm statistical conclusions from these data. In all models where the anti-TNFs are separated out individually (Models 2a and 2b), there is a clear indication that infliximab is associated with a higher risk of malignancy compared to the other two anti-TNFs. In Model 3b, infliximab with DMARD again appears to be the anti-TNF with the highest risk of malignancy. There is some contradiction in whether an additional DMARD reduces or increases the risk of malignancy when administered with an anti-TNF. In all models however, the use of a DMARD is associated with lower risk of malignancy when used without anti-TNFs, compared to placebo alone.

From a methodological point of view, this MTC has illustrated the importance of using a variety of models with different levels of combination of treatment arms. Different forms of combination can lead to results with different implications for clinical practice. Possibly the most important aspect is to incorporate clinical insight into which combinations are the most valid clinically (in terms of pharmacological action, clinical effects and so forth). These combinations of treatment arms can then be evaluated in the light of the appropriateness of the model, in terms of the number of treatment arms and the distribution of potentially sparse events across these treatment arms.

Heterogeneity is a topic not extensively dealt with in this example. Clinical heterogeneity is not a prime issue, largely due to the similarity of the studies,

all conducted in the USA, and on comparable patients, with a randomised controlled trial design. Statistical heterogeneity is intrinsically bound with the issue of sparse data, and the difficulty of achieving sufficient power to detect underlying differences with few events. Although this issue has not been specifically addressed, it is worthy of further consideration.

For many clinical purposes, the MTC with the lowest degree of combination may be the most useful, as the largest number of different treatments can be compared against each other, thus clarifying where any significant (statistically or clinically) results are to be found. However, a desire to maximise the number of treatments may lead to models with too few events to allow the model to be fitted, and to provide CrIs that are very wide and preclude making any firm inferences from such results. The interplay between number of treatments and number of events in a model, and how to most appropriately analyse the available data is an area that can be developed further.

The underlying assumptions regarding an MTC, as referred to in Section 9.2.4, are difficult to assess statistically, and in the specific case of adverse events, may be difficult to assess non-statistically, in terms of the plausibility of all studies coming from the same underlying distribution. If data are sparse, widely varying treatment effects may be seen across the different studies, but this would not preclude a common underlying distribution; but heterogeneity would require addressing. However, the lack of such a distribution, which would render meta-analysis inadvisable, could not be ruled out. The differences in results on excluding the study by Weinblatt *et al.* (2003, as cited by Bongartz *et al.* 2006) may point to a failure of the dataset to fulfil the underlying assumptions required for MTC meta-analysis. However, where data are sparse, the overriding consideration may be to gain a signal from the dataset, giving lower priority to the technical assumptions of the analysis.

Across the four studies (Bongartz *et al.* (2006), Leombruno *et al.* (2008), Bongartz *et al.* (2009) and the MTC analysis described above), the issue of malignancy in association with anti-TNFs has been approached in a variety of ways, including Bayesian and non-Bayesian meta-analysis methods, IPD and aggregate data, and MTC analysis. There are obvious advantages for some analyses, for example, the IPD analysis allows follow-up time to be accounted for, while the MTC analysis allows different treatments to be analysed separately, but compared directly and indirectly.

All the analyses have sent a signal that there is a non-significant increased risk of malignancy with anti-TNFs, which the MTC analysis has tried to analyse further by investigating individual anti-TNFs, but all analyses have the drawback of rare events, with the associated wide confidence/credible intervals and difficulties in interpreting the results.

Although there may be clinical issues regarding whether a single higher dose of anti-TNF is the same in terms of malignancy risk as two smaller doses, the dataset does not lend itself to such specific analyses and it was thought more important to concentrate on the more fundamental queries as described in Section 9.3.

One of the most difficult questions to address is that of which MTC model is the most appropriate; the answer is of course dependent on the clinical situation. If querying whether an anti-TNF in general is associated with increased risk of malignancy, this question would be best suited by either Models 1a and 1b. If the question relates to which anti-TNF may be associated with the highest risk, Models 2a/2b, 3a/3b and 4a/4b would be the most valuable. It is impossible to prescribe an anti-TNF without specifying the dose, hence Models 4a/4b can be argued to be the most useful in terms of clinical decision-making. The complexity of Models 5a/5b, compounded by practical difficulties of model-fitting, makes the interpretation of the results of these models difficult. The model including dose only (Section 9.7.12) is also beneficial in considering the influence of dose.

Further analysis of this dataset is performed in Chapter 10, which includes hierarchical models within the MTC modelling framework, the application of constraints to such models, and a sensitivity analysis across multiple prior distributions. These models extend the analyses of this chapter, clinically by assisting in answering some of the queries regarding choice of anti-TNF and dose, and statistically, by adding complexity to the models, but in response to clinical requirements.

As an extension to this work, a harm–benefit analysis on anti-TNF therapy, evaluating the risk of malignancy against the improvement in quality of life (QoL) as a result of the beneficial effect on rheumatoid arthritis, would be of interest. A particular benefit of this approach would be the ability to quantify the net benefit or reduction in QoL due to anti-TNF therapy. This would

render issues regarding statistical significance less important, which, due to the inevitable low power, is of little practical assistance in decision-making in this context. A harm–benefit analysis (although in a different clinical field) is presented in Chapter 11. The use of an MTC meta-analysis of an adverse event could then be fed into a harm–benefit model, to provide a result in terms of net QoL, as evaluated against the benefits of the intervention.

9.10 Summary

This chapter extends the investigations into the relationship between anti-TNFs and risk of malignancy in patients with rheumatoid arthritis that was commenced in Chapter 8. Using the dataset of 13 studies, including data on three anti-TNFs (etanercept, adalimumab and infliximab), analyses were conducted using MTC methods and extensions thereof, using Bayesian modelling methods in WinBUGS software.

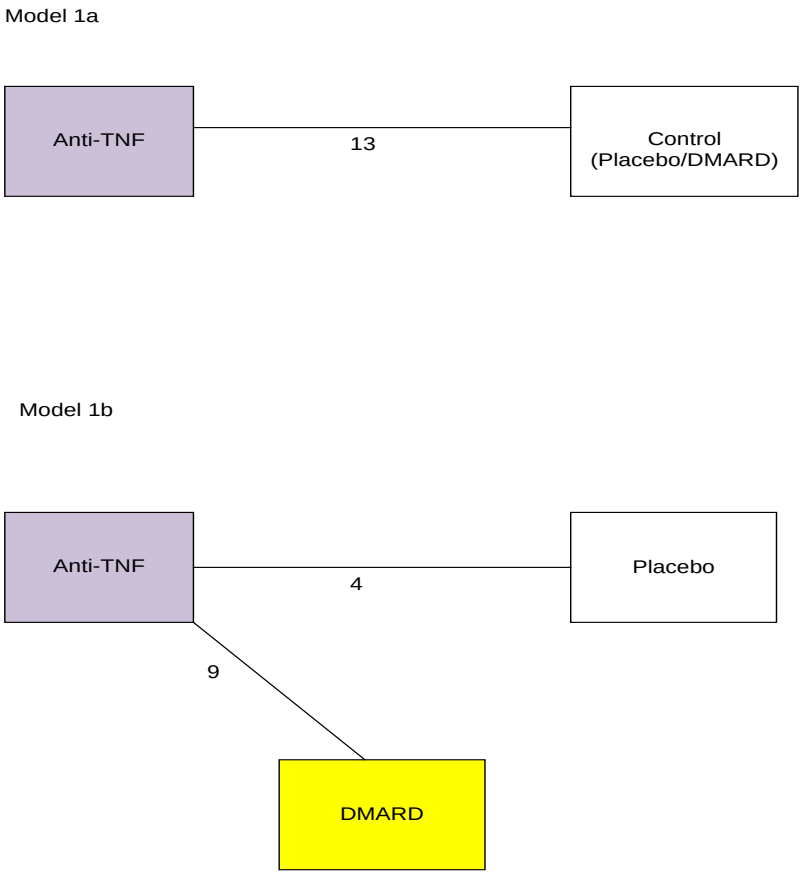
The initial analyses were conducted using MTC models increasing in complexity from a basic meta-analysis comparing anti-TNFs against non-anti-TNF controls, by incorporating individual anti-TNF drugs, additional DMARDs and dose of anti-TNF into the treatment definitions used in the MTC. At each increasing level of complexity the results were re-assessed for their implications into clinical treatment.

Sparsity of events across the dataset was a consideration throughout, and sensitivity analyses excluding a primary study that contributed to sparsity concerns were performed. Further extensions to the initial MTC models are considered in Chapter 10, including a sensitivity analysis across prior distributions used in the models, hierarchical modelling, and addition of constraints into the hierarchical models.

9.11 Mixed treatment comparison model diagrams

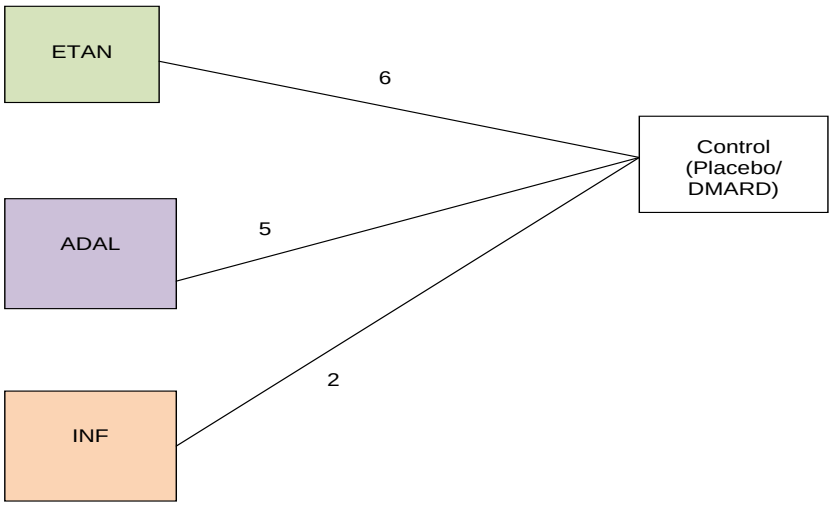
The following diagrams set out the MTC networks for the baseline models used in this chapter, as described in Section 9.5.

Figure 9.2: Network diagrams for Models 1a and 1b (described in Section 9.5.1).



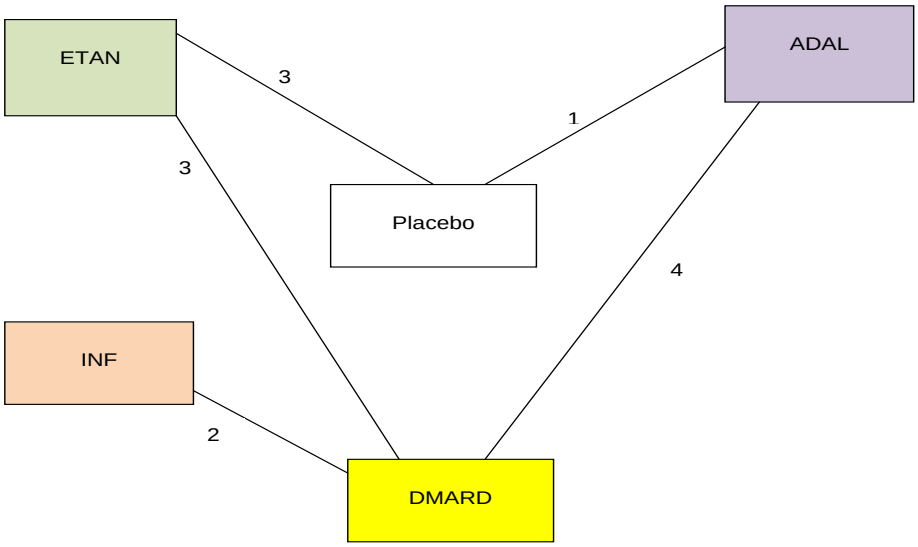
Anti-TNF: anti-tumour necrosis factor; DMARD: disease-modifying antirheumatic drug.

Figure 9.3: Network diagrams for Model 2a (described in Section 9.5.1).



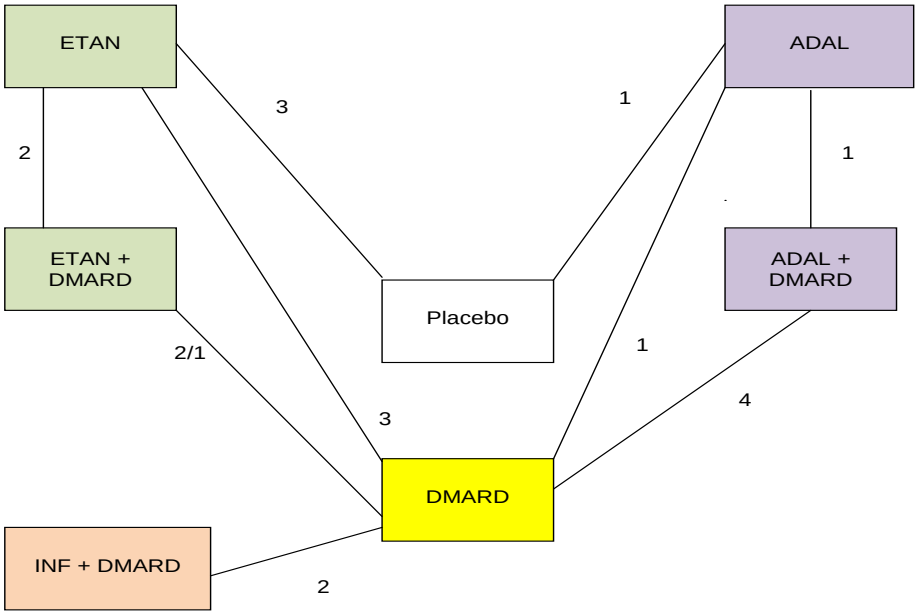
Adal: adalimumab; DMARD: disease-modifying antirheumatic drug; Etan: etanercept; Inf: infliximab.

Figure 9.4: Network diagrams for Model 2b (described in Section 9.5.1).



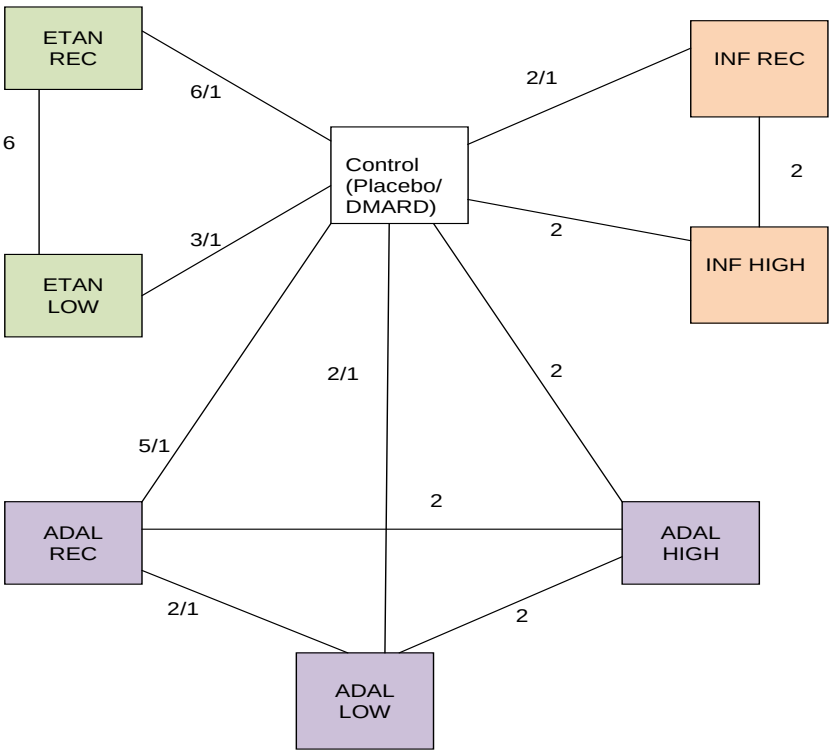
Adal: adalimumab; DMARD: disease-modifying antirheumatic drug; Etan: etanercept; Inf: infliximab.

Figure 9.5: Network diagrams for Model 3b (described in Section 9.5.1).



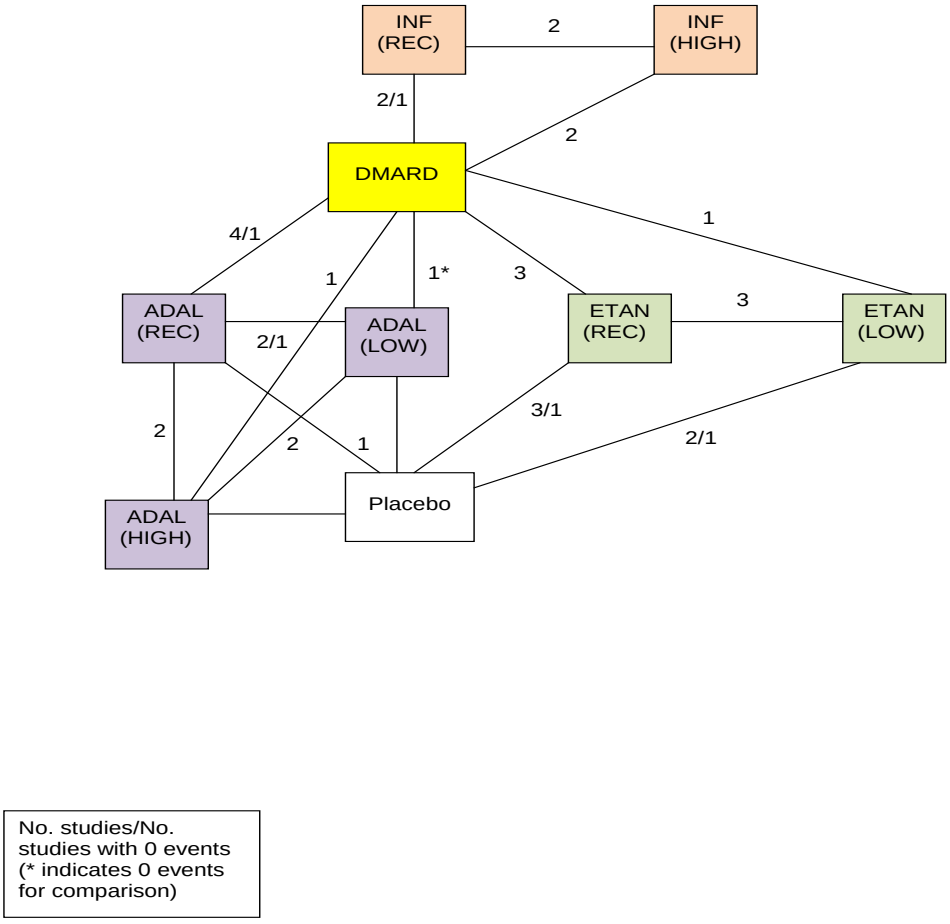
Adal: adalimumab; DMARD: disease-modifying antirheumatic drug; Etan: etanercept; Inf: infliximab.

Figure 9.6: Network diagrams for Model 4a (described in Section 9.5.1).



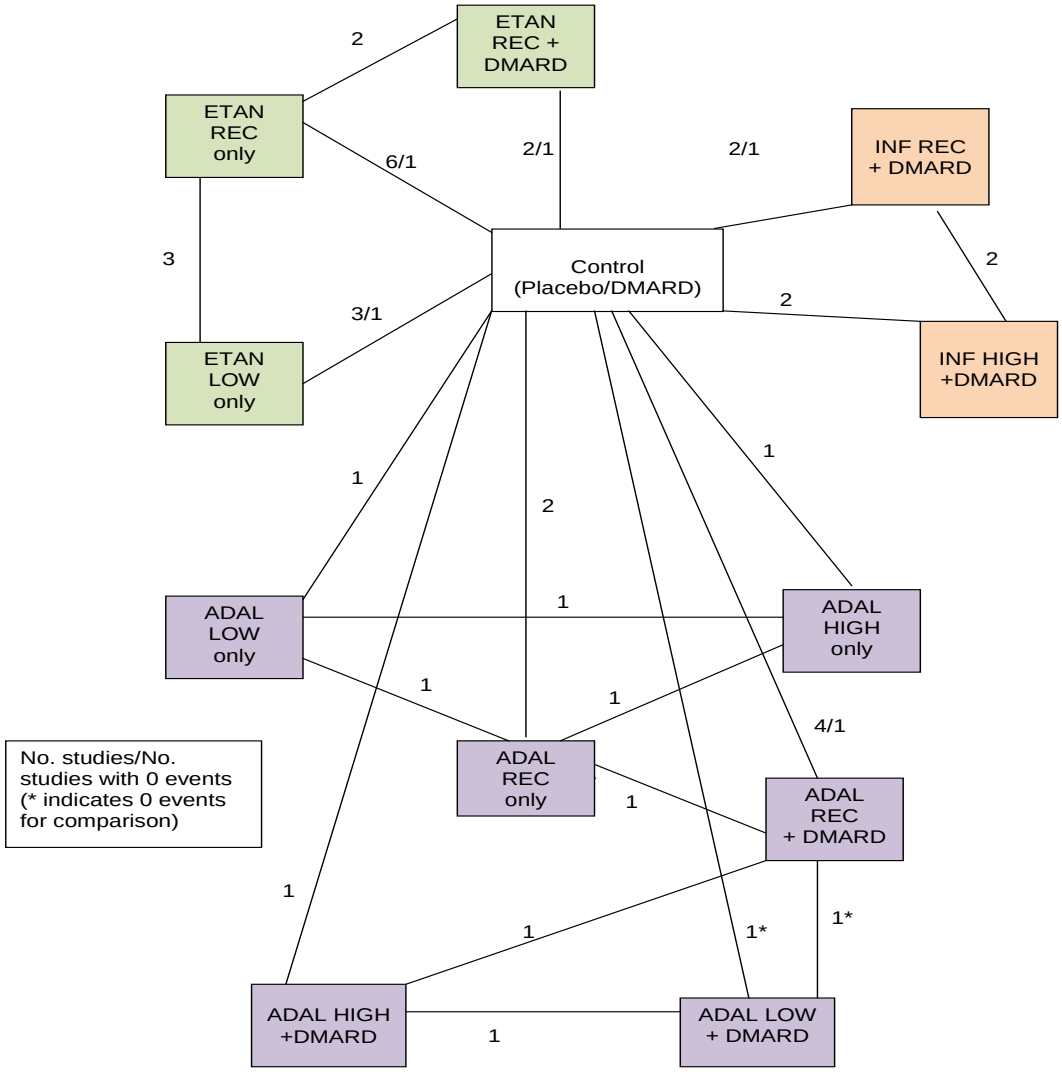
Adal: adalimumab; DMARD: disease-modifying antirheumatic drug; Etan: etanercept; Inf: infliximab.

Figure 9.7: Network diagrams for Model 4b (described in Section 9.5.1).



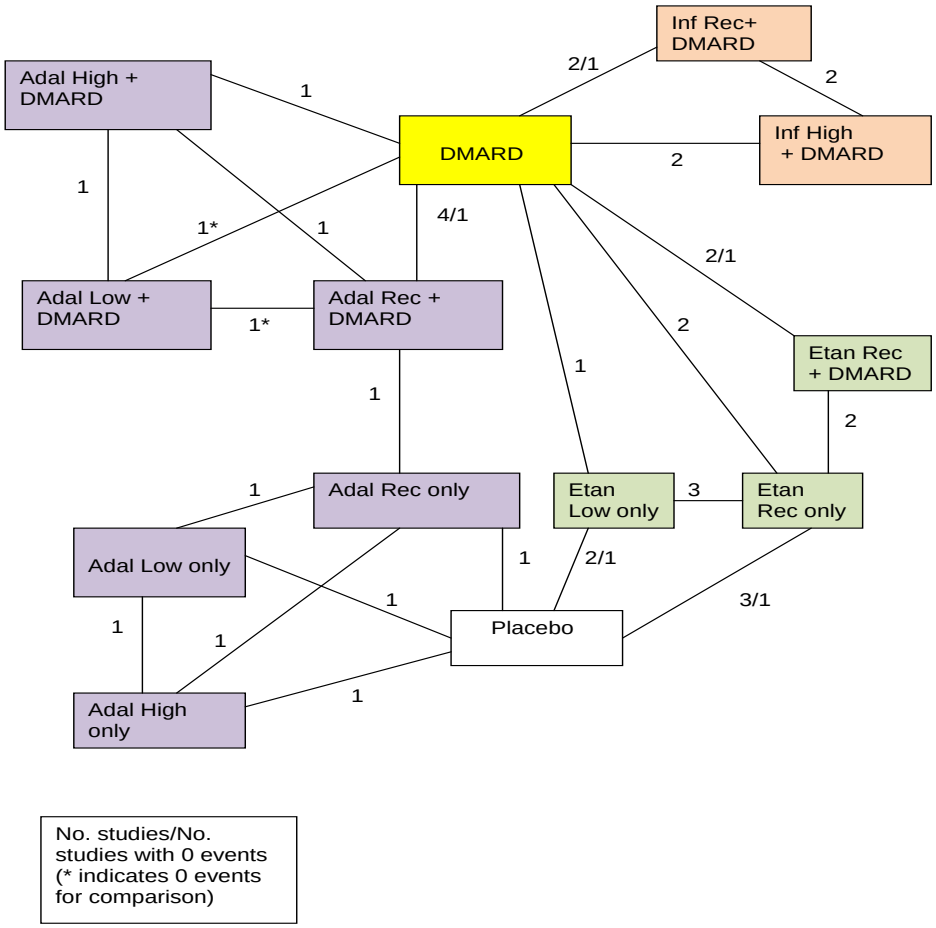
Adal: adalimumab; DMARD: disease-modifying antirheumatic drug; Etan: etanercept; Inf: infliximab.

Figure 9.8: Network diagrams for Model 5a (described in Section 9.5.1).



Adal: adalimumab; DMARD: disease-modifying antirheumatic drug; Etan: etanercept; Inf: infliximab.

Figure 9.9: Network diagrams for Model 5b (described in Section 9.5.1).



Adal: adalimumab; DMARD: disease-modifying antirheumatic drug; Etan: etanercept; Inf: infliximab.

10

Extensions to mixed treatment comparison models

10.1 Introduction

This chapter extends the mixed treatment comparison (MTC) analyses of Chapter 9, and uses the same clinical example, of anti-TNF therapy and the risk of malignancy in rheumatoid arthritis (RA). Although the analyses presented in Chapter 9 made some progress in interpreting the relationship between anti-TNF therapy and malignancy, the results did not demonstrate sufficient information regarding certain aspects of anti-TNF therapy, for example, the influence of individual anti-TNF and dose, to be valuable in clinical decision-making. The main aim of this chapter is to further refine and develop the MTC methods to attempt to shed further light on these two factors of anti-TNF therapy.

By developing methods specifically tailored to this clinical example, it may be feasible to apply such methods to other clinical examples, where certain aspects of a therapy require investigation. For example, interactions between different drugs may be an area where MTC analysis with hierarchical modelling could be of value, or MTC methods could be applied to demographic factors, such as sex, where the demographic factor could be viewed as an element of 'treatment'.

This chapter extends the work of Chapter 9 in three main ways:

1. use of hierarchical modelling to 'borrow strength' across studies with com-

mon elements of the treatment regime, e.g. using the same anti-TNF drug, or the same dose level across different anti-TNFs;

2. placing constraints on parameters within a hierarchical model to reflect prior beliefs regarding the relationship between aspects of the treatment regime (e.g. dose) and treatment effect; and
3. performing a sensitivity analysis across multiple prior distributions for certain parameters within the model.

These methods are discussed in more detail in Sections 10.2, 10.3.2, 10.3.3, for the hierarchical modelling methods, and Section 10.3.6 for the sensitivity analysis across prior distributions. This chapter is a direct continuation of Chapter 9, in that the same dataset (see Section 9.6) is used throughout.

10.2 Background of hierarchical models and constraints

A further extension of meta-analysis methods that may be applied to MTC models comes in the form of using hierarchical models, as discussed by Prevost *et al*. (2000). When using a random effects (RE) model, it may not be reasonable to assume that all studies are deriving their true underlying treatment effect from a common distribution. For example, there may be fundamental differences between studies that would suggest that this is not the case.

An example of data from different study types, such as trials and observational studies, is used by Prevost *et al*. (2000), but different treatments or study populations may have similar effects. One potential approach would be to use separate meta-analyses for the different study types, but this method would prevent exchange of information or 'borrowing strength' across the study types. An alternative approach would be to insert an additional layer of information into the meta-analysis model, allowing studies to be indexed by study type. In this way, the different study types are enabled to be derived from different distributions with different central locations. An additional level of heterogeneity is introduced to the model, that between studies of a certain type, in addition to the between-studies heterogeneity that existed in the original basic RE model.

An alternative approach would be to assume partial exchangeability across the studies, with some of the variability being modelled by covariates, and

some modelled using an RE model. This approach is discussed by Higgins *et al*. (2009).

The hierarchical model proposed by Prevost *et al*. (2000) has certain extensions that are potentially of interest to MTC models. The first of these relates to using different prior beliefs to inform the models. For example, using the anti-TNF and RA example, there may be a prior clinical belief that the specific anti-TNF drug is more highly associated with malignancy risk than dose, or that there is a dose–response relationship, such that the higher doses of all anti-TNFs are associated with higher risk of malignancy. To reflect these scenarios, a hierarchical model could be used to ensure that 'borrowing strength' across datapoints occurs more strongly between individual drugs, or between doses, as appropriate.

In the example used by Prevost *et al*. (2000), different priors can be applied to the heterogeneity that exists between studies of different design. Each overall relative treatment effect for each study type can be modelled as coming from a population with equal means, but the variances differ. Those study types with higher variance will in effect downgrade the weight of evidence provided by that study type and will move the overall (across all study types) relative treatment effect closer to those of other study types with smaller variances. A further extension to hierarchical models is the use of study-level covariates, such as age, to explain difference in relative treatment effects between age groups.

Sensitivity to prior distributions may also be an issue worthy of investigation in some models. The reason put forward by Prevost *et al*. (2000) is that prior distributions are often selected on a subjective or *ad hoc* basis, so therefore it is important to investigate the degree of influence the prior distribution may have on the results. Also, a sensitivity analysis is desirable due to the additional levels within the hierarchical model, with the added variance components. Any influence of the prior distribution on the outcome would be propagated throughout the model.

Another aspect of hierarchical modelling that may be of use in MTC modelling is the addition of a constraint to the priors on any stochastic nodes in the model (Prevost *et al*. 2000). Such constraints would need to be specific to the dataset being analysed to ensure they were appropriately applied to the model. The example put forth by Prevost *et al*. (2000) relates to a situation

where observational studies are less biased than trials, and hence the error terms for the trials are constrained to be smaller in magnitude than those for the observational studies (although the bias could occur in either direction of the outcome, it will be smaller in the trials compared to observational studies).

The use of constraints within a hierarchical model has been previously discussed by Prevost *et al.* (2000), using an example derived from breast cancer screening. In this model, a prior belief was that studies with randomisation were less biased than those that were non-randomised. This assumption was modelled by enforcing the error for randomised studies to be derived from a normal distribution centred on zero, with a mean of 1, truncated by the positive and negative magnitude of the error for non-randomised studies. The error for non-randomised studies was then constrained to fall within the tails of this distribution.

Alternatively, letting μ be the estimated mean of the population of study effects (for all study types), and letting θ_i be the mean estimated value for each study type i, where Study Type 1 is randomised studies, and Study Type 2 is non-randomised studies, then

$$|\mu - \theta_1| < |\mu - \theta_2| \tag{10.1}$$

or

$$|\varepsilon_1| < |\varepsilon_2|, \tag{10.2}$$

where ε_i refers to the error for each study type.

In effect, the bias for randomised studies is lower than that for non-randomised studies, although there is no way to ascertain a prior belief regarding the direction of the bias, just that the magnitude is greater for non-randomised compared to randomised studies.

10.3 Methods

10.3.1 Baseline methods and dataset creation

In this chapter, the same MTC networks (shown diagrammatically in Section 9.11) are used. The models in this chapter are developed from extension of those described in Section 9.5.2, by application of more complex statistical analyses. The modelling was implemented in WinBUGS 1.4, as discussed in Section 9.4.6. The dataset used is the same as that of Chapter 9, described in Section 9.6.

10.3.2 D. Random effects model with hierarchy on treatment effects

This model, and Model E (described in Section 10.3.3), are continuations of Models A–C, described in Section 9.4.

In this and subsequent models, the adjustment described in Section 9.4.4 for non-independent arms in multi-arm trials is employed when appropriate.

The most basic level of hierarchical model was employed in conjunction with Model 2a, described in Section 9.5.2. In this model, the three individual anti-TNFs were considered separately, compared against all non-anti-TNF arms. In the initial random effects (RE) model, the treatment effects d_k, for each non-baseline treatment k, were considered to come from a vague normal distribution, centred on zero. In the hierarchical model, the non-baseline d_k values also come from a normal distribution, but for each d_k, the distribution is centred on a different value, denoted μ_d, as follows:

$$d_k \sim \text{Normal}(\mu_d, \tau_d^2), \tag{10.3}$$

$$\mu_d \sim \text{Normal}(0, 1000), \tag{10.4}$$

$$\tau_d \sim \text{Uniform}(0, 2). \tag{10.5}$$

This hierarchy allows for greater variation in the values of the treatment effects, whilst still retaining a distributional connection.

A further hierarchical model was used to investigate the situation when an hierarchy of effects is placed on the model where both anti-TNF and dose are considered. This hierarchy can be envisaged with dose varying within drug, or with drug varying within dose. For the former model, each drug is modelled as coming from a different normal distribution, which is vague in all cases, but with a different mean in each case; drug is 'above' dose in the hierarchy. Each drug–dose combination is modelled as coming from a distribution determined by the drug. When the hierarchy has dose 'above' drug, each dose–drug combination is modelled as coming from a different vague normal distribution, determined by the dose. In other respects, the models remain as for the standard RE model.

For example, consider a hierarchy with dose j above individual drug k, and individual study i at the lowest level. At the bottom level of the hierarchy, the model is:

$$nd_{i,j,k} \sim \text{Normal}(d_{j,k}, \tau^2_{i,j,k}), \tag{10.6}$$

where the study-level log odds ratio (OR), by comparison to a baseline treatment, for a specific dose–drug combination for study i, denoted above by $nd_{i,j,k}$, is distributed normally with a mean of $d_{j,k}$, variance, $\tau^2_{i,j,k}$.

At the superseding level the model is:

$$d_{j,k} \sim \text{Normal}(\mu_j, \tau^2), \tag{10.7}$$

where μ_j refers to the mean log OR for each individual dose j, in this case recommended, low or high, $d_{j,k}$ refers to the log OR for the specific dose–drug combination, and is distributed normally, with a mean of μ_j and variance τ^2. Finally, at the highest level in the hierarchy, we have the distribution of μ_j, which requires a prior distribution. Again, vague prior distributions can be placed on the hyperparameters of μ_j, as in Equations 10.4–10.5.

10.3.3 E. Random effects model with hierarchy on treatment effects and constraints

The final variation on the MTC model comes with adding a constraint to the model parameters. In the example used here, there is a clinical argument that higher doses of anti-TNF are more likely to be associated with higher rates of malignancy than lower doses. This equates to a very strong prior belief, and this can be translated into the MTC model by means of adding constraints onto prior beliefs regarding the odds of malignancy at each dose level.

The example set out in the MTC is not directly comparable with the example used by Prevost *et al.* (2000), and hence a new approach to incorporating constraints within the model has been developed. In the MTC there are three dose levels, and the magnitude of effect on malignancy risk is believed to be such that the low dose has the lowest risk, the recommended dose has a risk greater than that of the lowest dose, but lower than that of the highest dose, which has the highest risk. This scenario differs from that of Prevost *et al.* (2000) in two ways. Firstly, there is a prior belief regarding the direction of the effect across the three doses, and secondly, there are three dose levels (as opposed to two study types) that must be correctly arranged in order of prior belief regarding magnitude of effect size.

To address this issue, the prior distributions placed on each dose level require truncation to ensure that the mean value of the prior distribution for odds of malignancy for the recommended dose is greater than that for the low dose, and lower than that for the high dose. Using a model with eight treatments, including a non-anti-TNF control, the odds of malignancy for low, recommended and high dose were placed in a hierarchy with individual anti-TNF below them in the hierarchy. Using recommended dose as the baseline, as this dose was present across all three anti-TNFs, the odds for malignancy in the low and high doses were set to be related to the odds for the recommended dose, by addition of a difference factor for each dose. The difference factor for the low dose was then set to be negative, based on a half-normal distribution truncated to be below zero, and similarly the difference factor for the high dose was set to be positive.

As a continuation of the model without constraints set out in Equations 10.6 and 10.7, the constraints can be added as follows:

$$\mu_2 = \mu_1 + \eta 1, \tag{10.8}$$

$$\mu_3 = \mu_1 + \eta 2, \tag{10.9}$$

$$\eta_1 \sim \text{Normal}(0, 10\ 000)\text{I}(,0), \tag{10.10}$$

$$\eta_2 \sim \text{Normal}(0, 10\ 000)\text{I}(0,), \tag{10.11}$$

where μ_j refers to log OR of malignancy in the three dose groups (recommended $= 1$, low $= 2$ and high $= 3$), η_1 and η_2 are the differences in odds between the recommended dose and the low and high groups respectively, I(,0) indicates that the normal distribution is truncated above zero (can take only negative values), and I(0,) indicates that the normal distribution is truncated below zero (can take only positive values).

It is important to note that the parameters for normal distributions placed on the η_1 and η_2 stochastic nodes are themselves defined as numbers and not hyperparameters with distributions placed on them. This is a crucial distinction, as the values for the hyperparameters would then be influenced by the truncation placed on η_1 and η_2, which may constrict them to artificial values and invalidate the model.

10.3.4 Application of hierarchical mixed treatment comparison models

The basic hierarchical models (Model D; Section 10.3.2) are applied in two scenarios, firstly for the scenario outlined in Model 2a (Section 9.5.2), in which individual nodes in the MTC network are defined by anti-TNF only, and secondly for the scenario outlined in Model 4a (Section 9.5.2) in which individual nodes are defined by the combination of individual anti-TNF and dose. Hence, the modelling scenario of Model 4a, whereby two parameters are used to define a treatment, allows the addition of an extra level within the hierarchy. Model 5a, in which treatment nodes are defined on three parameters, individual anti-TNF, dose and additional disease-modifying anti-rheumatic (DMARD), would allow the inclusion of another level within the hierarchical model, but due to potential

difficulties, both with fitting the model, and interpretation of results, this added level within the hierarchy was not attempted.

The hierarchical model with constraints (Model E; Section 10.3.3) is used with regard to Model 4a only, as the aim was to evoke a situation whereby both drug and dose had an influence on malignancy, but with dose as the more prominent factor, and an obvious 'ascending order' from low to high in terms of malignancy risk.

10.3.5 Summary of hierarchical mixed treatment comparison models

The MTC models described above are summarised below.

1. Random effects model with hierarchy on treatment effects: random effects model where treatments are defined according to different parameters e.g. by anti-TNF and dose, with modelling of treatment effects according to the defined hierarchy.
2. Random effects model with hierarchy on treatment effects, with constraints: random effects model where treatments are defined according to different parameters e.g. by anti-TNF and dose, with modelling of treatment effects according to the defined hierarchy, using prior beliefs to impose constraints on the distribution of different treatment effects at the same level of the hierarchy.

These methods extend those set out in Section 9.4.5, where the definitions of a 'treatment' and 'dose' are provided. The basic MTC models from which these models were developed are described in Section 9.4.1, with the MTC networks set out in Sections 9.5.1 and 9.5.2. The constraints mentioned above are discussed in more detail in Section 10.3.3.

10.3.6 Alternative prior distributions

Being mindful of the sparsity of the dataset, with the associated risk of any supposedly 'vague' or 'non-informative' priors in actuality exerting influence over the dataset, it was decided to perform a sensitivity analysis by including a range of alternative priors on two separate MTC models, both using random effects.

The Prior Sets are described below. The prior distributions across models are discussed further in Section 9.4.6.

A. μ: normal distribution, mean 0, precision 0.0001; d: normal distribution, mean 0, precision 0.001 or 0.0001 depending on model; standard deviation: uniform distribution, parameters 0,2; (standard priors).

B. μ: normal distribution, mean 0, precision 0.000001; d: normal distribution, mean 0, precision 0.000001; standard deviation: uniform distribution, parameters 0,2.

C. μ: normal distribution, mean 0, precision 0.0001; d: normal distribution, mean 0, precision 0.001 or 0.0001 depending on model; standard deviation: uniform distribution, parameters 0,5.

D. μ: normal distribution, mean 0, precision 0.000001; d: normal distribution, mean 0, precision 0.000001; standard deviation: uniform distribution, parameters 0,5.

E. μ: normal distribution, mean 0, precision 0.0001; d: normal distribution, mean 0, precision 0.001 or 0.0001 depending on model; standard deviation: half-normal distribution, mean 0, precision 0.001.

F. μ: normal distribution, mean 0, precision 0.000001; d: normal distribution, mean 0, precision 0.000001; standard deviation: half-normal distribution, mean 0, precision 0.001.

G. μ: normal distribution, mean 0, precision 0.0001; d: normal distribution, mean 0, precision 0.001 or 0.0001 depending on model; standard deviation set to equal $1/\sqrt{\tau}$, τ has gamma distribution with parameters 0.001, 0.001.

H. μ: normal distribution, mean 0, precision 0.000001; d: normal distribution, mean 0, precision 0.000001; standard deviation set to equal $1/\sqrt{\tau}$, τ has gamma distribution with parameters 0.001, 0.001.

The models selected for sensitivity analysis across priors were Models 2b and 4b. These models were chosen for comparison purposes. Model 2b was a simpler model with fewer treatments, did not require adjustment across multi-arm trials, and had less data sparsity than Model 4b, which had a larger number of treatments in a more complex model, increased sparsity of data and correlation for multi-arm trials. Model 5b was considered as a potential model for sensitiv-

Table 10.1: Median log odds ratios for hierarchical model based on Model 2a with log odds ratio for each anti-TNF assumed to come from same distribution.

Treatment	Median† LOR*	95% CrI
Etanercept	0.947	-0.013; 2.239
Adalimumab	0.939	-0.042; 2.249
Infliximab	1.139	-0.094; 3.113

† median of posterior mean distribution; * baseline for LOR is Control; CrI: credible interval; LOR: log odds ratio.

ity to priors but due to concerns regarding possible convergence issues was not selected.

For all sets of priors the burn-in was 10 000 iterations, convergence having been attained at this point, followed by a sample of 50 000 iterations. The exceptions to this rule were Prior Set E, for which convergence was poor at 10 000 iterations, therefore a burn-in of 20 000 iterations was preferred, and Prior Set F which required a burn-in of 50 000 iterations.

10.4 Results and initial discussion

10.4.1 Use of hierarchical models

In order to investigate any possible differences in outcome by 'borrowing strength' across different treatment categories that are in some way similar, such as containing the same drug or dose, several hierarchical models have been used. In all examples with hierarchical models, RE models (which are by definition hierarchical) are used.

The simplest of these involved a model where the treatment effects for three drugs were assumed to come from some overall distribution for anti-TNFs in general. The results of this model are set out in Table 10.1 for comparison to the non-hierarchical model.

These results can be compared with those set out in Table 9.4, for Model 2a. The results are as would be expected, in that the hierarchical model brings together the posterior distributions for the log ORs. For example, from Model 2a (non-hierarchical), the smallest median log OR (using the RE model) is 0.875,

Table 10.2: Median log odds ratios for hierarchical model based on Model 4a with log odds ratio for each anti-TNF assumed to come from same distribution.

Treatment	Median† LOR*	95% CrI	CrI width
Etanercept (Rec)	0.862	-0.048; 1.981	2.029
Etanercept (Low)	0.843	-0.230; 2.104	2.334
Adalimumab (Rec)	0.833	-0.090; 1.99	2.080
Adalimumab (Low)	0.824	-0.691; 2.218	2.909
Adalimumab (High)	0.931	-0.254; 2.501	2.755
Infliximab (Rec)	0.746	-0.922; 2.231	3.153
Infliximab (High)	1.153	0.059; 3.139	3.198

† median of posterior mean distribution; * baseline is Control; CrI: credible interval; LOR: odds ratio; Rec: recommended.

for adalimumab, while the highest median log OR was 1.928 for infliximab. The range of median log ORs is clearly narrower using the hierarchical model.

Having established that the hierarchical models work as would be expected, the next step was to use these models in two separate hierarchies, one with drug at the upper level, with dose categories below drug in the hierarchy, the second with dose at the upper level, with drug below dose. In the first model, the different treatments including the same drug are assumed to come from an overarching distribution (doses within drug are exchangeable), in the second, the different treatments including the same drug are assumed to come from an overarching distribution (drugs within dose are exchangeable).

Looking first at the model where all treatments using the same drug are considered to be derived from the same distribution, and with non-anti-TNF controls, the median log ORs are set out in Table 10.2.

Comparing these results with those set out in Model 4a, Table 9.7, for the RE version of Model 4a, looking at adalimumab, the median log OR for the low dose was 0.290, with the median log OR for the high dose at 1.413 (the corresponding value for the recommended dose lying between these two extremes). It is therefore evident that applying the hierarchy on drug brings the results for the three different dose levels for adalimumab closer together. Similarly, for infliximab the median log ORs for the non-hierarchical model at the recommended dose was -0.127 compared to 0.746 in the model above, effectively going from an estimated lower incidence of malignancy to higher incidence of malignancy. The sum of summed deviances for this model was 33.61, compared to 34.28 for the equivalent non-hierarchical RE model.

Table 10.3: Median log odds ratios for hierarchical model based on Model 4a with log odds ratio for each dose level assumed to come from same distribution.

Treatment	Median† LOR*	95% CrI	CrI width
Etanercept (Rec)	0.828	-0.066; 1.890	1.956
Etanercept (Low)	0.838	-0.273; 2.097	2.370
Adalimumab (Rec)	0.813	-0.064; 1.897	1.961
Adalimumab (Low)	0.829	-0.725; 2.240	2.965
Adalimumab (High)	1.154	-0.076; 2.806	2.882
Infliximab (Rec)	0.675	-0.792; 1.942	2.734
Infliximab (High)	1.322	0.226; 3.137	3.363

† median of posterior mean distribution; * baseline is Control; CrI: credible interval; LOR: odds ratio; Rec: recommended.

An alternative hierarchical model enforces the same dose level across different anti-TNFs to be derived from the same distribution for log OR (drugs within dose are exchangeable). The results are shown in Table 10.3.

The total sum of deviances for this model was 33.13, compared to 34.28 for the non-hierarchical model, and 33.61 for the model with dose exchangeable within drug in the hierarchy.

10.4.2 Use of models with constraints on prior distributions

The model using constraints on the LOR between different doses of anti-TNF presented some difficulties in execution, namely that the use of multiple chains with different initial values did not appear to be viable. However, on using one chain, convergence was achieved by 10 000 iterations, and a sample size of a further 50 000 iterations was derived. The results are shown in Table 10.4.

It is useful to compare the results from the model with constraints to those derived from the equivalent non-constrained model, a hierarchical model with drugs exchangeable within dose, without constraints, which are set out in Table 9.7.

The total sum of deviances for the non-constraint model was 33.13, compared to 32.75 for the model with constraints (for 35 datapoints). It is also useful to compare the underlying log ORs for the three dose levels across the two models, as set out in Table 10.5.

Table 10.4: Median log odds ratios for hierarchical model based on Model 4a with log odds ratio for each dose level assumed to come from same distribution, plus comparison of prior distributions with and without constraints.

	No constraints		Constraints	
Treatment	Median† LOR*	95% CrI	Median† LOR*	95% CrI
Etanercept (Rec)	0.828	-0.066; 1.890	0.721	-0.161; 1.861
Etanercept (Low)	0.838	-0.273; 2.097	0.346	-1.032; 1.653
Adalimumab (Rec)	0.813	-0.064; 1.897	0.761	-0.153; 1.953
Adalimumab (Low)	0.829	-0.725; 2.240	0.254	-1.555; 1.724
Adalimumab (High)	1.154	-0.076; 2.806	1.930	0.327; 3.774
Infliximab (Rec)	0.675	-0.792; 1.942	0.666	-0.891; 2.094
Infliximab (High)	1.322	0.226; 3.137	2.153	0.745; 4.065

† median of posterior mean distribution; * baseline is Control; CrI: credible interval; LOR: odds ratio; Rec: recommended.

Table 10.5: Underlying LORs for each dose level, comparing models with and without constraints.

Dose	Constraints	Median† LOR*	95% CrI
Rec	No	0.815	0.207; 1.936
Rec	Yes	0.764	-0.238; 2.026
Low	No	0.871	-0.370; 2.192
Low	Yes	0.184	-0.519; 1.347
High	No	1.143	-0.044; 2.724
High	Yes	2.079	0.684; 4.057

† median of posterior mean distribution; * baseline is Control; CrI: credible interval; LOR: odds ratio; Rec: recommended.

As there was not a strong clinical indication *a priori* that any of the anti-TNFs presented a higher risk of malignancy than any others, there was no reason to perform a constraints model based on anti-TNF.

10.4.3 Comparisons across models

The models described above display a wide variety of levels of complexity, and additional modelling features. To assist in comparing these models, results from multiple treatments are set out in Tables 10.6–10.8, each table including results relating to treatments including one of the three anti-TNFs. These cross-model comparisons would be useful in comparing treatments within and between anti-TNFs, and could be used in conjunction with clinical background knowledge. Both adalimumab and infliximab have results that are significant statistically, and potentially clinically, which vary according to the model selected; this phenomenon is perhaps most dramatically demonstrated for infliximab, in Table 10.8. High-dose infliximab has significant results for both hierarchical models, and for the model with constraints, but when using the straightforward model using dose (Model 4a), the results are not significant. Hence, a careful consideration of whether hierarchical models and the use of constraints are appropriate, both clinically and statistically, is required.

Table 10.6: Comparison of effects of treatments including etanercept at increasing levels of specificity. Random effects models only. Baseline for comparison is non-anti-TNF control unless stated otherwise.

	median † LOR (95% CrI)								
Model	1a	2a	3b*	4a	5a	HM 2a	HM 4a, drug*	HM 4a, dose**	HM constraints
Treatment									
All anti-TNF	0.908 (0.176; 1.995)	NA	NA	NA	NA	NA	NA	NA	NA
Etan	NA	0.910 (-0.314; 2.598)	0.839 (-1.160; 3.409)	NA	NA	0.947 (-0.013; 2.239)	NA	NA	NA
Etan + D	NA	NA	0.917 (-1.565; 3.460)	NA	NA	NA	NA	NA	NA
Etan (Rec)	NA	NA	NA	0.885 (-0.467; 2.536)	0.894 (-0.449; 2.670)	NA	0.862 (-0.048; 1.981)	0.828 (-0.066; 1.890)	0.721 (-0.161; 1.861)
Etan (Low)	NA	NA	NA	0.824 (-0.972; 2.966)	0.812 (-0.990; 3.115)	NA	0.843 (-0.230; 2.104)	0.838 (-0.273; 2.097)	0.346 (-1.032; 1.653)
Etan (Rec) + D	NA	NA	NA	NA	0.774 (-1.562; 3.152)	NA	NA	NA	NA

† median of posterior mean distribution; * drug is above dose in hierarchy; ** dose is above drug in hierarchy; CrI: credible interval; LOR: odds ratio; Rec: recommended; shading highlights 'significant' results.

Table 10.7: Comparison of effects of treatments including adalimumab at increasing levels of specificity. Random effects models only. Baseline for comparison is non-anti-TNF control unless stated otherwise.

	median † LOR (95% CrI)								
Model	1a	2a	3b*	4a	5a	HM 2a	HM 4a, drug*	HM 4a, dose**	HM constraints
Treatment									
All anti-TNF	0.908 (0.176; 1.995)	NA	NA	NA	NA	NA	NA	NA	NA
Adal	NA	0.875 (-0.378; 2.599)	0.424 (-2.009; 3.287)	NA	NA	0.939 (-0.042; 2.249)	NA	NA	NA
Adal + D	NA	NA	1.036 (-0.523; 3.099)	NA	NA	NA	NA	NA	NA
Adal (Rec)	NA	NA	NA	0.812 (-0.519; 2.561)	NA	NA	0.833 (-0.90; 1.99)	0.813 (-0.064; 1.897)	0.761 (-0.153; 1.953)
Adal (Low)	NA	NA	NA	0.290 (-3.366; 3.431)	NA	NA	0.824 (-0.691; 2.218)	0.829 (-0.725; 2.240)	0.254 (-1.555; 1.724)
Adal (High)	NA	NA	NA	1.413 (-1.292; 4.452)	NA	NA	0.931 (-0.254; 2.501)	1.154 (-0.076; 2.806)	1.930 (0.327; 3.774)
Adal (Rec) + D	NA	NA	NA	NA	0.893 (-0.659; 2.885)	NA	NA	NA	NA
Adal (Low) + D	NA	NA	NA	-42.08 (-166.5; 51.43)	NA	NA	NA	NA	NA
Adal (High) + D	NA	NA	NA	NA	51.58 (3.274; 193.6)	NA	NA	NA	NA

† median of posterior mean distribution; * drug is above dose in hierarchy; ** dose is above drug in hierarchy; CrI: credible interval; LOR: odds ratio; Rec: recommended; shading highlights 'significant' results.

Table 10.8: Comparison of effects of treatments including infliximab at increasing levels of specificity. Random effects models only. Baseline for comparison is non-anti-TNF control unless stated otherwise.

	median † LOR (95% CrI)								
Model	1a	2a	3b*	4a	5a	HM 2a	HM 4a, drug	HM 4a, dose	HM constraints
Treatment									
All anti-TNF	0.908 (0.176; 1.995)	NA	NA	NA	NA	NA	NA	NA	NA
Inf	NA	1.928 (-0.410; 5.655)	NA	NA	NA	1.139 (-0.094; 3.113)	NA	NA	NA
Inf + D	NA	NA	2.025 (-0.513; 5.848)	NA	NA	NA	NA	NA	NA
Inf (Rec)	NA	NA	NA	-0.127 (-4.121; 3.865)	NA	NA	0.746 (-0.922; 2.231)	0.675 (-0.792; 1.942)	0.666 (-0.891; 20.94)
Inf (High)	NA	NA	NA	2.303 (-0.053; 5.974)	NA	NA	1.153 (0.059; 3.139)	1.322 (0.226; 3.137)	2.153 (0.745; 4.065)
Inf (Rec) + D	NA	NA	NA	NA	-0.097 (-4.092; 3.904)	NA	NA	NA	NA
Inf (High) + D	NA	NA	NA	NA	2.299 (-0.106; 6.062)	NA	NA	NA	NA

† median of posterior mean distribution; * drug is above dose in hierarchy; ** dose is above drug in hierarchy; CrI: credible interval; LOR: odds ratio; Rec: recommended; shading highlights 'significant' results.

10.4.4 Alternative prior distributions

The results for the sensitivity analysis across prior distributions set out in Section 10.3.6 are presented in this section.

The results for Model 2b are set out in Table 10.9. Selected results for specified treatments are also presented for Model 4b, with the same set of priors as above. In this model, the three adalimumab dose levels were selected as treatments for comparison of priors. This was due to the existence of some sparsity of data across the three doses, and therefore these treatments may be sensitive to the selected prior. The results for these models are set out in Table 10.10.

The priors impact primarily on the posterior distribution for the standard deviation in each model, and through the standard deviation exert an influence on the OR and log OR. Therefore, to understand the influence of any prior on the results, it is helpful to consider the values of the standard deviation for each set of priors, and in particular the shape of the posterior density. Table 10.11 sets out the relevant values for the standard deviation. Densities for the standard deviation for Model 2b with the various priors are set out in Figure 10.1, and for the Model 4b the densities are set out in Figure 10.2 (see Section 10.7).

Table 10.9: Median log odds ratios for Model 2b with priors A–H.

Prior set	Median† LOR	95% CrI	Median† LOR	95% CrI	Median† LOR	95% CrI
	Etanercept against placebo		Adalimumab against placebo		Infliximab against placebo	
A	0.683	0.032; 3.073	0.553	-0.192; 3.289	1.587	0.407; 6.117
B	0.673	-1.209; 3.086	0.574	-1.615; 3.340	1.533	-1.869; 5.978
C	0.780	-1.394; 4.374	0.674	-1.959; 4.558	1.655	-2.389; 7.318
D	0.831	-1.388; 4.552	0.704	-1.98; 4.774	1.653	-2.469; 7.762
E	0.836	-1.499; 5.114	0.742	-2.174; 5.348	1.768	-2.819; 8.350
F	0.865	-1.56; 6.116	0.726	-2.341; 6.402	1.751	-3.082; 9.757
G	0.631	-1.092; 3.043	0.530	-1.451; 3.211	1.507	-1.616; 6.054
H	0.588	-1.080; 3.058 H	0.471	-1.465; 3.179	1.453	-1.626; 5.782

† median of posterior mean distribution; CrI: credible interval; LOR: log odds ratio.

Table 10.10: Median log odds ratios for Model 4b with priors A–H: adalimumab (recommended) compared to placebo.

Prior set	Median† LOR	95% CrI	Median† LOR	95% CrI	Median† LOR	95% CrI
	Adalimumab (recommended) against placebo		Adalimumab (low) placebo		Adalimumab (high) against placebo	
A	0.463	-1.889; 3.195	0.071	-4.079; 3.625	1.219	-1.827; 4.696
B	0.451	-1.965; 3.189	0.041	-4.157; 3.618	1.214	-1.877; 4.738
C	0.491	-2.629; 4.338	0.088	-4.745; 4.810	1.406	-2.229; 6.669
D	0.507	-2.635; 4.419	0.131	-4.705; 4.802	1.429	-2.185; 6.764
E	0.479	-3.054; 5.099	0.027	-5.281; 5.495	1.442	-2.353; 8.403
F	0.541	-3.429; 5.467	0.076	-5.743; 5.810	1.487	-2.501; 8.994
G	0.366	-1.876; 3.082	-0.069	-3.977; 3.374	0.954	-1.830; 4.688
H	0.290	-1.920; 2.958	-0.125	-4.449; 3.483	0.896	-1.839; 4.707

† median of posterior mean distribution; CrI: credible interval; LOR: log odds ratio.

Table 10.11: Median standard deviations for Models 2b and 4b with Prior Sets A–H.

Model 2b		
Prior set	Median sd	95% CrI
A	0.906	0.064; 1.914
B	0.910	0.063; 1.913
C	1.109	0.077; 3.972
D	1.159	0.073; 4.137
E	1.176	0.054; 5.640
F	1.246	0.065; 6.925
G	0.385	0.033; 2.729
H	0.385	0.033; 2.659
Model 4b		
A	0.956	0.044; 1.927
B	0.995	0.072; 1.934
C	1.297	0.095; 4.269
D	1.326	0.100; 4.296
E	1.383	0.108; 6.409
F	1.469	0.107; 7.473
G	0.438	0.33; 3.016
H	0.412	0.031; 3.139

CrI: credible interval; sd: standard deviation.

10.5 Further discussion and conclusions

10.5.1 Alternative prior distributions

The results for alternative Prior Sets are set out in Tables 10.9–10.10. The associated densities for the standard deviation are set out in Figures 10.1 and 10.2.

Use of different priors did have some influence over the central estimates and associated CrIs. For example, Model 2b, comparing etanercept with placebo, Prior Set A resulted in the narrowest CrI (Table 10.9), and yielded a significant result, whereas all other Prior Sets produced non-significant results. A similar effect of priors was seen when comparing infliximab with placebo, also in Model 2b (Table 10.9).

There was also a consistent tendency for priors that used a gamma distribution (Prior Sets G and H) on the precision to result in lower central estimates for the LOR values, due to the increased right skewness that this distribution enforces on the standard deviation density. When the dataset includes few events, it may be more vulnerable to the influences of the prior, hence choice of prior is very important in these cases, and a sensitivity analysis to assess the effects of multiple priors is very important.

10.5.2 Hierarchical models and constraints

The hierarchical models indicated that the effects of 'borrowed strength' across different treatments within a sub-category, such as drugs within dose or vice versa, can also bring results that differ statistically and may also lead to differences in clinical conclusions.

The phrase 'borrowing strength' implies that it is in some way a positive thing for different but related subgroups to be analysed in such a way that they may influence each other, and reduce the width of associated plausible intervals. In a scenario where sparsity of data is not an issue this may be the case, but when there are few events across a dataset, it may be possible that 'strength' is not so much 'borrowed' from one subgroup to the next, but that 'strength' is inflicted between subgroups in a way that may not always be conducive to valid results.

For example, in the MTC performed above, when an hierarchy with dose assumed interchangeable within individual anti-TNF, 'strength' was inflicted on the recommended dose of infliximab from the high dose, producing a considerable difference in median LOR (-0.127 in the non-hierarchical model to 0.746 in the hierarchical model). Whilst CrIs were wide in both cases, the use of the hierarchical model narrowed the CrI somewhat. If however, the order of the hierarchy was flawed, in that dose is in fact the most important factor in malignancy risk (whereas this model places individual anti-TNF as the most important factor, above dose in the hierarchy), this hierarchical model may in fact be misrepresenting the recommended dose of infliximab as having a greater association with malignancy than is actually the case.

The ordering of hierarchical models is therefore of great importance and may be more influential when there are few events and/or few studies in the dataset, re-

sulting in wide intervals and central estimates that may be more easily influenced by borrowed strength across subgroups.

The work of Walsh & Mengersen (2007) suggests that the hierarchy should be determined empirically using a rule that the highest level of the hierarchy should be that with the greatest degree of variability between its groups, and then with degree of variability within each group becoming successively smaller as the levels of the hierarchy descend. In practice, this should result in narrower confidence or credible intervals for each measurement, if the hierarchy follows this rule.

In a simulation example used by Walsh & Mengersen (2007) the mean value for each observation was centred in the same place for either of two hierarchical models, whereas the width of the confidence interval (CI) was consistently narrower for one of the two models (using a quarter of the CI to be analagous to the standard error of the mean). Using a clinical example, one hierarchical model provided a narrower CI for the overall mean estimate of the parameter of interest, although CIs for each measurement were not greatly different between the two hierarchies.

Based on the work of Walsh & Mengersen (2007), there was no clearly preferable hierarchical model in terms of whether dose or drug should be the higher of the two levels. Considering the width of the CrIs as set out in Tables 10.2 and 10.3, there was no strong pattern regarding CrI width between the two models. However, the overall sparsity of events across the 13 studies may have precluded the emergence of a clear hierarchical structure due to dilution of the data. Therefore, based on these results it is difficult to determine which hierarchical model may yield the more valid results, or indeed if a hierarchical structure is in any way preferable to a non-hierarchical model including dose and drug in a non-connected way. The benefit of narrower CrIs may be offset by the disadvantages of model uncertainty.

Other aspects that may influence the hierarchy of a model are theoretical issues, considerations of study design, and physical arguments (Walsh & Mengersen, 2007). These areas are not pursued further by these authors, but it is assumed that theoretical issues relate to prior knowledge regarding the way the different levels of the hierarchy inter-relate, design considerations refer to constructing the modelling hierarchy in accordance with the way a study has been designed,

and physical arguments may indicate that there are definite physical restrictions that impose themselves on the way in which a hierarchy can be constructed.

With regard to adverse events, hierarchical models have the potential to play a very important role. For example, there may be hierarchies related to treatment, such as drug, dose and duration of treatment. Other hierarchies may relate to indication for treatment, or demographic factors such as age and sex. Considering ways to construct such hierarchies to make most valid use of the available data could be a very valuable area for future study within the field of adverse events, and indeed meta-analyses of other types of primary data.

One way in which an hierarchical model may be developed further, by using prior clinical beliefs regarding the nature of the hierarchy between drug and dose, is to add in constraints to the model. This method worked well with this example, as the three dosage levels lent themselves to an evident clinical assumption, namely that higher doses, of all three anti-TNFs, would result in increased risk of malignancy. This is a very strong prior belief to place on a model parameter, but it can be justified as reasonable in this instance.

The results were considerably different from those of the comparable model without the constraint. Not only were the 'best' and 'worst' values constricted to be in the expected order of size for each drug, but the log ORs were also very different, and gave more clear-cut results. For example, in the constraints model the log OR for the highest dose of adalimumab was 1.930 (95% CrI 0.327; 3.774), compared to 1.154 (95% CrI -0.076; 2.806). Hence, the result has become 'significant' with the use of the constraints model. As can be seen in Table 10.5, the model without constraints indicated a higher median LOR for malignancy in the low dose than the recommended dose. The constraints model however, enforced the expected ordering of LOR increasing accordingly with anti-TNF dose. The constraints model also enforced the expected hierarchy of results onto the probabilities for 'best' and 'worst', an effect that was not seen consistently in the hierarchical non-constrained model.

This result adds support to the argument that the high dose drugs are associated with the highest risk of malignancy, whereas there is no strong evidence to support increased risk for low or recommended doses. One caveat however, is that the sparsity of events across the dataset may indicate that the data in this example are more easily overwhelmed by the prior than would be the case in

a dataset with a larger number of events and hence a stronger input into the posterior model. With a dominant prior, there may be a risk of imposing a particular result that the data do not in fact support.

10.5.3 Final conclusions

The use of hierarchical models is arguably an appropriate one in the circumstances of this clinical situation, in terms of allowing 'borrowing strength' across treatments with equivalent parameters. The estimates of treatment effect from these models may be considered preferable to those from the non-hierarchical models.

It is, however, in the use of constraints, that such models appear to make the greatest contribution to the analysis. By allowing a strong prior viewpoint to be an influential aspect of the model, in that the low doses (across all anti-TNFs) have a weaker influence on malignancy risk than high doses, and high doses have a stronger influence, the results are more emphatic, and provide a clear response to the question (Table 10.5). With the sparsity of events across this dataset, the option of adding constraints to the model is appropriate, both statistically and clinically. It is statistically appropriate in terms of making best use of the data, by applying not only 'borrowing strength' within the dataset, but also placing what is effectively a prior distribution on parameters within the model, to ensure their relative magnitude. From a clinical perspective, it is reasonable to assume a dose–response effect, based on previous analyses (Bongartz *et al*. 2006 and Leombruno *et al*. 2008, see Section 9.8.6).

The only caveat of using hierarchical models with constraints is that the constraint may dominate the data, especially where events or primary studies are scarce. However appropriate the constraint may seem, there is concern that its effect will exert undue influence on the results. It is clear from Table 10.5 that the use of constraints has significantly altered the results, in terms of both the point estimates and uncertainty estimates.

Further investigation into the use of constraints within a hierarchical model, with the aim of discovering how strongly they can influence the outcome of a model, with varying amounts of data, differences in construction of MTC network, and strength of treatment effect, would be an interesting field of endeavour.

Simulation studies to incorporate these factors would be an appropriate way to evaluate the use of constraints within a hierarchical model framework.

As can be seen in Section 10.4.4, the choice of prior distribution, even when supposedly 'vague', can be influential in the model. In these analyses, the sensitivity analysis across prior distributions was applied only to selected non-hierarchical MTC models. It would be interesting to extend this work by a sensitivity analysis across the hierarchical models; as a function of these models, uncertainty parameters can be propagated through the model (Section 10.2), thus a prior distribution may have increased influence.

Within the context of a harm–benefit model, as described in the next chapter, an MTC model with a hierarchical structure (with or without constraints) could be used to inform the model, for the adverse events, and if appropriate for beneficial effects also.

10.6 Summary

This chapter extends the MTC models of Chapter 9 by applying a hierarchical modelling structure on certain models, and then by applying constraints to selected hierarchical models, whilst using the same dataset. These methods are novel approaches to adverse events data, and are particularly appropriate to the clinical questions posed by this dataset. These models support the best use of a dataset with sparse events, where it is difficult to derive a clear signal, due to lack of power. In this context, estimates of treatment effects, and ranking of treatments, become more important in understanding the clinical picture.

A sensitivity analysis of prior distributions is also performed, using standard MTC models as described in Chapter 9.

Potential ways to extend this model would include simulation modelling to evaluate the effects of placing a constraint on a hierarchical model, and including a range of prior distributions in the hierarchical model for a sensitivity analysis.

10.7 Densities for standard deviation of alternative priors

The posterior densities on standard deviation for Models 2b and 4b, across Prior Sets A–G are shown below in Figures 10.1 and 10.2.

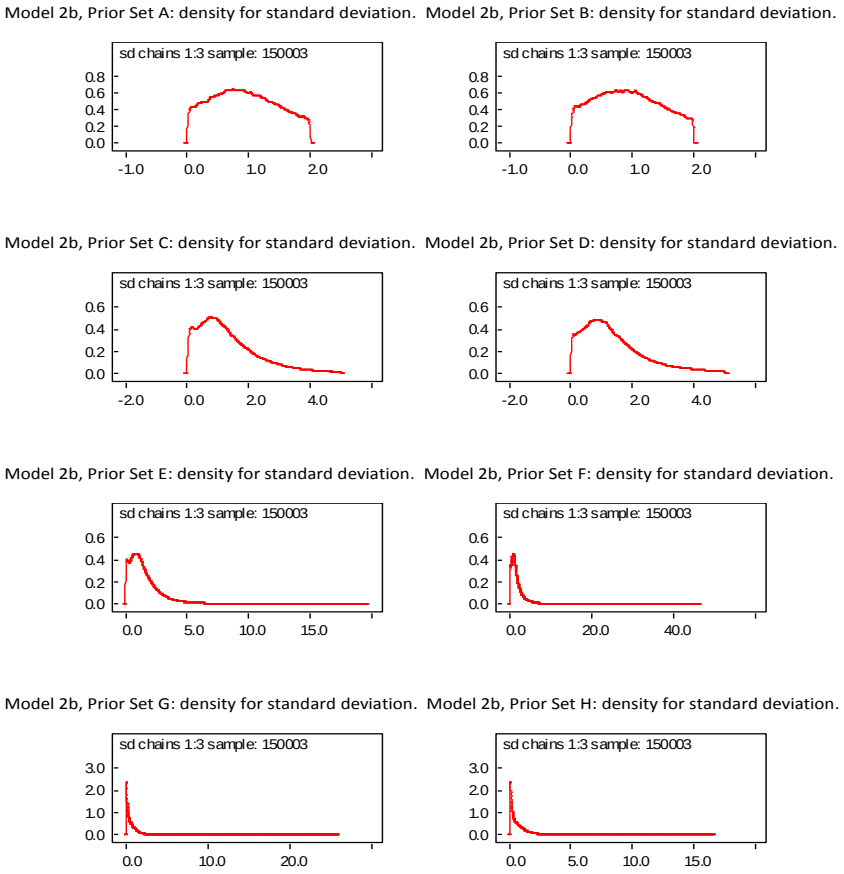


Figure 10.1: Densities for standard deviation of alternative priors for Model 2b.

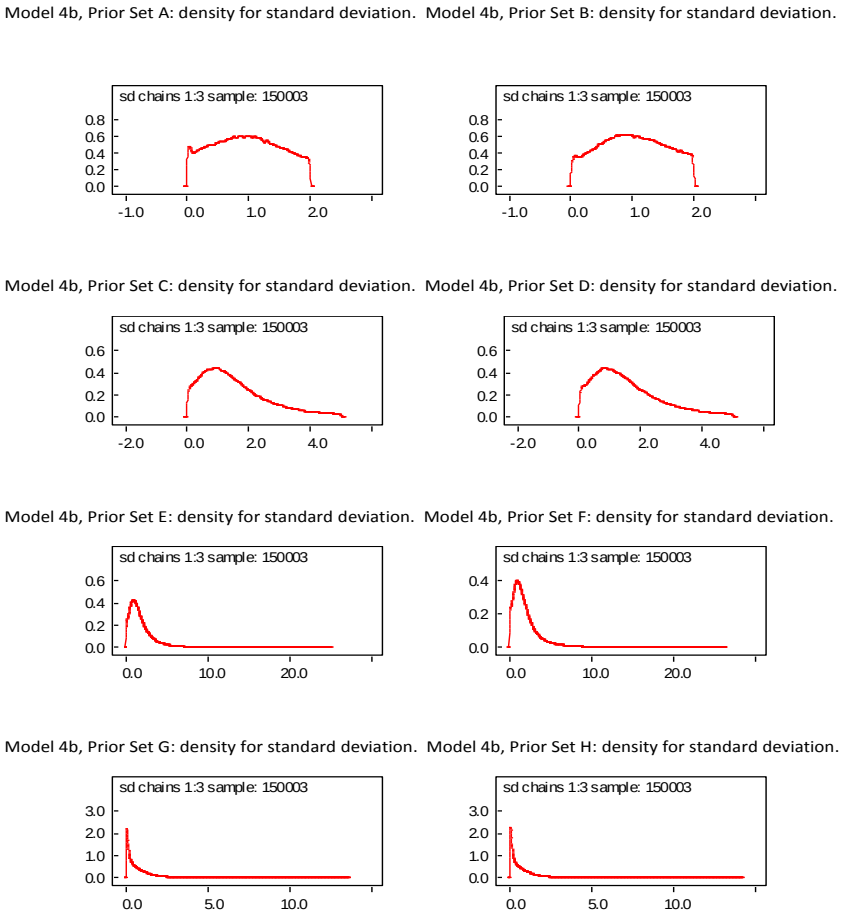


Figure 10.2: Densities for standard deviation of alternative priors for Model 4b.

11

Net clinical benefit models: case-study using tamoxifen for prevention of breast cancer recurrence

11.1 Introduction

There are instances where a drug (or other intervention) may be deemed so harmful to any individual who may use it that it is justified to prohibit its use altogether, regardless of any possible beneficial effects it may have for some patients. For example, ximelagatran, a thrombin inhibitor with efficacy in stroke prevention, was withdrawn due to hepatotoxicity (Agnelli *et al*. 2009). In many cases, however, a blanket ban on a drug is not warranted, because any potential adverse events are not sufficiently common or severe to justify the loss of beneficial effects.

In less extreme cases it may be possible to exclude a certain drug from general use due to the availability of other drugs with similar efficacy but a superior safety profile.

It is in such situations that it is important to quantify potential benefits and harms, possibly varying across different patient characteristics, to allow clinical decision-making at an individual level. Furthermore, it is rarely valuable to consider only adverse events associated with any particular intervention, as the

aim of the intervention is to create a positive outcome, and only once this is established can potential adverse effects be considered that may offset this outcome.

In this section, the example of tamoxifen has been used, as it has been an established treatment to prevent recurrence of breast cancer, but is also associated with endometrial cancer and other possible harms. Net-benefit models were developed that bring together a variety of evidence on the effects of tamoxifen, translated into the 'common currency' of quality of life (QoL), thus converting into one metric the analyses of benefits and harms.

11.2 Overview of net-benefit models and quality of life measurement

11.2.1 Net-benefit models for medical decision-making

An early approach to net benefit modelling was provided by Glasziou & Irwig (1995). These authors were interested in the question of applying treatments to particular patient subgroups. They rejected the method of considering the entry criteria for the initial clinical trials, and then determining whether an individual patient would be sufficiently similar to these criteria. They put forward the idea of assessing, for an individual patient, the potential benefits and harms of a treatment. Their argument was that for most patients, potential harms will be fixed (although this may not be the case for every treatment), whereas patients will vary widely in their potential for deriving a positive benefit (people with the greatest risk or greatest severity of disease would stand to gain most benefit).

Net benefit would be assessed by a meta-analysis of randomised trials, as argued by Glasziou & Irwig (1995), whereas risk to individual patients would be derived from multivariate analyses of cohort study data. However, the major assumptions of fixed adverse events and constant reduction in relative risk (beneficial effect of the treatment across all patients) would need verification. Citing earlier work by Lubsen & Tijssen (1989), Glasziou & Irwig (1995) reported that benefits and harms can be analysed separately, with the underlying assumption that at some point, for the patients with the least to gain from treatment, the

potential for harm will balance the potential for benefit. The mainstay of the model is to convert relative benefits into absolute benefits, which are likely to vary according to the initial level of patient risk. By incorporating a fixed harm, a net-benefit equation is derived:

$$\text{Net benefit} = (\text{risk} \times \text{reduction in relative risk}) - \text{harm}. \qquad (11.1)$$

A four-stage model is then discussed by Glasziou & Irwig (1995). Stage 1 involves estimation of benefits and harms, preferably by use of a randomised trial or meta-analysis of such trials. Stage 2 involves checking assumptions of relative benefit and absolute harm. Relative risk reduction may vary with underlying risk (possibly for many clinical reasons), also absolute harm may be dependent on risk. Investigation into these assumptions may be performed by plotting risk reductions (or increases, as appropriate) against underlying risk [of disease being treated], for both harms and benefits. If an intervention has both positive and negative effects on the same outcome, then both of these influences would need to be modelled simultaneously.

Stage 3 in this model is to balance out the benefits and harms. In their model, Glasziou & Irwig (1995) consider quality of life (QoL) issues to develop an equivalence between deaths caused by intracranial haemorrhages and thromboembolic strokes prevented. Whilst their models used average values for QoL, the authors make the valid point that different individuals will place varying values on QoL following different medical events and conditions. The final stage (Stage 4) in their model returns to the concept of predicting risk at the level of the individual patient, by identifying specific risk factors, and using these to estimate risk across all risk factors. The authors suggest that population-based cohorts may be preferable to trials in eliciting information regarding risk factors, due to the processes of a trial in determining eligibility and obtaining consent, which may have an influence on risk factors for adverse events.

The model used by Glasziou & Irwig (1995) was a quantitative model to evaluate clinical benefit, but did not include Bayesian methods. A re-evaluation of the examples used by Glasziou & Irwig (1995) was performed by Sutton *et al*. (2005), using Bayesian methods, the primary benefit of which is that they can add in uncertainty around individual parameters in the model. To illustrate where uncertainty can be brought into the model, Glasziou & Irwig (1995) admit that,

when using QoL values, individual patients will vary in the QoL ratings for certain conditions, but such variations are not incorporated into the model.

By placing uncertainty around average estimates, the uncertainty can be incorporated into the overall model, producing a result that can be applied over all patients receiving a treatment. Whilst it would be ideal to perform an individual analysis for each patient, incorporating that patient's specific values for QoL and individual risk factors, this is not always feasible, and a model incorporating uncertainty is preferable to one that uses fixed values that will not be applicable to all patients.

Using the terminology of *net clinical benefit* (NCB) as the evaluation of balance of risks, Sutton *et al.* (2005) put forward a simple equation to calculate this metric:

$$\begin{aligned} \text{Expected NCB} &= \text{Expected benefit from treatment} \\ &\quad -\text{Expected harm from treatment.} \end{aligned} \tag{11.2}$$

When there are multiple benefits and/or harms, these can be added together across the relevant clinical outcomes, as in Equation 11.3.

$$\begin{aligned} \text{Expected NCB} &= \Sigma(\text{Expected benefits from treatment}) \\ &\quad -\Sigma(\text{Expected harms from treatment}). \end{aligned} \tag{11.3}$$

A further issue is that risk of harms and potential for benefit may not be constant across all patients but may vary according to patient characteristics, for example, patients with the greatest propensity for disease may stand to gain the most benefit from treatment. In many meta-analyses, the outcome metric is on a ratio scale, such as an odds ratio (OR) or relative risk (RR). These relative outcome metrics are often considered constant across varying levels of patient risk. Hence, to evaluate a net clinical benefit for a patient at a given level of risk, it is necessary to convert the RR (for benefit) onto an absolute scale, such a reduction in relative risk (RRR, which is equal to 1 – RR). (It is assumed that the RR will be less than 1 for a beneficial treatment. Relative risk increase (RRI), equal to RR – 1, would be used in cases where the RR was greater than 1.) The RRR can then be multiplied against the patient's risk (of the disease that is being treated) for each individual patient.

This then gives the following formula for NCB:

$$\text{Expected NCB} = (\text{Patient risk} \times \text{RRR}) - \text{Expected harm from treatment.} \tag{11.4}$$

If benefit equals harms then NCB is equal to zero; if benefit exceeds harms then NCB is positive and if harms exceed benefit then NCB is negative and the treatment would be regarded as inadvisable for a patient with the level of risk modelled in the formula. Assuming relative risk remains constant across all patient factors, then patient risk has a linear relationship with absolute benefit, and assuming expected harms from treatment also remain constant for all patients, then patient risk also has a linear relationship with NCB. If the assumption of constant RR for benefit is invalid, then the linear relationship between NCB and patient risk would not occur. RCT data and meta-analyses of such data can be used to establish the relative risk for benefits, and can also be used to identify any patient characteristics influencing RR by use of subgroup analysis and meta-regression. Once risk factors have been identified, the ways in which these risk factors may influence relative risk for individual patients can be used in modelling NCB.

Re-analysing the data used by Glasziou & Irwig (1995), Sutton *et al*. (2005) used Bayesian methodology, which allows each parameter in the model to be represented by a probability distribution with its own degree of uncertainty. This allows more flexibility in applying the results, which are expressed as credible intervals (CrIs).

The authors set out to evaluate the quantity they refer to as 'net clinical benefit', describing the difference between expected treatment benefit and expected treatment harm. In their example of evaluating the benefit of stroke prevention with warfarin therapy against the risk of fatal intracranial hemorrhage, the authors stress the importance of putting both benefits and harms on the same scale. Bayesian methods were used to combine the preventative effects of warfarin on stroke, hemorrhage, a quality of life outcome and risk of stroke in different subgroups in a multiparameter evidence synthesis model. The net clinical benefit can then be synthesised as a posterior distribution having incorporated data on risks and benefits, for different subgroups, based on level of risk.

As both of the main outcomes of this model involve a risk of death, it is straightforward to see how the risks and benefits can be 'traded off' against each other. In a scenario where the risks and benefits are not so diametrically opposed, some framework in terms of quality of life for risks and benefits would need to be developed for each intervention. There may be several adverse events for each intervention (and possibly more than one treatment effect although this is likely to be a less commonplace situation), which should be especially borne in mind.

The necessary data to inform the model can either be derived from single studies or from multiple studies using meta-analysis methods within the NCB model, another benefit of using a Bayesian approach. A QoL model can then be used to evaluate harms and benefits on the same scale, ensuring that patients who die (and hence have a QoL of zero following death) are included in the QoL calculations. The importance of sensitivity analyses to assess the effect of the multiple assumptions that are of necessity included in such a complex model was also highlighted.

A similar model balancing harms and benefits of hormone replacement therapy (HRT), which has multiple potential benefits and harms, was carried out by Minelli *et al*. (2004). Using quality-adjusted life-years (QALYs) to translate benefits and harms onto a common scale, these authors developed a decision model based on that of Glasziou & Irwig (1995), with the aim of identifying a level of baseline risk of breast cancer (the most significant potential harm) in women using HRT. Above this level of baseline risk of breast cancer, the potential benefits of HRT would be outweighed by risk of breast cancer (and other harms including coronary heart disease, pulmonary embolism and stroke). Hence, the same model was evaluated using multiple baseline risks. Again, this model relied heavily on the validity of multiple assumptions across the different aspects of the model. The first assumption was that the QoL values for risks and benefits were indeed interchangeable and could be added and subtracted for the various harms and benefits. (This assumption highlights the subjective nature of QoL as a tool for decision modelling, which will be discussed further in Section 11.2.2).

11.2.2 Use of quality of life in medical evaluations

The conversion of health benefits conferred by medical interventions into economic evaluations allows these interventions to be evaluated for cost-benefits, to determine whether an intervention can be administered cost-effectively to a population or to compare different interventions for the same condition in terms of cost-effectiveness. A cost–utility analysis places the unit of measurement of benefit (or, indeed, disbenefit) to the individual patient in terms of QALYs. As discussed by Torrance (1986), the QALY is useful when there are multiple outcomes, outcomes that include both morbidity and mortality, and when a treatment is to be compared to other treatments that have already been assessed using QALYs.

There are several ways that quality of life can be assessed quantitatively, as described by Torrance (1986). The simplest involves a rating scale, whereby the individual allocates a rating to each health state, for example between 0 (least preferred health state, for example death) and 1 (most strongly preferred health state, or perfect health). The value derived is effectively a weighting for the quality of health in a certain state, also known as a utility. A visual analogue scale may be used, which was the preferred method for the EuroQoL, whereby participants rated health states on a scale of 0 to 100 (The EuroQoL Group 1990).

A slightly more complex method is the standard gamble, in which patients are offered a choice of outcomes following a certain treatment, with associated probabilities for each outcome, for example, being healthy (with probability p) or death (with probability $p\text{-}1$). These probabilities are compared to a different alternative (for example, no treatment) which has a certain outcome, for example a chronic state of poor health. The probabilities following the treatment can be varied until there is no preference for either alternative treatment, at which point the preference (utility) for the chronic health state compared to being healthy is taken as p.

A third method is that of the time trade-off, whereby the respondant is asked to determine how much time in a chronic health state would be the equivalent to a lesser time in full health, from which a utility can be derived. Other methods exist, for example, equivalence techniques and ratio scaling (discussed

by Torrance (1986), citing previous authors) but the three methods described above are among the most prominent.

In evaluating the validity of methods of health state evaluation, Torrance (1986) argues that these methods are valid, as long as the participants are appropriate, the health state descriptions are adequate, the questions are framed in a balanced way, and that the measurement techniques are valid and reliable. The participants may be selected in a variety of ways, for example, members of the public, patients with specific health conditions, or healthcare professionals. Hence, when using utilities for different conditions, it should be borne in mind that there may have been differences in the populations used to determine the utility values, and that it may not be valid to directly compare utilities across multiple conditions.

11.3 Clinical context: tamoxifen for recurrence of breast cancer

Tamoxifen has both anti-estrogenic and estrogenic effects, and it is the anti-estrogen effects that make it useful in breast cancer sufferers, with the aim of reducing recurrence of the disease. However, there are many adverse events associated with tamoxifen, some of which are themselves life-threatening, for example, increased risk of endometrial cancer, pulmonary embolism and cerebrovascular accident.

In the light of concerns regarding the adverse effects of tamoxifen, Braithwaite *et al*. (2003) performed a meta-analysis of vascular and neoplastic outcomes using randomised controlled trials where tamoxifen was compared to a non-tamoxifen control group. In the majority of their primary studies, the indication for taking tamoxifen was the prevention of recurrence of breast cancer in patients who already had the disease, but there were some primary studies where the indication for use of tamoxifen was prevention of breast cancer in women in high risk groups, as well as some trials that were unrelated to breast cancer. Based on their results, tamoxifen was associated with a statistically significant increase in risk of stroke, deep vein thrombosis, gastrointestinal cancers, and endometrial cancers. However, there was a significant reduction in deaths due to myocardial infarction.

For a clinician or patient with breast cancer attempting to determine whether tamoxifen treatment would be beneficial, this meta-analysis addresses only one side of the issue – that surrounding adverse events, without attempting to quantify them against the beneficial effects of tamoxifen in reducing breast cancer recurrence. The modelling outlined in this chapter aims to combine the data regarding tamoxifen as a beneficial agent of reducing breast cancer recurrence, against one of the adverse events most strongly associated with tamoxifen use: endometrial cancer. In this way, a quantitative analysis can be made of whether tamoxifen is advisable for women with breast cancer recurrence, with regard to the risk of endometrial cancer.

11.4 Methods 1: Dataset creation

11.4.1 Required data

This example involves an assessment of the increased QoL for the reduced recurrence of breast cancer, which is balanced against the loss of QoL due to increased risk of endometrial cancer. Hence, two sets of data are required, one for each section of the model.

The items of data required for the recurrence of breast cancer aspect are set out below.

1. Relative risk (RR) of breast cancer recurrence for tamoxifen users compared to non-tamoxifen users.
2. Average risk of breast cancer recurrence.
3. Average risk of death after breast cancer recurrence.
4. QoL during breast cancer recurrence.

The items of data required for the risk of endometrial cancer are shown below.

1. Relative risk of endometrial cancer for tamoxifen users compared to non-tamoxifen users.
2. Average relative risk of endometrial cancer in general population and associated precision.

3. Average risk of death due to endometrial cancer.
4. QoL during endometrial cancer.

With the elements of data above, it is possible to develop a model to evaluate the changes in QoL due to risk of breast cancer recurrence and endometrial cancer, for tamoxifen users compared to non-tamoxifen users, as well as modelling the associated uncertainty around the different values as they inform the model.

11.4.2 Data sources and extraction

Breast cancer recurrence

The relative risk of breast cancer recurrence for tamoxifen users compared to non-tamoxifen users was derived from a study by the Early Breast Cancer Trialists' Collaborative Group (EBCTCG) from 1998; the data selected for use related to patients who had used tamoxifen for an average of 3 years or more, with a median of 5 years. Nine trials were available, with data regarding numbers of patients and events for tamoxifen and non-tamoxifen arms.

The underlying average risk of breast cancer recurrence (for non-tamoxifen patients) was also determined within the model, by use of the EBCTCG (1998) data. Using the model, the RR for breast cancer was determined, as well as the RRR.

Death due to breast cancer recurrence

The next stage in the model was to include the risk of death due to breast cancer recurrence in those patients who did experience a recurrence of disease. Data regarding risk of deaths due to breast cancer were derived from Schairer *et al*. (2004); these authors provide data regarding cause of death in patients with different stages of disease, and subdivided by ethnic group and age group. Data in this example were selected for localised disease and combined data for both white and black patients. In the primary model, the age range 50–55 was the target age, which was most closely approximated by the 50–59 year age group in Schairer *et al*. (2004). For an older age range, data for the '70 years and older' age range was used, again combining ethnic groups who were diagnosed with localised disease.

The numbers of deaths due to breast cancer were then modelled using a binomial distribution, which then allowed a beta prior distribution (with values restricted between 0 and 1) to be placed on the probability of the binomial distribution.

$$r.br \sim \text{Binomial}(p.br, n.br), \tag{11.5}$$

and

$$p.br \sim \beta(1, 1), \tag{11.6}$$

where *p.br* is the probability of dying from breast cancer recurrence (and is equal to the value of *specific.deaths.br*), *r.br* is the actual number of deaths due to breast cancer recurrence as derived from the appropriate data, and *n.br* is the actual number of participants in each arm of the trial from which the data were derived.

With the data described above, it was then possible to model the number of deaths prevented by tamoxifen, by multiplying the RRR for breast cancer for tamoxifen users compared to non-tamoxifen by the average risk of breast cancer and the probability of death due to breast cancer recurrence. Alternatively:

$$\text{dead.br} = \text{RRR.br} \times \text{ave.risk.br} \times \text{specific.deaths.br}, \tag{11.7}$$

where *dead.br* is the number of deaths prevented by use of tamoxifen, RRR.br is the relative risk reduction of breast cancer recurrence for tamoxifen users compared to non-tamoxifen controls, and specific.deaths.br refers to the probability of dying of breast cancer recurrence should this occur.

Quality of life due to breast cancer recurrence

Quality of life during breast cancer recurrence was provided by Tengs & Wallace (2000), using the QoL weighting for breast cancer after surgery, after first recurrence (stated as 0.85) with a standard deviation as quoted for a different but similar condition, breast cancer with breast-conserving therapy, disease-free 1 year later, duration of remaining life (stated as 0.115). From these values, using the 'method of moments' it is possible to calculate the parameters for a beta distribution to be applied to QoL.

Using the method of moments, we have

$$\alpha = \text{mean} \times (((\text{mean} \times (1 - \text{mean})) / \text{variance}) - 1), \quad (11.8)$$

and

$$\beta = 1 - \text{mean} \times (((\text{mean} \times (1 - \text{mean})) / \text{variance}) - 1). \quad (11.9)$$

Risk of endometrial cancer using tamoxifen

Several sources of data were also required to estimate the risk of endometrial cancer with associated risk of death and reduction in quality of life. The earlier meta-analysis by Braithwaite *et al*. (2003) supplied data on risk of endometrial cancer in tamoxifen users with breast cancer, as opposed to non-tamoxifen-users. The meta-analysis by Braithwaite *et al*. (2003) included trials of tamoxifen for two separate indications; treatment of breast cancer in those already having the condition, and breast cancer risk reduction. Data on 23 trials were presented for risk of endometrial cancer, one of which was a breast cancer risk reduction trial and so was excluded (IBIS-1 trial 2002, as mentioned in Figure 4, Braithwaite *et al*. (2003)).

As the meta-analysis by Braithwaite *et al*. (2003) did not include data on the number of cases of endometrial cancer, these data were elicited directly from the authors. For two of the studies there was a discrepancy between the total number of participants as stated in the Braithwaite *et al*. (2003) reference, and the numbers provided by the authors. One of these studies was the South-Swedish Trial, reported by Rutqvist *et al*. (1995). This reference also reported data from the Stockholm Trial and Danish Breast Cancer Group Trial. The correct numbers for each study arm were derived from Rutqvist *et al*. (1995) with regard to the South-Swedish Trial. It was also noted that the numbers for the individual trial arms for the Danish Breast Cancer Group Trial differed between those provided and those stated by Rutqvist *et al*. (1995), and it was decided to use those reported in the primary study.

The other study where there appeared to be contradictory data between the published reference and the data supplied by Braithwaite *et al*. was the Scottish Trial of Adjuvant Tamoxifen, necessitating data to be derived directly from the

primary study (Stewart 1992). A later publication based on the same trial (Scottish Trial of Adjuvant Tamoxifen) was unusable for the purpose of a meta-analysis on endometrial cancer, as the data were presented in such a way as to make it impossible to discern whether cases of endometrial cancer occurred in patients using tamoxifen (Stewart *et al.* 2001).

For another four studies listed in the Braithwaite *et al.* (2003) reference, no data were supplied by the authors, so the primary references were required to provide data. These studies included Klijn *et al.* (2000), reporting on a study for the European Organization for Research and Treatment of Cancer – Breast Cancer Cooperative Group (designated EORTC.b by Braithwaite *et al.*); the report on the Danish Breast Cancer Co-operative Group study known as DBCG 82B, reported by Andersson *et al.* (1999); a report by Fisher *et al.* (2001) on the National Surgical Adjuvant Breast and Bowel Project B-23; and the report by The Arimidex, Tamoxifen Alone or in Combination (ATAC) Trialists' Group (2002).

For most of the trials, there were no data regarding women who had previously undergone a hysterectomy, with the exception of the ATAC Trialists' Group study of 2002, which provided data of women who had not previously had a hysterectomy – these data were used as they provide a more appropriate denominator, although it is recognised that this may be inconsistent with other primary data.

Overall, 22 primary studies were included in the current meta-analysis of endometrial cancer in tamoxifen and non-tamoxifen groups. The data for eight of these were derived directly from the primary studies, or primary studies were used to confirm data provided by Braithwaite *et al.*, whilst the remaining 14 were provided exclusively by the authors of the earlier meta-analysis. It should be mentioned that the control groups varied in their treatments across different primary studies. For some studies, data on the duration of tamoxifen and follow-up time were also available.

Underlying risk of endometrial cancer

There is evidently an underlying risk of developing endometrial cancer in the general population (those who do not have breast cancer) and this needs to be accounted for in the risk–benefit model. Data from Cancer Research UK[1] indicated an incidence rate for women aged 50–54 years of 29.8/100 000, derived from an actual number of cases of 552 in 2005. For women in the age range of 70–74 years and above, the equivalent figures were an incidence rate of 76.5/100 000, based on 965 cases.

Risk of death following endometrial cancer

The negative impact due to endometrial cancer is evaluated in terms of quality of life and risk of death. The latter issue is discussed by Bergman *et al.* (2000), and data can be extracted from this primary case–control study, combining data for tamoxifen users and non-users, informing a binomial distribution and beta distribution for probability of death due to endometrial cancer. These data are not categorised by age.

Quality of life with endometrial cancer

A recent Health Technology Assessment (HTA) report (Hind *et al.* 2007) provided QoL values for a base case [for the HTA report this referred to a scenario where it was assumed that the benefits of tamoxifen subsided over a 5-year period beyond the therapy duration] as well as parameters on a beta distribution for QoL following endometrial cancer, completing all the required data for harms due to this condition.

Data for extensions to the model

Extensions to this model are also discussed below. These extensions focus on concerns regarding the duration of tamoxifen treatment, and the duration of follow-up for participants in the primary studies. Duration of tamoxifen treatment was available for 19 primary studies, 14 of which had the necessary data within the spreadsheet provided by Braithwaite *et al.* (2003). The data

[1] Cancer Research UK (2009). Available at [November 2009]: http://www.cancerresearchuk.org/

for four of these studies was derived directly from the original references, while in the case of one study a discrepancy was noted in the duration of tamoxifen treatment between the provided spreadsheet and the primary study. This discrepancy was noted whilst verifying data on cases and numbers of participants for three studies reported by one reference, where discrepancies had been noted between the figures in the spreadsheet and those reported in the meta-analysis by Braithwaite *et al*. (2003).

Duration of follow-up was available for 20 of the primary studies, with data being derived from the spreadsheet of Braithwaite *et al*. (2003) for 15 of these. For the remaining five studies, the necessary data were derived from a primary reference related to the study. For the majority of studies, the duration of follow-up was expressed as a median, in four cases as a mean, and once as a uniform duration. (No attempt has been made to treat these three metrics differently in the analysis.)

Dataset for endometrial cancer occurrence

The full dataset for endometrial cancer occurrence used in this harms and benefits model is set out below in Table 11.1.

A meta-analysis was performed using this dataset, with a Mantel–Haenszel (M–H) model, using the relative risk outcome metric, with a standard 0.05 continuity correction (See Chapters 3 and 5 for further details of these aspects of meta-analysis). This meta-analysis yielded a pooled OR of 3.04 (95% 2.13; 4.34); the associated forest plot is shown in Figure 11.1.

11.5 Methods 2: Modelling methods for net-benefit model

11.5.1 Modelling benefits with regard to breast cancer recurrence

The model used allowed for benefits and harms to be amalgamated in terms of QoL, and then the difference between the two to be determined. The EBCTCG (1998) study provided information on the RR of breast cancer recurrence between tamoxifen and non-tamoxifen users, as well as the innate risk for non-tamoxifen users, leading to an evaluation of the number of cases that would be prevented by use of tamoxifen. Non-informative priors are placed on the underlying log RR of breast cancer recurrence in tamoxifen users compared

to non-users and on the mean logit for recurrence in the non-tamoxifen patients, as well as on associated standard deviations.

A proportion of the cases of recurrence would then lead to death, which is incorporated into the model by the use of the data from Schairer *et al.* (2004), informing a binomial distribution for deaths relating to breast cancer recurrence.

In this model, the potential benefits from tamoxifen occur in relation to two types of patients. Firstly, there are those who would have had recurrence of their breast cancer, but not died (within a 5-year period) as a result of the recurrence, and secondly, there are those who would have had a recurrence and would have died within 5 years in consequence.

The increases in QoL (compared to those patients who do have a recurrence of breast cancer) for these two patient types can be summed to produce an overall benefit for tamoxifen with regard to reduction in breast cancer recurrence. For this model, the benefits and harms are evaluated over a 5-year period. An assumption is made at this point that patients who would have died due to their recurrence had a mean lifespan of 2.5 years, out of the total five-year period of interest.

For the patients who would have had a recurrence and died within 5 years, they receive the 2.5 years additional life at reduced QoL (5-2.5 years), plus the additional life they would have received after the 5-year period, during which time a default assumption is made that the QoL reverts to 1. This aspect of the model is accounted for when considering the total deaths prevented. Similarly, those patients who would have had a recurrence but not died within the 5-year period receive their 5 years of reduced QoL, plus the rest of life with a QoL equal to 1.

Uncertainty around quality of life for those patients who would benefit from tamoxifen was modelled using a beta distribution with parameters derived from Tengs & Wallace (2000); (Section 11.4.2).

11.5.2 Modelling harms with regard to endometrial cancer

Turning attention to the harms due to increased risk of endometrial cancer, the meta-analysis by Braithwaite *et al.* (2003) informs the logit of the probability of endometrial cancer in tamoxifen users compared to non-tamoxifen users, at the

individual study level. From this point, the relative risk of endometrial cancer in tamoxifen users compared to non-tamoxifen users, at the individual study level, can be estimated. The underlying mean relative risk is then estimated, using non-informative priors for the study-level risk of endometrial cancer in the baseline group, the underlying mean relative risk and the standard deviation for the study-level relative risk of endometrial cancer.

This allows the calculation of the relative risk increase of endometrial cancer for breast cancer patients using tamoxifen, compared to those who do not use tamoxifen. The next stage is to infer the underlying risk of developing endometrial cancer in the general population, over a 5-year period; this is done using the data provided by Cancer Research UK. The actual number of cases derived was 552, from which value the precision of the underlying incidence of endometrial cancer can be derived (where standard error of an incidence is the reciprocal of the square root of the number of cases; Clayton & Hills 1993). Using this data, the underlying risk level of endometrial cancer over 5 years can be estimated.

The number of additional cases of endometrial cancer occurring in the population using tamoxifen can then be estimated, leading to the QoL assessments for patients who die within the 5-year period and those who survive the 5 years with endometrial cancer. In this phase of the model, QoL data from Hind (2007) are used, as well as data on risk of death due to endometrial cancer, derived from Bergman *et al*. (2000).

At this stage of the model, the total harm due to risk of endometrial cancer can be estimated based on the number of deaths that result due to tamoxifen use, and the reduced QoL of those patients who develop endometrial cancer, but do not die within a 5-year period. Thus, quantitative evaluations of both harm and benefit relating to QoL over a 5-year period, have been estimated.

Table 11.1: Dataset for tamoxifen and endometrial cancer risk. Data are derived from spreadsheet provided by Braithwaite *et al.* unless marked by *, in which case they are derived from a primary reference.

Study name	Year(s)	No. cases in tamoxifen group	Total no. in tamoxifen group	No. cases in non-tamoxifen group	Total no. in non-tamoxifen group	Tamoxifen duration (months)	Follow-up duration (months)
Ingle	1988	0	71	0	75	15	60
NCCTG	1989	0	198	0	202	15	63
GROCTA	1992	1	171	0	165	60	60
Scot	1992	1	374	1	373	60	47
Christie	1992	1	481	1	480	12	Unknown
ECOG1178	1993	1	85	1	83	24	120
Wisc	1994	0	70	0	70	60	60
SWOG	1994	3	303	0	300	12	78
Barner	1994	0	86	0	81	82	75
Stockholm	1995	23	1372	4	1357	24*	108
S.Swed	1995	4*	239*	2*	236*	12	108
EORTC.E	1995	0	37	0	37	Unknown	45
DBCG	1995	7*	864*	2*	846*	12	96
Gunderson	1995	0	180	0	170	24	76
B14.96	1996	21	1422	3	1437	Unknown	Unknown
BrPr	1998	4	1238	1	1233	70	70
P1	1998	37	6576	18	6599	47.7	54.5
DBCG82b	1999	1*	320*	0*	314*	12	146.4
B24	1999	7	899	2	899	60	74
EORTCb	2000	0*	53*	0*	54*	Unknown	87.6*
B23	2001	5*	990*	1*	992*	60*	65*
ATAC	2002	6*	2240*	3*	2228*	60*	33.3*

In order to more fully account for the effect of tamoxifen treatment on deaths, the influence of this treatment on death in the longer term can be evaluated. Given that most patients will be in the age range of 40 and upwards, the majority will die within a period of 50 years. Assuming that QoL will be reduced by approximately 3% per year, deaths that are prevented by tamoxifen will accrue QoL, but the amount of QoL will reduce on a year-by-year basis. This total additional QoL (assuming that tamoxifen has a positive effect on death, by preventing more deaths due to breast cancer recurrence than are caused due to endometrial cancer) accrued can then be added into the overall model.

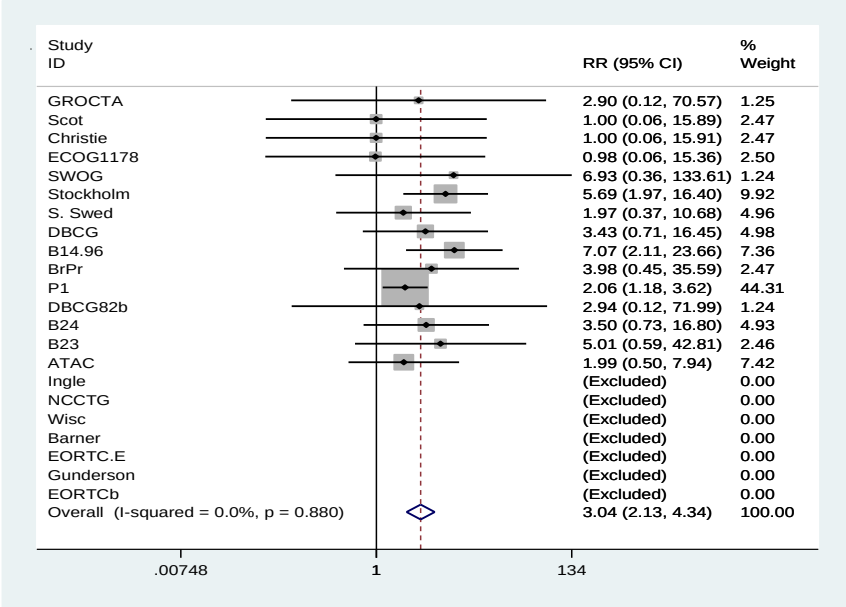
11.5.3 Modelling net benefit

Finally, the net benefit due to tamoxifen, based on the harms due to endometrial cancer being subtracted from the overall benefit due to reduced breast cancer recurrence, and adding in the QoL accrued due to deaths prevented by tamoxifen, can be evaluated.

The harm and benefit estimates in the model are calculated using the number of cases of disease (breast cancer recurrence or endometrial cancer) prevented or caused by tamoxifen, based on the RRR or RRI, and the average population risk. The model considers whether those cases would or would not have resulted in death within five years. The QoL gained or lost associated with each condition is aggregated over the total cases for each disease condition. For those patients whose death is prevented by tamoxifen (within the 5-year period under evaluation) their QoL for the next 50 years of life is added to the model, with the QoL being discounted for increasing loss of QoL due to age as well as the reduced value of QoL experienced in the future compared to QoL in the present.

Hence, the net benefit is a numerical concept relating QoL (gain or loss), duration of QoL and the number of individuals experiencing the changes in QoL for each treatment group. The net benefit is therefore the aggregated difference in QoL comparing the tamoxifen and non-tamoxifen groups on a difference scale. A positive net benefit indicates that there is an overall gain in QoL for the tamoxifen group compared to the non-tamoxifen group. To conceptualise this clinically, at the level of the individual, if the overall net benefit is 0.7, an 'average' woman in this population would gain one additional year of life with a QoL of 0.7 (or the equivalent), when using tamoxifen, compared to not using

Figure 11.1: Forest plot of a meta-analysis using the dataset described in Table 11.1, using a Mantel–Haenszel model, 0.05 continuity correction (studies with zero events in one arm only).



See Section 11.4.2, Table 11.1 and Braithwaite *et al*. 2003 for further details of the primary studies.

tamoxifen. A negative net benefit would indicate that there is a loss in QoL over time for the patients using tamoxifen.

In all models, non-informative priors were used where required on all stochastic nodes.

The separate benefits and harms of the model are set out in Table 11.2. The benefits and harms can then be aggregated and the difference between them evaluated, to determine whether there is an overall harm or benefit due to tamoxifen, in terms of QoL.

11.5.4 Extensions to the model

This model can be extended by bringing in duration of tamoxifen treatment, and duration of follow-up, which were available for the meta-analysis of tamoxifen and occurrence of endometrial cancer by Braithwaite *et al*. (2003). Inclusion of these variables allows the investigation of whether or not they have any bearing on tamoxifen safety with regard to endometrial cancer. For those studies where duration of tamoxifen treatment or duration of follow-up is not available, a normal distribution can be placed on the relevant variable, with a vague distribution on its mean, and a uniform distribution on its standard deviation, based on the standard deviations of those studies where the appropriate observation is available.

11.5.5 Sensitivity analyses

The various risk factors impacting on breast cancer recurrence were studied by Brewster *et al*. (2008), in a large study of women who were followed up after primary breast cancer at 5 and 10 years. In their sample of 2838 women, the median age was 50 years, among women ranging from 21 to 87 years. Considering women in three age bands, 35 years or younger, 36 to 59 years and 60 years and older, age alone (in a univariate analysis) had no association with risk of breast cancer recurrence. Given that breast cancer tends to be less aggressive in older patients, it is likely that those in older age brackets, such as 70 years and older, are at less risk of recurrence than younger patients. Hence, different recurrence risks for different age brackets have not been used in this model as a sensitivity analysis.

In the same study (Brewster *et al*. 2008), the pathological factors most strongly associated with increased risk of recurrence were hormone receptor status being negative, and higher grade or more advanced stage of tumour, (as well as receipt of endocrine therapy). Any sensitivity analyses based on these pathological factors would be more advantageous in a clinical model than use of age.

Risk of death due to breast cancer, following recurrence, may vary with different age groups, and in this study, two different age groups were contrasted using data from Schairer *et al*. (2004). The age groups cited in the study reported by these authors were 50–59, which was used to approximate the range 50–54 years, and 70 years or older, which was used for the age range 70–74 years.

For endometrial cancer, there is a major increase in risk with increasing age, following the menopause. Other risk factors relate to nulliparity or low parity, having diabetes or having a high body mass index. Further details regarding risk factors for endometrial cancer are available from Cancer Research UK (Footnote 1, page 316). There is also evidence for genetic risk factors, as endometrial cancer can run in families, for example, the condition hereditary nonpolyposis colon cancer (HNPCC); (Dunlop *et al*. 1997). Any of these risk factors would be of use in a clinical model. However, in this example, sensitivity analyses have been restricted to different age ranges, using the data on endometrial cancer risk provided by Cancer Research UK. The data from Cancer Research UK was presented in 5-year age bands, with the age bands of 50–54, and 70–74, being used.

Table 11.2: Description of harm–benefit model.

Benefits – prevention of breast cancer recurrence			Harms – causation of endometrial cancer		
Element of model	Parameter of model	Quality of life conversion	Element of model	Parameter of model	Quality of life conversion
Cases of breast cancer recurrence PREVENTED that would have resulted in death within 5 years	RRR × average risk × specific deaths	(1–QoL*)× 5	Cases of endometrial cancer CAUSED that would have resulted in death within 5 years	RRI × average risk × specific deaths	(1–QoL*)× 5
Cases of breast cancer recurrence PREVENTED that would NOT have resulted in death within 5 years	(RRR × average risk × specific deaths)–(RRR × average risk)	(1–QoL*)× 2.5	Cases of endometrial cancer CAUSED that would NOT have resulted in death within 5 years	(RRI × average risk × specific deaths)–(RRI × average risk)	(1–QoL*)× 2.5
Net death prevented[†] = deaths due to breast cancer recurrence that were prevented – deaths due to endometrial cancer that were caused (assume QoL = 1 for subsequent life gained/lost for deaths prevented/caused)					
Final model					
Net benefit = total benefit - total harm[‡] + net death prevented					

*A QoL of 1 is assumed for patients post-breast cancer awaiting recurrence; [†]Net death prevented is a benefit if positive and a harm if negative; [‡]Total harm is QoL lost; QoL: quality of life; RRI: relative risk increase; RRR: relative risk reduction.

11.5.6 Summary of net benefit models

A summary of the specific net benefit models analysed is given below.

1. Model A. No covariates included. Age group for risk of death following breast cancer recurrence data and risk of endometrial cancer occurrence: 50–54.
2. Model B. No covariates included. Age group for risk of death following breast cancer recurrence data and risk of endometrial cancer occurrence: 70–74.
3. Model C. Duration of tamoxifen with regard to endometrial cancer included as covariate. Age group for risk of death following breast cancer recurrence data and risk of endometrial cancer occurrence: 50–54.
4. Model D. Duration of follow-up with regard to endometrial cancer included as covariate. Age group for risk of death following breast cancer recurrence data and risk of endometrial cancer occurrence: 50–54.
5. Model E. Duration of tamoxifen and duration of follow-up, both with regard to endometrial cancer, included as covariates. Age group for risk of death following breast cancer recurrence data and risk of endometrial cancer occurrence: 50–54.

These results of these five models are displayed in Table 11.3.

11.5.7 Implementation using WinBUGS

All models were fitted using WinBUGS 1.4, with a burn-in period of 10 000 iterations followed by a further 50 000 iterations. Only one chain was used for each model, with convergence being checked by history trace.

11.6 Results

Results are presented in Table 11.3, where all models are described according to the age range for age-specific data, and any additional covariates (in relation to endometrial cancer) included in the model, as described in Section 11.5.6.

11.7 Discussion

11.7.1 Discussion of results

With the basic model, (Model A), the median net benefit was 0.615 on the QoL scale, and from the 95% CrI (0.359; 1.011), it is very unlikely that tamoxifen in this age category confers any harmful effect, when considering the interaction between benefits due to reduced breast cancer recurrence and harms due to endometrial cancer. The probability of harm, at 0.0005, is also very low, implying that very few women would be unfortunate enough to be disbenefited by use of tamoxifen.

When including the individual covariates of duration of tamoxifen and duration of follow-up in the same age bracket, (Models C and D), it appears very unlikely that either of these factors has any significant interplay with the model, as the coefficients for both have wide CrIs that include 0.

However, of the two covariates, there is evidence to indicate that duration of follow-up may exert the stronger influence, as the lower bound of the 95% CrI more closely approaches 0 (95% CrI -0.009; 0.030). This is possibly because a longer follow-up period allows longer time for any endometrial cancers to develop and be detected. Longer follow-up for trials involving women using tamoxifen may therefore be advisable to improve the possibility of detecting potential cases of endometrial carcinoma. However, when both duration of tamoxifen and length of follow-up are included in the model (Model E), there is no clear evidence that either exerts a stronger influence than the other or indeed that either of these variables plays a significant role in modifying the influence of endometrial carcinoma within the model.

For women in older age ranges (Model B) where some of the data relate to women aged 70 and over, the net benefit is reduced, to a median value of 0.504 on the QoL scale (95% CrI 0.283; 0.879). The total deaths prevented appears to be reduced for the older age bracket (median 0.395, 95% CrI 0.240; 0.0.556, compared with 0.500, 95% CrI 0.307; 0.700), thus lowering the net benefit. Even in this age group however, the benefits of tamoxifen clearly outweigh the possible risk of harm due to endometrial cancer.

Table 11.3: Results for tamoxifen and endometrial cancer harm–benefit analysis.

Model	A	B	C	D	E
Age group*	50–59	70–79	50–59	50–59	50–59
Duration of tamoxifen**	NA	NA	Yes	NA	Yes
Duration of follow-up**	NA	NA	NA	Yes	Yes
RRI.ec	2.323	2.268	1.84	1.734	1.948
	(1.09; 4.676)	(1.098; 4.725)	(0.414; 5.499)	(0.604; 4.100)	(0.433; 6.380)
RRR.br	0.448	0.447	0.447	0.448	0.446
	(0.324; 0.529)	(0.323; 0.529)	(0.321; 0.530)	(0.320; 0.530)	(0.320; 0.529)
Average risk of Ca breast	0.433	0.432	0.432	0.432	0.432
	(0.312; 0.563)	(0.312; 0.561)	(0.313; 0.563)	(0.314; 0.561)	(0.312; 0.560)
Deaths prevented	0.019	0.016	0.019	0.019	0.019
(Ca breast)	(0.012; 0.027)	(0.010; 0.022)	(0.012; 0.027)	(0.012; 0.027)	(0.012; 0.027)
Deaths caused	0.0002	0.0006	0.0002	0.00018	0.003
(Ca endometrium)	(0.0001; 0.0005)	(0.0003; 0.0014)	(0.00005; 0.0006)	(0.00006; 0.00046)	(0.0006; 0.008)
Net deaths prevented	0.019	0.015	0.019	0.019	0.012
	(0.012; 0.027)	(0.009; 0.021)	(0.012; 0.027)	(0.012; 0.027)	(0.00003; 0.019)
Probability of harm (mean)	0.0005	0.0004	0.0003	0.0003	0.0005
Risk level (Ca endometrium)	0.0015	0.0038	0.0015	0.0015	0.0015
	(0.0014; 0.0016)	(0.0036; 0.0041)	(0.0014; 0.0016)	(0.0014; 0.0016)	(0.0014; 0.0016)
Total death	0.500	0.395	0.500	0.500	0.498
	(0.307; 0.700)	(0.240; 0.556)	(0.308; 0.701)	(0.308; 0.700)	(0.307; 0.695)
Net benefit	**0.615**	**0.504**	0.614	**0.616**	**0.613**
	(0.359; 1.011)	**(0.283; 0.879)**	**(0.361; 1.006)**	**(0.362; 1.011)**	**(0.360; 1.003)**
Beta1†	NA	NA	-0.008	NA	0.006 (-0.032; 0.045)
			(-0.033; 0.022)		
Beta2‡	NA	NA	NA	0.012	0.014 (-0.016; 0.043)
				(-0.009; 0.030)	

All values are median (95% credible interval), unless stated otherwise; † Coefficient for duration of tamoxifen; ‡ Coefficient for duration of follow-up; *Age ranges refer to risk of death following breast cancer recurrence and risk of endometrial cancer occurrence; **Durations relate to occurrence of endometrial cancer; NA: Not applicable.

11.7.2 Discussion of methodological issues

There are several concerns regarding the development of this model and data sources used to inform it.

The construction of an overall harms and benefits model such as in this example necessitates the use of primary data from several disparate sources, that were never intended to be brought together to inform one model. Therefore, an obvious concern is that data that may have differing levels of validity are being treated as equally valid in this model (although this is an important issue with any modelling exercise).

One area in which the model is slightly unrepresentative is in the way that the two types of patients who are prevented from having a recurrence of breast cancer are treated by the model. Patients who would have had a recurrence of breast cancer and died within 5 years (of commencement of tamoxifen) have the same amount of additional life with a QoL of 1, which is the same as the people who would have had a recurrence but not died within 5 years. In actuality, the patients who would have died within 5 years should receive a longer period of life with a QoL of 1, but this difference is not incorporated into the model.

It is also the case that different types of data may refer to different patient groups. For example, no attempt has been made in this study to differentiate between pre- and post-menopausal patients, although in some instances, relevant information was available, for example, regarding some of the primary studies used by Braithwaite *et al.* (2003). Information regarding menopausal status may not be available for all types of primary data, bringing up the question of whether it should be used when available. In this model, there may be wide discrepancies in the types of patients involved in the primary data sources, hence creating concern regarding the validity of combining different data into a single model. However, this concern may be outweighed by the benefits of creating a quantitative harm–benefit model.

Another issue to point out is that women who have previously undergone a hysterectomy have no risk of developing endometrial cancer. Therefore, the model would be better informed by data regarding risk of endometrial cancer which excludes such women from the denominators, to avoid artificially reducing risk of endometrial cancer in women who do have an intact uterus. Also, the

control arms across different studies within the meta-analysis by Braithwaite *et al*. (2003) consisted of several different treatments across the primary studies, and an assumption has been made that these adjuvant treatments have no influence on endometrial cancer risk.

The use of QoL data is another source of uncertainty. In this model, QoL is evaluated for a 5-year period in terms of events (breast cancer recurrences and cases of endometrial cancer) prevented, which may or may not have led to deaths, and the added QoL incurred.

Such data can only be subjective, even if based on QoL data aggregated across many different patients. By using QoL as a 'common currency' in a model intended to inform clinical decision-making for individual patients, there is a risk of including subjectivity that may not be appropriate for an individual patient. In the light of this, a model that does not include QoL but considers only risk of death for both breast cancer recurrence and endometrial cancer, with life expectancy for patients who do eventually die, may be more helpful.

This aspect of the model is covered by the total deaths prevented, which start from the commencement of tamoxifen therapy and are continued for a 50-year period. The total deaths prevented is in effect the number of prevented deaths and associated QoL at a fixed value of 1, which is then discounted over the 50-year period to account for deaths during this time and associated reduction in QoL due to ageing. It is also possible that this aspect of the model is over-generous in its allocation of QoL and attrition rate of participants due to death.

As a final step, inclusion of other harms and possibly other potential benefits due to use of tamoxifen in patients with breast cancer would be a useful extension to this model. Amongst the additional potential harms associated with tamoxifen discussed by Braithwaite *et al*. (2003) are cerebrovascular accidents, pulmonary emboli, and gastrointestinal cancers. In these cases, the increased risk of harm appeared to be less than for endometrial cancer, but their inclusion into a harm–benefit model would serve to reduce the net benefit due to tamoxifen. Conversely, there was evidence from the study by Braithwaite *et al*. (2003) that there were other, unintended, benefits due to tamoxifen, such as reduced risk of myocardial infarction (MI), and these could also be included as potential benefits, despite the fact that the tamoxifen was not being taken with the aim of reducing MI risk.

Another area for exploration would be the harms and benefits of tamoxifen when used for other indications, for example, prevention of primary breast cancer for women in high-risk categories.

11.7.3 Final considerations

This case study extends the work of previous chapters, both in terms of methodology and conceptual issues. For example, the work of Chapter 7 largely related to issues of sparsity of events, and the difficulty of receiving a clear signal from the dataset in such circumstances. The analyses were unable to confirm or allay concerns regarding an increased risk of suicidality for users of paroxetine. An extension of these methods into a harm–benefit model, taking into account the benefits to QoL in terms of relief from symptoms, set against loss of QoL due to suicide, would facilitate decision-making, especially if such analyses were performed across a range of anti-depressant drugs, and including covariates such as age where possible.

The IPD models of Chapter 8 would be especially appropriate with time-to-event outcomes, which could be converted to a QoL scale, whilst the mixed treatment comparison (MTC) models (Chapters 9 and 10) would facilitate comparison across multiple treatments, for both benefits and harms. These models, and in particular hierarchical models (Chapter 10) would enable the best use to be made of a diverse dataset, within each outcome and for each treatment, prior to combining the results of these models into a net–benefit model.

Despite the limitations of this case-study, it nevertheless serves as an example of how an evidence synthesis of adverse events data (using methods discussed in this thesis) can be combined with information on potential benefits in a more meaningful way. This approach to decision-modelling brings together risk of adverse events and potential for positive therapeutic effect, quantifies them both by means of QoL scales, and produces an overall result, with uncertainty sources incorporated throughout.

A wider variety of approaches to meta-analysis methods for the adverse events data can facilitate such harm–benefit modelling by providing appropriate statistical methodologies for the dataset available, which could be evaluated by simulation studies. Input from a clinical perspective could also be incorporated to assist in appropriate modelling.

11.8 Summary

A harm–benefit model for tamoxifen use when indicated for prevention of breast cancer recurrence was developed. Data from multiple disparate sources informed different elements of the model, including relative risk of breast cancer recurrence (for tamoxifen users compared to non-users), underlying risk for breast cancer recurrence, risk of death following breast cancer recurrence, QoL following breast cancer recurrence, relative risk of endometrial cancer (for tamoxifen users compared to non-users), underlying risk of endometrial cancer (in the general population), risk of death following endometrial cancer, and QoL following endometrial cancer.

Models were also fitted using data relevant to different age brackets for comparison, and using additional covariates in relation to risk of endometrial cancer, for example, duration of tamoxifen use and duration of follow-up.

Results indicated that the net benefit from using tamoxifen appeared to be positive regarding the interplay between prevention of breast cancer recurrence and endometrial cancer, with a commensurately low probability of harm. This benefit related to tamoxifen was seen in age ranges 50–59 and 70 years upwards, although it was somewhat reduced in the older age bracket. There was no strong evidence to indicate that duration of follow-up or duration of tamoxifen treatment was a significant factor in the model. Although duration of follow-up did appear to have some influence, the evidence was not conclusive

Further extensions to the model could include incorporation of additional harms and benefits within the model, and consideration of different model formats.

12

Discussion and development

12.1 Overview

From a clinical perspective, the potential risks of deleterious effects of medical interventions that are intended to be beneficial are a serious cause for concern. The use of evidence synthesis methods to make maximum use of all available data concerning adverse events would be desirable, and the need for evidence synthesis focusing on this area has been highlighted elsewhere (Loke *et al*. 2008). The work of Loke *et al*. (2008) is primarily directed towards the non-statistical aspects of evidence synthesis such as search strategies (also considered by Golder *et al*. 2006a; 2006b; 2006c; 2008). Hence, there was a requirement for further research into the statistical methods that may be appropriate for adverse events meta-analyses. This area has been considered previously (Sutton *et al*. 2002) and there is scope to extend this research.

The four main aims set out in Section 1.1 have been fulfilled in the following ways. Methodological issues related to evidence synthesis and meta-analysis have been described in Chapter 3, and, from a Bayesian perspective, in Chapter 4. This background in pre-existing methods was then used as the foundation for a discussion of the specific features of adverse events datasets that could prove challenging when attempting evidence synthesis, in the context of a specific clinical question (Chapter 5).

This review of statistical methodology was supported by a systematic literature review of 166 published meta-analyses in which the primary outcome was an adverse (or unintended) event (Chapter 6). The motivation behind this review was to discover the statistical methods already employed in meta-analyses of adverse events data, and to establish the ways in which the methodological challenges discussed in Chapter 5 had been identified and approached. These chapters also encompass the first three of the narrower objectives outlined in Section 1.1.

Some specific clinical issues that were potential case-studies were set out in Section 1.4.2. The second chapter of this thesis has the function of investigating in more detail the nature and extent of the clinical problem of adverse events.

Chapters 7–10 use case-studies to address different areas of statistical methodology in relation to adverse events. The main area of investigation in Chapter 7 is that of a dataset with sparse events, and this problematic issue is tackled by multiple meta-analyses using different methodologies.

Chapters 8–10 are linked by the use of a common clinical case-study: the potential increased risk of malignancy due to use of anti-tumour necrosis factor (anti-TNF) drugs in rheumatoid arthritis. Chapter 8 concentrates on one specific drug in this class, etanercept, and uses a dataset of individual patient data (IPD) from multiple clinical trials. Again, sparsity of events is a major issue, as well as use of IPD with time-to-event data for all participants. This clinical theme is expanded in Chapter 9, by the inclusion of two further anti-TNF drugs, plus additional aspects of therapy such as concomitant use of different drugs used in RA therapy, and the dose of the specific anti-TNF, which is an essential element of prescribing. As IPD were not available for this dataset, aggregate data were used throughout. From a statistical perspective, the methodology was extended by the use of mixed treatment comparison (MTC) methods.

The final chapter in this clinical area was Chapter 10, which incorporated further complexity into the MTC models, and conducted a sensitivity analysis across prior distributions for selected MTC models developed in Chapter 9.

The final aim proposed in Section 1.1 was to combine both adverse events data and data related to positive treatment outcomes in one overarching model. This model would use a quantitative assessment of the influence on quality of life (QoL), as a result of the intervention, with regard to the increased QoL due

to the therapeutic effect and the reduced QoL due to the adverse effects. The results of such a harm–benefit analysis would be of value in clinical decision-modelling. This aim is the subject of Chapter 11, the final case-study, which presents a harm–benefit model based on the effects of tamoxifen therapy in patients with breast cancer, with regard to the increased risk of endometrial cancer, offset against a reduced risk in recurrence of the original breast cancer.

Referring back to the specific objectives for this thesis, in Section 1.1, the fourth, fifth and sixth objectives are included within Chapters 7–11. The seventh objective is addressed to some extent within Chapter 11, but could be investigated more thoroughly. Areas for further extension to the work of this thesis are discussed in Section 12.4.

The aspects of this thesis that are novel, or extend previous work, are set out below:

1. the review of evidence synthesis methods previously employed in the context of adverse events data (Chapter 6);
2. the comprehensive comparison of meta-analysis methods applied to the same dataset, for which sparsity of events is an issue (Chapter 7);
3. fitting a random effects meta-analysis model to individual patient data with a time-to-event outcome, using a Poisson formulation (Chapter 8);
4. application of mixed treatment comparison techniques to adverse outcome data with sparsity of events (Chapter 9);
5. using hierarchical modelling in conjunction with mixed treatment comparisons, to enable 'borrowing strength' across datapoints, and comparing treatment-based differences in hierarchical structure within an MTC network (Chapter 10);
6. application of constraints on prior distributions to hierarchical models, as a means of incorporating prior beliefs into the model (Chapter 10); and
7. the use of net-benefit modelling, by integrating meta-analyses of adverse events data within a harm–benefit model for clinical decision-making.

These elements are discussed in more detail in Sections 12.2–12.3.

12.2 Review of meta-analysis methods and meta-analyses previously performed on adverse events data

An initial review of meta-analysis methods (Chapters 3 and 4) was followed by a discussion of the specific ways that meta-analysis methods may be difficult to implement with regard to adverse events data (Chapter 5). Such a review immediately provided a wealth of issues for meta-analysis challenges, and ideas for development of methods that would be highly pertinent to adverse events data. It also became apparent that the field of adverse events is in itself very diverse in clinical terms, and that more specific characterisation of the clinical issues would be necessary. Statistical aspects of adverse events data that would apply in certain clinical instances would not be relevant for other circumstances. In some cases, an issue highly applicable to adverse events would also be applicable in other areas, such as clinical efficacy. An example would be that of class effects (of a class of drug), which could be applied in terms of adverse events or positive events; other examples would be dose–response effects or effects over periods of time.

From a review of previous meta-analysis studies regarding adverse events (Chapter 6), the most outstanding issue was that in the majority of cases, no special consideration had been given by the authors to the fact that the outcome was one of adverse events.

This calls into question the usefulness of bringing together data from disparate clinical issues, with the only common theme being that of adverse events, when this common factor was not specifically acknowledged by the original authors of these papers. Although it is possible to develop a composite picture of what has been done in previous studies, it is difficult to perceive this composite as an overview of adverse events meta-analyses when there was little conception of performing an adverse events meta-analysis in the minds of the original authors. Rather, it represents a disparate grouping of studies with little in common, in terms of meta-analysis methods, and only a tenuous clinical connection.

Above all, such a review indicates that a greater understanding of meta-analysis issues in general, and in application to specific clinical areas, is required.

As a methodological issue, when considering areas for research within the framework of meta-analysis and evidence synthesis, both the clinical area and potential

statistical challenges should be considered in conjunction. This would ensure that the statistical issues can be considered on an equal par with the clinical aspects, and avoid clinical areas that are too disparate in terms of their statistical requirements for evidence synthesis to be considered in general terms of 'adverse events'.

12.3 Case-studies using adverse events data

Four separate case-studies of adverse events meta-analysis have been performed:

1. selective serotonin re-uptake inhibitors and suicide risk;
2. etanercept and malignancy risk in rheumatoid arthritis sufferers using individual patient data;
3. anti-TNFs and malignancy risk in rheumatoid arthritis sufferers using mixed-treatment comparisons; and
4. tamoxifen and risk of endometrial cancer in breast cancer sufferers.

These case-studies are discussed below, with regard to the methodological issues highlighted by each one, with commentary on the important issues raised by each model and the ways in which they may be used and developed for adverse events evidence synthesis.

12.3.1 Selective serotonin re-uptake inhibitors and suicide risk

The aim of this initial case-study (presented in Chapter 7) was to re-evaluate data concerning the use of selective serotonin re-uptake inhibitors and any association with suicidal ideation, deliberate self-harm and completed suicides, especially in children and adolescents (Gunnell *et al.* 2005; Gibbons *et al.* 2006; Hammad *et al.* 2006). Using data from Glaxo-Smith-Kline (the manufacturers of paroxetine), a series of meta-analyses using several outcome metrics and different continuity corrections were performed. Bayesian methods were also used to contrast with traditional statistical methods.

This case-study set out the importance of multiple analyses to compare different outcome metrics, and use of continuity corrections, also the importance of comparing traditional with Bayesian statistics. Ideally, statistical background

knowledge, initial scrutiny of the dataset and understanding of the clinical issues involved should inform *a priori* the statistical methods. For example, knowledge of the baseline risk could inform whether the outcome metric would be best described on a difference or relative scale, or knowledge of imbalanced treatment arms could determine what type of continuity correction may be most suitable. Previous knowledge from other studies in a similar area could inform prior distributions for a Bayesian analysis.

In reality however, an iterative process of multiple analyses would help to provide a more complete picture, and with hindsight could help to identify the statistical methods most suited to a clinical problem.

12.3.2 Etanercept and malignancy risk in rheumatoid arthritis sufferers

Etanercept is an anti-TNF drug commonly used as a long-term treatment for patients with rheumatoid arthritis (RA); however, there were concerns regarding its safety for malignancy. To investigate this issue, individual patient data (IPD) were obtained from the relevant pharmaceutical companies (Amgen and Wyeth). Using this IPD, multiple models including both fixed effect (FE) and random effects (RE) were used to contrast results, also considering issues such as excluding certain types of cancer and different durations of follow-up (see Chapter 8). The results of this study have been published elsewhere (Bongartz *et al*. 2009). From a statistical perspective, the novel aspect of these analyses was the RE model fitted using a Poisson generalised linear model, as described in Section 8.2.3.

This case-study highlighted the value of using IPD for adverse events, especially in the light of sparse events in clinical trials. The use of IPD comes into its own when knowledge of individual cases is of importance, for example, considering patient-specific factors for each case, that would be lost when using aggregate data. In this example, evidence synthesis may be best presented as a hybrid between statistical methods and clinical narrative of individual cases, which can add depth to the analysis, and highlight certain types of patient, or certain types of adverse event that would merit further investigation.

12.3.3 Anti-TNFs and malignancy risk in rheumatoid arthritis sufferers using mixed treatment comparisons

The work on etanercept as a potential risk for malignancy in rheumatoid arthritis sufferers was extended by adding data for two other anti-TNFs, adalimumab and infliximab. Notably, these were of a different pharmacological class of anti-TNF. Previous research had associated these drugs with malignancy and infections (Bongartz *et al.* 2006; Leombruno *et al.* 2008).

The three drugs involved were further complicated by use of dose and additional anti-rheumatic drugs, combining to produce treatments that at their highest level of complexity were determined by anti-TNF, dose and presence of additional anti-rheumatoid drugs. This case-study is discussed in Chapters 9 and 10.

The clinical issues presented by this problem, effectively a comparison of drugs within a class, but with issues relating to dose and use of concurrent medication (using drugs of other classes), were in themselves an opportunity to use mixed-treatment comparison (MTC) models, which have been described by Lu & Ades (2004) and Caldwell *et al.* (2005), and are discussed in Sections 9.2 and 9.4.

The characterisation of different treatments by varying combinations of factors naturally lent itself to the use of mixed-treatment comparison (MTC) models, which, when combined with additional problems of sparsity of events, and in some cases sparsity of trials, resulted in difficulties of analysis. The importance of the individual MTC network in influencing the statistical results was demonstrated, as variable results were arrived at depending on the network used. However, the great advantage of using these MTC modelling, fitted using Bayesian methods, is the ability to include studies with sparse events. The initial MTC networks and models are discussed in Chapter 9.

The MTC modelling is extended in Chapter 10, with the use of hierarchical models and constraints (Prevost *et al.* 2000) to allow the 'borrowing of strength' (Higgins & Whitehead 1996) across treatments that are in some way connected (for example by dose of anti-TNF, which connects different members of the anti-TNF group). The use of hierarchical modelling with regard to an MTC, in

association with the clinical dimension of adverse events, sparsity of data and trials, is the novel aspect of this chapter.

12.3.4 Tamoxifen and risk of endometrial cancer in breast cancer sufferers

In the final case-study, a harms and benefits model was developed (Chapter 11), using tamoxifen, which reduces risk of recurrence of breast cancer in patients already suffering from this disease, counterbalanced against risk of endometrial cancer. The concept and methods of net-benefit modelling are put forward by Glasziou & Irwig (1995) and extended into Bayesian methodology by Sutton *et al*. (2005), as discussed in Section 11.2.1. The complexity of modelling harms and benefits leads to the question of how accurate can such a model be, with the aim of applying the model to clinical decision-making at the level of the individual patient.

There are many levels of uncertainty in such a model, with a major issue of the validity of bringing together data from disparate sources, that were never intended to be used in conjunction to inform statistical modelling or clinical decision-making. Certain aspects of the data are inherently subjective, such as quality of life (QoL) data. A further issue is that of temporality: how far into the future (following the treatment period) can such a model be extended with any degree of validity?

Modelling of harms against concomitant benefits is perhaps the most clinically useful of all the methodology presented in this work, in view of the fact that efficacy of an intervention has already been demonstrated. The harms resulting from an intervention can only be interpreted clinically in the light of the benefits that a patient would be eligible to gain from the treatment, and if quantitatively expressed, can be conceptualised as a 'negative benefit' to set against the positive benefit.

However, in some instances it is more highly relevant to prove that a clinically significant risk of adverse outcomes does in fact exist at all, prior to attempting to quantify such a risk against the potential benefits.

12.4 Potential extensions of the current work

Based on the initial areas identified in Chapter 5 as being methodological areas of special relevance to evidence synthesis for adverse events, only some of these have been addressed. The previous case-studies have covered aspects of IPD, sparse data, subgroup analysis, influence of timing of analysis and duration of treatment, issues regarding dosage, class effects in comparison to individual drug effects, and harm–benefit analysis. There is much scope to extend many of these case-studies to further investigate these areas, and to include additional case-studies to explore other ways in which these issues could be considered and dealt with.

The use of IPD with adverse events data has much potential for future development. The case-study of Chapter 8, which uses IPD, is the only one that does not incorporate Bayesian methods in any way, and Bayesian approaches to use of IPD would be an area of interest to explore.

It would have been particularly useful to incorporate IPD in conjunction with the MTC modelling and hierarchical modelling used in conjunction with the MTC analyses. A harm–benefit analysis would also be facilitated by use of IPD, and, ideally, the studies would provide IPD for each patient for both therapeutic effects and adverse events. This form of data would be of the best quality for a harm–benefit analysis. With information available for benefits and harms for each patient, it would be possible to derive QoL data from each patient directly, and to investigate patient-level covariates that are associated with both harms and benefits.

The MTC modelling has raised particular issues regarding the definition of treatments according to different parameters, and the difficulties of choosing the most appropriate model, especially where sparsity of events becomes an issue for the models of highest complexity. Hierarchical Bayesian methodology is another area of particular interest that lends itself to adverse events modelling, promoting the integration of available data in ways that can militate against the effects of data sparsity.

One area of particular relevance that has not been considered within this thesis is that of combining data from randomised and non-randomised (observational) studies. The example of hierarchical modelling described by Prevost *et*

al. (2000) addresses the issue of differential degrees in bias between the two study types, and this approach could be applied to adverse events analyses. An MTC model could also be created if there were multiple treatment types, with a hierarchical model placed on study type and treatment type.

Another way of assimilating information from observational studies would be to use data from such studies to inform prior distributions for Bayesian meta-analyses of randomised trials, in the way that data from the meta-analysis by Kaizar *et al*. (2006) were used to form a prior distribution (see Section 7.3.1). One of the most fundamental ways to use external data in a meta-analysis is that of narrative review, which can then be transformed into a prior distribution with the aid of clinical expertise. Although the focus of this thesis has been on quantitative methods, it should be remembered that evidence synthesis can benefit from non-quantitative methods to provide depth of understanding to a clinical problem and assist in formulating an answer.

The scenario of adverse events data being available for the same drug used in the treatment of different medical conditions has not been addressed in this thesis, and is an area worthy of consideration. The use of MTC modelling could again form part of the solution, as the disease/treatment combination could be considered as a 'node' within the MTC network. This approach would facilitate understanding of the different profiles of potential harm across different diseases, whilst allowing 'borrowing of strength' across the dataset.

There are also many potential ways to extend the harm–benefit modelling, for example by including multiple harms and benefits in the model, and exploring ways to 'individualise' the model for patient-specific decision-modelling; as medicine becomes more oriented toward personalising treatments, patient-specific modelling has the potential to become increasingly important.

A harm–benefit model can be adapted to different subgroups of patients (for example, different age ranges) by substituting appropriate data values, but perhaps the most relevant use of such a model would be in individual decision-making, whereby a patient would be able to input her or his own values for the subjective QoL elements, thus reducing the uncertainty of this element of the model. In areas where the generic data provide clear-cut conclusions of unacceptable risk for certain patient groups, this type of modelling could also be of use, but

the risk of adverse events against potential benefit would need to be high to counterbalance against the modelling uncertainty.

Within a harm–benefit model there is wide scope for different meta-analysis methods for incorporating data, for example the MTC modelling, with use of hierarchies and constraints, or using external data to formulate a prior distribution. As discussed above, IPD would also be of benefit, especially if data exist on harms and benefits for individual patients.

All of the case-studies in this work are based on drug interventions, in part because data regarding adverse events of such interventions are relatively easy to obtain. The eighth objective set out in Section 1.1 was to apply the methods used in this thesis to non-pharmaceutical interventions, such as surgical procedures or medical devices. A particularly important area to investigate for adverse events would be public health interventions, for example, fluoridation of water supplies, vaccinations (a topical example being those for 'swine flu' (influenza virus H1N1)), or public 'education' programmes. As large numbers of people are exposed to these interventions, it is important to thoroughly investigate any potential adverse effects that may be caused.

Cost issues are of particular importance to adverse events analyses, in two main ways. Firstly, there are issues of cost-effectiveness, whereby the financial costs of adverse events can be incorporated into a cost-effectiveness model based on a benefit–harm approach. Secondly, there are potential costs to pharmaceutical companies and healthcare providers regarding litigation by patients who have suffered harm due to an intervention. Costs in these cases are comprised of compensation paid to the patient (or the patient's family) and costs of legal proceedings. In the case of state-funded healthcare, such as the National Health Service, these costs will ultimately be met by the taxpayer. As well as financial costs, organisations are also concerned about damage done to the reputation of the organisation in areas where patient safety is concerned. Effective ways to investigate the potential risk of interventions are therefore beneficial to patients, to the pharmaceutical company and to healthcare providers.

The harm–benefit model described in Chapter 10 could be extended to incorporate the dimension of cost. By definition however, in a state-funded health system, or in a system of private health provision, costs are met not by the individual patient but by an aggregation of tax-payers or insurance holders.

Hence, modelling at the level of cost-effectiveness must be done by definition at the level of the population. This introduces a risk that individual patients may be denied a treatment that would be effective (and indeed maybe even cost-effective) for that individual. Use of individualised harm–benefit modelling would help to avoid such occurrences.

When evaluating different statistical methods, simulation studies can be employed, whereby different methods are applied to a common dataset, which has been simulated according to certain, known, specifications; for example, event rates and treatment effects. For the purposes of evidence synthesis, a dataset could be simulated comprising IPD from multiple studies, or by simulating the aggregate data for each study. In this situation, the methodologies can be compared against each other, in terms of their ability to accurately analyse the dataset. For example, the multiple meta-analysis methods compared in Chapter 7 could be evaluated using simulation methods. This would allow the methods to be evaluated using parameters such as bias, coverage and statistical power, as described by Bradburn *et al.* (2007) in Section 5.2.1.

12.5 Conclusions

From a clinical standpoint, an analysis of adverse events can impact on the following decisions (using a pharmaceutical intervention as an example):

1. whether or not to use a drug at all (i.e. to ban the drug on safety grounds and/or cost-effectiveness grounds);
2. to recommend the drug for use in certain subgroups or to avoid in certain subgroups;
3. select drugs for specific monitoring or further research where there are safety concerns;
4. to make a decision at the level of the individual patient whether the potential for benefit outweighs the risk of harm;
5. to select the drug with the most positive benefit–harm balance within a particular class or for a particular indication; and

6. to determine that if a drug is to be used, then actions should be taken to prevent adverse effects or alleviate their influence, or at the very least warn patients of the possibility of adverse events.

For the first and third items in this list, choice of metric is highly important, as well as combining data from all available sources and considering temporal effects on the outcome. For the second item, subgroup analysis is clearly important, as well as concerns regarding indication for an intervention.

For the fourth, fifth and sixth items, the modelling issues become more complex, with harm–benefit models being essential, and with the inclusion of MTC analysis for selection among a class of drugs. Knowledge of risk of adverse events can then be used to inform clinical decision-making at the level of whether or not to use a particular drug, and whether to take any further action to alleviate potential harms.

To paraphrase a traditional adage, it is not what you look for, it is the way you look for it that counts. The evidence synthesis of adverse events data provides a strong example of this. The methods used can be highly complex and can have a strong influence on the results of any evidence synthesis performed. In view of this, it is very important to be explicit about which methods are being used and why they were chosen. Models should be carefully evaluated, not only for the algebraic functionality, but for any assumptions being made, sources of uncertainty, the choice of outcome metric and the choice of prior distributions in a Bayesian model.

Ideally, evidence synthesis methods for adverse events data will continue to be developed, with this specific clinical aspect in mind. The development of statistical methods closely aligned with a clinical framework would be highly beneficial for individual patient care, promotion of public health, and cost-effectiveness.

Appendix A: References included in review of previous meta-analyses with an adverse event as primary outcome

S1. Alfirevic Z, Sundberg K & Brigham S (2003). Amniocentesis and chorionic villus sampling for prenatal diagnosis. *Cochrane Database of Systematic Reviews*, Issue 3, Art. No.: CD003252. DOI: 10.1002/14651858.CD003252.

S2. Allen DB, Mullen M & Mullen B (1994). A meta-analysis of the effect of oral and inhaled corticosteroids on growth. *Journal of Allergy and Clinical Immunology*, 93(6): 967–976.

S3. Allison DB, Mentore JL, Heo M, Chandler LP, Cappelleri JC, Infante MC & Weiden PJ (1999). Antipsychotic-induced weight gain: a comprehensive research synthesis. *American Journal of Psychiatry*, 156(11): 1686–1696.

S4. Anderson JW, Kendall CWC & Jenkins DJA (2003). Importance of weight management in Type 2 diabetes: review with meta-analysis of clinical studies. *Journal of the American College of Nutrition*, 22(5): 331–333.

S5. Anderson-Hanley C, Sherman ML, Riggs R, Agocha VB & Compas BE (2003). Neuropsychological effects of treatments for adults with cancer: a meta-analysis and review of the literature. *Journal of the International Neuropsychological Society*, 9(7): 967–982.

S6. Ashcroft DM, Chapman SR, Clark WK & Millson DS (2001). Upper gastroduodenal ulceration in arthritis patients treated with celecoxib. *Annals of Pharmacotherapy*, 35(7-8): 829–834.

S7. Ashraf E, Cooper S, Kellstein D & Jayawardena S (2001). Safety profile of nonprescription ibuprofen in the elderly osteoarthritis patient: a meta-analysis. *Inflammopharmacology*, 9(1-2): 35–41.

S8. Banks E (2001). Hormone replacement therapy and the sensitivity and specificity of breast cancer screening: a review. *Journal of Medical Screening*, 8(1): 29–35.

S9. Barbui C & Saraceno B (1996). Low-dose neuroleptic therapy and extrapyramidal side effects in schizophrenia: an effect size analysis. *European Psychiatry*, 11(8): 412–415.

S10. Barone JE, Tucker JB, Rassias D & Corvo PR (2001). Routine perioperative pulmonary artery catheterization has no effect on rate of complications in vascular surgery: a meta-analysis. *American Surgeon*, 67(7): 674–679.

S11. Ben-David S, Einarson T, Ben-David Y, Nulman I, Pastuszak A & Koren G (1995). The safety of nitrofurantoin during the first trimester of pregnancy: meta-analysis. *Fundamental and Clinical Pharmacology*, 9(5): 503–507.

S12. Bender BG, Berning S, Dudden R, Milgrom H & Tran ZV (2003). Sedation and performance impairment of diphenhydramine and second-generation antihistamines: a meta-analysis. *Journal of Allergy and Clinical Immunology*, 111(4): 770–776.

S13. Berglundh T, Persson L & Klinge B (2002). A systematic review of the incidence of biological and technical complications in implant dentistry reported in prospective longitudinal studies of at least 5 years. *Journal of Clinical Periodontology*, 29(Supplement 3): 197–212.
S14. Bernal-Delgado E, Latour-Perez J, Pradas-Arnal F & Gomez-Lopez LI (1998). The association between vasectomy and prostate cancer: a systematic review of the literature. *Fertility and Sterility*, 70(2): 191–200.
S15. Brown JS, Sawaya G, Thom DH & Grady D (2000). Hysterectomy and urinary incontinence: a systematic review. *Lancet*, 356(9229): 535–539.
S16. Brown S, Small R, Faber B, Krastev A & Davis P (2002). Early postnatal discharge from hospital for healthy mothers and term infants. *Cochrane Database of Systematic Reviews*, Issue 3, Art. No.: CD002958. DOI: 10.1002/14651858.CD002958.
S17. Brumback BA, Holmes LB & Ryan LM (1999). Adverse effects of chorionic villus sampling: a meta-analysis. *Statistics in Medicine*, 18(16): 2163–2175.
S18. Cannegieter SC, Rosendaal FR & Briet E (1994). Thromboembolic and bleeding complications in patients with mechanical heart valve prostheses. *Circulation*, 89(2): 635–641.
S19. Caraballo PJ, Gabriel SE, Castro MR, Atkinson EJ & Melton LJ (1999). Changes in bone density after exposure to oral anticoagulants: a meta-analysis. *Osteoporosis International*, 9(5): 441–448.
S20. Cardwell ME, Siviter G & Smith AF (2005). Non-steroidal anti-inflammatory drugs and perioperative bleeding in paediatric tonsillectomy. *Cochrane Database of Systematic Reviews*, Issue 2, Art. No.: CD003591.
DOI: 10.1002/14651858.CD003591.pub2.
S21. Caro-Paton T, Carvajal A, Martin de Diego I, Martin-Arias LH, Alvarez Requejo A & Rodriguez Pinilla E (1997). Is metronidazole teratogenic: a meta-analysis. *British Journal of Clinical Pharmacology*, 44(2): 179–182.
S22. Chan W-S, Ray J, Wai EK, Ginsburg S, Hannah ME, Corey PN & Ginsberg JS (2004). Risk of stroke in women exposed to low-dose oral contraceptives: a critical evaluation of the evidence. *Archives of Internal Medicine*, 164(7): 741–747.
S23. Chang CH, Chen KY, Lai MY & Chan KA (2002). Meta-analysis: ribavirin-induced haemolytic anaemia in patients with chronic hepatitis C. *Alimentary Pharmacology and Therapeutics*, 16(9): 1623–1632.
S24. Chapron C, Fauconnier A, Goffinet F, Breart G & Dubuisson JB (2002). Laparoscopic surgery is not inherently dangerous for patients presenting with benign gynaecologic pathology: results of a meta-analysis. *Human Reproduction*, 17(5): 1334–1342.
S25. Choi PT, Galinski SE, Takeuchi L, Lucas S, Tamayo C & Jadad AR (2003). PDPH is a common complication of neuraxial blockade in parturients: a meta-analysis of obstetrical studies. *Canadian Journal of Anaesthesia*, 50(5): 460–469.

S26. Col NF, Hirota LK, Orr RK, Erban JK, Wong JB & Lau J (2001). Hormone replacement therapy after breast cancer: a systematic review and quantitative assessment of risk. *Journal of Clinical Oncology*, 19(8): 2357–2363.
S27. Collaborative Group on Hormonal Factors in Breast Cancer (1996). Breast cancer and hormonal contraceptives: collaborative reanalysis of individual data on 53 297 women with breast cancer and 100 239 women without breast cancer from 54 epidemiological studies. *Lancet*, 347(9017): 1713–1727.
S28. Correll CU, Leucht S & Kane JM (2004). Lower risk for tardive dyskinesia associated with second-generation antipsychotics: a systematic review of 1-year studies. *American Journal of Psychiatry*, 161(3): 414–425.
S29. Coughlin SS, Giustozzi A, Smith SJ & Lee NC (2000). A meta-analysis of estrogen replacement therapy and risk of epithelial ovarian cancer. *Journal of Clinical Epidemiology*, 53(4): 367–375.
S30. Cutler C, Giri S, Jeyapalan S, Paniagua D, Viswanathan A & Antin JH (2001). Acute and chronic graft-versus-host disease after allogeneic peripheral-blood stem-cell and bone marrow transplantation: a meta-analysis. *Journal of Clinical Oncology*, 19(16): 3685–3691.
S31. Derry S & Loke YK (2000). Risk of gastrointestinal haemorrhage with long term use of aspirin: meta-analysis. *BMJ*, 321(7270): 1183–1187.
S32. Dezfulian C, Lavelle J, Nallamothu BK, Kaufman SR & Saint S (2003). Rates of infection for single-lumen versus multilumen central venous catheters: a meta-analysis. *Critical Care Medicine*, 31(9): 2385–2390.
S33. DiMatteo MR, Morton SC, Lepper HS, Damush TM, Carney MF, Pearson M & Kahn KL (1996). Cesarean childbirth and psychosocial outcomes: a meta-analysis. *Health Psychology*, 15(4): 303–314.
S34. Dolovich LR, Addis A, Vaillancourt RJ, Power BJ, Koren G & Einarson TR (1998). Benzodiazepine use in pregnancy and major malformations of oral cleft: meta-analysis of cohort and case-control studies. *BMJ*, 317(7162): 839–843.
S35. Dong EW, Connelly JE, Borden SP, Yorzyk W, Passov DG, Kupelnick B, Luo DH & Ross SD (1997). A systematic review and meta-analysis of the incidence of cancer in randomized, controlled trials of verapamil. *Pharmacotherapy*, 17(6): 1210–1219.
S36. Doren M, Nilsson JA & Johnell O (2003). Effects of specific post-menopausal hormone therapies on bone mineral density in post-menopausal women: a meta-analysis. *Human Reproduction*, 18(8): 1737–1746.
S37. Douketis JD, Ginsberg JS, Holbrook A, Crowther M, Duku EK & Burrows RF (1997). A reevaluation of the risk for venous thromboembolism with the use of oral contraceptives and hormone replacement therapy. *Archives of Internal Medicine*, 157(14): 1522–1530.
S38. Doyle LW & Davis PG (2000). Postnatal corticosteroids in preterm infants: systematic review of effects on mortality and motor function. *Journal of Paediatrics and Child Health*, 36(2): 101–107.

S39. Duffy G & Neal KR (1996). Differences in post-operative infection rates between patients receiving autologous and allogeneic blood transfusion: a meta-analysis of published randomized and nonrandomized studies. *Transfusion Medicine*, 6(4): 325–328.
S40. Egger M, Davey Smith G, Stettler C & Diem P (1997). Risk of adverse effects of intensified treatment in insulin-dependent diabetes mellitus: a meta-analysis. *Diabetic Medicine*, 14(11): 919–928.
S41. Eikelboom JW, Mehta SR, Pogue J & Yusuf S (2001). Safety outcomes in meta-analyses of phase 2 vs phase 3 randomized trials: intracranial hemorrhage in trials of bolus thrombolytic therapy. *JAMA*, 285(4): 444–450.
S42. Etminan M, Samii A, Takkouche B & Rochon PA (2001). Increased risk of somnolence with the new dopamine agonists in patients with Parkinson's disease: a meta-analysis of randomised controlled trials. *Drug Safety*, 24(11): 863–868.
S43. Etminan M, Gill S & Samii A (2003). Comparison of the risk of adverse events with pramipexole and ropinirole in patients with Parkinson's disease: a meta-analysis. *Drug Safety*, 26(6): 439–444.
S44. Etminan M, Gill S & Samii A (2003). Effect of non-steroidal anti-inflammatory drugs on risk of Alzheimer's disease: systematic review and meta-analysis of observational studies. *BMJ*, 327(7407): 128–131.
S45. Fernandez E, La Vecchia C, Balducci A, Chatenoud L, Franceschi S & Negri E (2001). Oral contraceptives and colorectal cancer risk: a meta-analysis. *British Journal of Cancer*, 84(5): 722–727.
S46. Furberg CD, Psaty BM & Meyer JV (1995). Nifedipine: dose-related increase in mortality in patients with coronary heart disease. *Circulation*, 92(5): 1326–1331.
S47. Gabriel Sánchez R, Carmona L, Roque M, Sánchez Gomez LM & Bonfill X (2005). Hormone replacement therapy for preventing cardiovascular disease in post-menopausal women. *Cochrane Database of Systematic Reviews*, Issue 2, Art. No.: CD002229. DOI: 10.1002/14651858.CD002229.pub2.
S48. Garcia Rodriguez LA (1997). Nonsteroidal antiinflammatory drugs, ulcers and risk: a collaborative meta-analysis. *Seminars in Arthritis and Rheumatism*, 26(6 Supplement 1): 16–20.
S49. Garcia Rodriguez LA, Hernandez-Diaz S & de Abajo FJ (2001). Association between aspirin and upper gastrointestinal complications: systematic review of epidemiologic studies. *British Journal of Clinical Pharmacology*, 52(5): 563–571.
S50. Garg PP, Kerlikowske K, Subak L & Grady D (1998). Hormone replacement therapy and the risk of epithelial ovarian carcinoma: a meta-analysis. *Obstetrics and Gynecology*, 92(3): 472–479.
S51. Gillum LA, Mamidipudi SK & Claiborne Johnston S (2000). Ischemic stroke risk with oral contraceptives: a meta-analysis. *JAMA*, 284(1): 72–78.
S52. Gordon PV, Young ML & Marshall DD (2001). Focal small bowel perforation: an adverse effect of early postnatal dexamethasone therapy in extremely low birth weight infants. *Journal of Perinatology*, 21(3): 156–160.

S53. Grady D, Gebretsadik T, Kerlikowske K, Ernster V & Petitti D (1995). Hormone replacement therapy and endometrial cancer risk: a meta-analysis. *Obstetrics and Gynecology*, 85(2): 304–313.

S54. Graham GD (2003). Tissue plasminogen activator for acute ischemic stroke in clinical practice: a meta-analysis of safety data. *Stroke*, 34(12): 2847–2850.

S55. Greenland S, Satterfield MH & Lanes SF (1998). A meta-analysis to assess the incidence of adverse effects associated with the transdermal nicotine patch. *Drug Safety*, 18(4): 297–308.

S56. Grodstein F, Newcomb PA & Stampfer MJ (1999). Postmenopausal hormone therapy and the risk of colorectal cancer: a review and meta-analysis. *American Journal of Medicine*, 106(5): 574–582.

S57. Grossman E & Messerli FH (1997). Effect of calcium antagonists on plasma norepinephrine levels, heart rate, and blood pressure. *American Journal of Cardiology*, 80(11): 1453–1458.

S58. Guise JM, Berlin M, McDonagh M, Osterweil P, Chan B & Helfand M (2004). Safety of vaginal birth after cesarean: a systematic review. *Obstetrics and Gynecology*, 103(3): 420–429.

S59. Gupta P & Sachdev HP (2003). Safety of oral use of nimesulide in children: systematic review of randomized controlled trials. *Indian Pediatrics*, 40(6): 518–531.

S60. Hart RG, Benavente O & Pearce LA (1999). Increased risk of intracranial hemorrhage when aspirin is combined with warfarin: a meta-analysis and hypothesis. *Cerebrovascular Diseases*, 9(4): 215–217.

S61. Hauth JC, Goldenberg RL, Parker CR, Cutter GR & Cliver SP (1995). Low-dose aspirin: lack of association with an increase in abruptio placentae or perinatal mortality. *Obstetrics and Gynecology*, 85(6): 1055–1058.

S62. He J, Whelton PK, Vu B & Klag MJ (1998). Aspirin and risk of hemorrhagic stroke: a meta-analysis of randomized controlled trials. *JAMA*, 280(22): 1930–1935.

S63. Hébert-Croteau N (1998). A meta-analysis of hormone replacement therapy and colon cancer in women. *Cancer Epidemiology, Biomarkers & Prevention*, 7(8): 653–659.

S64. Heisel O, Heisel R, Balshaw R & Keown P (2004). New onset diabetes mellitus in patients receiving calcineurin inhibitors: a systematic review and meta-analysis. *American Journal of Transplantation*, 4(4): 583–595.

S65. Henk JM (1997). Controlled trials of synchronous chemotherapy with radiotherapy in head and neck cancer: overview of radiation morbidity. *Clinical Oncology*, 9(5): 308–312.

S66. Hennessy S, Berlin JA, Kinman JL, Margolis DJ, Marcus SM & Strom BL (2001). Risk of venous thromboembolism from oral contraceptives containing gestodene and desogestrel versus levonorgestrel: a meta-analysis and formal sensitivity analysis. *Contraception*, 64(2): 125–133.

S67. Henry D & McGettigan P (2003). Epidemiology overview of gastrointestinal and renal toxicity of NSAIDs. *International Journal of Clinical Practice*, 135(Supplement): 43–49.
S68. Hoes AW, Grobbee DE, Peet TM & Lubsen J (1994). Do non-potassium sparing diuretics increase the risk of sudden cardiac death in hypertensive patients: recent evidence. *Drugs*, 47(5): 711–733.
S69. Huang JQ, Sridhar S & Hunt RH (1999). Gastrointestinal safety profile of nabumetone: a meta-analysis. *American Journal of Medicine*, 107(6A): 55S–61S.
S70. Humphrey LL, Takano LMA & Chan BKS (2002). *Postmenopausal Hormone Replacement Therapy and Cardiovascular Disease [Systematic evidence review no. 10]*. Rockville, MD, USA: Agency for Healthcare Research and Quality.
S71. Hwang BF & Jaakkola JJ (2003). Water chlorination and birth defects: a systematic review and meta-analysis. *Archives of Environmental Health*, 58(2): 83–91.
S72. Impicciatore P, Choonara I, Clarkson A, Provasi D, Pandolfini C & Bonati M (2001). Incidence of adverse drug reactions in paediatric in/out-patients: a systematic review and meta-analysis of prospective studies. *British Journal of Clinical Pharmacology*, 52(1): 77–83.
S73. Ismail AI & Bandekar RR (1999). Fluoride supplements and fluorosis: a meta-analysis. *Community Dentistry and Oral Epidemiology*, 27(1): 48–56.
S74. Ivanov R, Allen J & Calvin JE (2000). The incidence of major morbidity in critically ill patients managed with pulmonary artery catheters: a meta-analysis. *Critical Care Medicine*, 28(3): 615–619.
S75. Janowsky EC, Kupper LL & Hulka BS (2000). Meta-analysis of the relation between silicone breast implants and the risk of connective tissue diseases. *NEJM*, 342(11): 781–790.
S76. Johnson AG, Nguyen TV & Day RO (1994). Do nonsteroidal anti-inflammatory drugs affect blood pressure: a meta-analysis. *Annals of Internal Medicine*, 121(4): 289–300.
S77. Johnston SC, Colford JM, Jr. & Gress DR (1998). Oral contraceptives and the risk of subarachnoid hemorrhage: a meta-analysis. *Neurology*, 51(2): 411–418.
S78. Jolles BM & Bogoch ER (2006). Posterior versus lateral surgical approach for total hip arthroplasty in adults with osteoarthritis. *Cochrane Database of Systematic Reviews*, Issue 3, Art. No.: CD003828. DOI: 10.1002/14651858.CD003828.pub3.
S79. Jones G, Riley M, Couper D & Dwyer T (1999). Water fluoridation, bone mass and fracture: a quantitative overview of the literature. *Australian and New Zealand Journal of Public Health*, 23(1): 34–40.
S80. Jonker-Pool (2001). Sexual functioning after treatment for testicular cancer: review and meta-analysis of 36 empirical studies between 1975-2000. *Archives of Sexual Behavior*, 30(1): 55–74.

S81. Kellstein DE, Waksman JA, Furey SA, Binstok G & Cooper SA (1999). The safety profile of nonprescription ibuprofen in multiple-dose use: a meta-analysis. *Journal of Clinical Pharmacology*, 39(5): 520–532.

S82. Kemmeren JM, Algra A & Grobbee DE (2001). Third generation oral contraceptives and risk of venous thrombosis: meta-analysis. *BMJ*, 323(3705): 131–134.

S83. Khader YS, Rice J, John L & Abueita O (2003). Oral contraceptives use and the risk of myocardial infarction: a meta-analysis. *Contraception*, 68(1): 11–17.

S84. Ko DT, Hebert PR, Coffey CS, Sedrakyan A, Curtis JP & Krumholz HM (2002). Beta-blocker therapy and symptoms of depression, fatigue, and sexual dysfunction. *JAMA*, 288(3): 351–357.

S85. Ko DT, Hebert PR, Coffey CS, Curtis JP, Foody JM, Sedrakyan A & Krumholz HM (2004). Adverse effects of beta-blocker therapy for patients with heart failure: a quality overview of randomized trials. *Archives of Internal Medicine*, 164(13): 1389–1394.

S86. Kongnyuy EJ, Norman RJ, Flight IHK & Rees MCP (1999). Oestrogen and progestogen hormone replacement therapy for peri-menopausal and post-menopausal women: weight and body fat distribution. *Cochrane Database of Systematic Reviews*, Issue 3, Art. No.: CD001018. DOI: 10.1002/14651858.CD001018.

S87. Kozer E, Nikfar S, Costei A, Boskovic R, Nulman I & Koren G (2002). Aspirin consumption during the first trimester of pregnancy and congenital anomalies: a meta-analysis. *American Journal of Obstetrics and Gynecology*, 187(6): 1623–1630.

S88. Lazarou J, Pomeranz BH & Corey PN (1998). Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *JAMA*, 279(15): 1200–1205.

S89. LeBlanc ES, Janowsky J, Chan BK & Nelson HD (2001). Hormone replacement therapy and cognition: systematic review and meta-analysis. *JAMA*. 285(11): 1489–1499.

S90. Lee TM, Yip JT & Jones-Gotman M (2002). Memory deficits after resection from left or right anterior temporal lobe in humans: a meta-analytic review. *Epilepsia*, 43(3): 283–291.

S91. Lee A, Cooper MC, Craig JC, Knight JF & Keneally JP (2004). Effects of nonsteroidal anti-inflammatory drugs on postoperative renal function in adults with normal renal function. *Cochrane Database of Systematic Reviews*, Issue 2, Art No.: CD002765. DOI: 10.1002/14651858.CD002765.

S92. Leipzig RM, Cumming RG & Tinetti ME (1999). Drugs and falls in older people: a systematic review and meta-analysis. I. Psychotropic drugs. *Journal of the American Geriatrics Society*, 47(1): 30–39.

S93. Leipzig RM, Cumming RG & Tinetti ME (1999). Drugs and falls in older people: a systematic review and meta-analysis. II. Cardiac and analgesic drugs. *Journal of the American Geriatrics Society*, 47(1): 40–50.

S94. Lethaby A, Suckling J, Barlow D, Farquhar CM, Jepson RG & Roberts H (2004). Hormone replacement therapy in postmenopausal women: endometrial hyperplasia and irregular bleeding. *Cochrane Database of Systematic Reviews*, Issue 3, Art. No.: CD000402. DOI: 10.1002/14651858.CD000402.

S95. Lipworth BJ (1999). Systemic adverse effects of inhaled corticosteroid therapy: a systematic review and meta-analysis. *Archives of Internal Medicine*, 159(9): 941–955.

S96. Liu EH & Sia AT (2004). Rates of Caesarean section and instrumental vaginal delivery in nulliparous women after low concentration epidural infusions or opioid analgesia: systematic review. *BMJ*, 328(7453): 1410–1412.

S97. Loke YK, Derry S & Pritchard-Copley A (2002). Appetite suppressants and valvular heart disease: a systematic review. *BMC Clinical Pharmacology*, 2(6).

S98. MacLennan SC, MacLennan AH & Ryan P (1995). Colorectal cancer and oestrogen replacement therapy: a meta-analysis of epidemiological studies. *Medical Journal of Australia*, 162(9): 491–493.

S99. Marcolongo R, Frediani B, Biasi G, Minari C & Barreca C (1999). A meta-analysis of the tolerability of amtolmetin guacil, a novel, effective nonsteroidal anti-inflammatory drug, compared with established agents. *Clinical Drug Investigation*, 17(2): 89–96.

S100. Mardirosoff C, Dumont L, Boulvain M & Tramer MR (2002). Fetal bradycardia due to intrathecal opioids for labour analgesia: a systematic review. *BJOG: an International Journal of Obstetrics and Gynaecology*, 109(3): 274–281.

S101. McAlister FA, Clark HD, Wells PS & Laupacis A (1998). Perioperative allogeneic blood transfusion does not cause adverse sequelae in patients with cancer: a meta-analysis of unconfounded studies. *British Journal of Surgery*, 85(2): 171–178.

S102. Mehta SR, Eikelboom JW & Yusuf S (2000). Risk of intracranial haemorrhage with bolus versus infusion thrombolytic therapy: a meta-analysis. *Lancet*, 356(9228): 449–454.

S103. Miller J, Chan BK & Nelson H (2002). *Hormone Replacement Therapy and Risk of Venous Thromboembolism. [Systematic evidence review no. 11]*. Rockville, MD, USA: Agency for Healthcare Research and Quality.

S104. Moiniche S, Romsing J, Dahl JB & Tramer MR (2003). Nonsteroidal antiinflammatory drugs and the risk of operative site bleeding after tonsillectomy: a quantitative systematic review. *Anesthesia and Analgesia*, 96(1): 68–77.

S105. Mol BW, Ankum WM, Bossuyt PM & Van der Veen F (1995). Contraception and the risk of ectopic pregnancy: a meta-analysis. *Contraception*, 52(6): 337–341.

S106. Muldoon MF, Manuck SB, Mendelsohn AB, Kaplan JR & Belle SH (2000). Cholesterol reduction and non-illness mortality: meta-analysis of randomised clinical trials. *BMJ*, 322(7277): 11–15.

S107. Nalysnyk L, Fahrbach K, Reynolds MW, Zhao SZ & Ross S (2003). Adverse events in coronary artery bypass graft (CABG) trials: a systematic review and analysis. *Heart*, 89(7): 767–772.

S108. Nanda K, Bastian LA, Hasselblad V & Simel DL (1999). Hormone replacement therapy and the risk of colorectal cancer: a meta-analysis. *Obstetrics and Gynecology*, 93(5 Part 2 Supplement S): 880–888.

S109. Nazareth I, Lewin J & King M (2001). Sexual dysfunction after treatment for testicular cancer: a systematic review. *Journal of Psychosomatic Research*, 51(6): 735–743.

S110. O'Connell D, Robertson J, Henry D & Gillespie W (1998). A systematic review of the skeletal effects of estrogen therapy in postmenopausal women. II. An assessment of treatment effects. *Climacteric*, 1(2): 112–123.

S111. Ofman JJ, MacLean CH, Straus WL, Morton SC, Berger ML, Roth EA & Shekelle P (2002). A metaanalysis of severe upper gastrointestinal complications of nonsteroidal antiinflammatory drugs. *Journal of Rheumatology*, 29(4): 804–812.

S112. Ofman JJ, Maclean CH, Straus WL, Morton SC, Berger ML, Roth EA & Shekelle PG (2003). Meta-analysis of dyspepsia and nonsteroidal antiinflammatory drugs. *Arthritis and Rheumatism*, 49(4): 508–518.

S113. Oger E & Scarabin PY (1999). Assessment of the risk for venous thromboembolism among users of hormone replacement therapy. *Drugs and Aging*, 14(1): 55–61.

S114. Olsen O (1997). Meta-analysis of the safety of home birth. *Birth*, 24(1): 4–13.

S115. Park-Wyllie L, Mazzotta P, Pastuszak A, Moretti ME, Beique L, Hunnisett L, Friesen MH, Jacobson S, Kasapinovic S, Chang D, Diav-Citrin O, Chitayat D, Nulman I, Einarson TR & Koren G (2000). Birth defects after maternal exposure to corticosteroids: prospective cohort study and meta-analysis of epidemiological studies. *Teratology*, 62(6): 385–392.

S116. Pfahlberg A, Hassan K, Wille L, Lausen B & Gefeller O (1997). Systematic review of case-control studies: oral contraceptives show no effect on melanoma risk. *Public Health Reviews*, 25(3-4): 309–315.

S117. Pladevall-Vila M, Delclos GL, Varas C, Guyer H, Brugues-Tarradellas J & Anglada-Arisa A (1996). Controversy of oral contraceptives and risk of rheumatoid arthritis: meta-analysis of conflicting studies and review of conflicting meta-analyses with special emphasis on analysis of heterogeneity. *American Journal of Epidemiology*, 144(1): 1–14.

S118. Raaijmakers E & Engelen AM (2002). Is sensorineural hearing loss a possible side effect of nasopharyngeal and parotid irradiation: a systematic review of the literature. *Radiotherapy and Oncology*, 65(1): 1–7.

S119. Ray JG, O'Brien TE & Chan WS (2001). Preconception care and the risk of congenital anomalies in the offspring of women with diabetes mellitus: a meta-analysis. *QJM: An International Journal of Medicine*, 94(8): 435–444.

S120. Reid FD, Mercer PM, Harrison M & Bates T (1996). Cholecystectomy as a risk factor for colorectal cancer: a meta-analysis. *Scandinavian Journal of Gastroenterology*, 31(2): 160–169.

S121. Reynolds F, Sharma SK & Seed PT (2002). Analgesia in labour and fetal acid-base balance: a meta-analysis comparing epidural with systemic opioid analgesia. *BJOG: an International Journal of Obstetrics and Gynaecology*, 109(12): 1344–1353.
S122. Robinson JW, Dufour MS & Fung TS (1997). Erectile functioning of men treated for prostate carcinoma. *Cancer*, 79(3): 538–544.
S123. Ross SD, Kupelnick B, Kumashiro M, Arellano FM, Mohanty N & Allen IE (1997). Risk of serious adverse events in hypertensive patients receiving isradipine: a meta-analysis. *Journal of Human Hypertension*, 11(11): 743–751.
S124. Ross SD, Akhras KS, Zhang S, Rozinsky M & Nalysnyk L (2001). Discontinuation of antihypertensive drugs due to adverse events: a systematic review and meta-analysis. *Pharmacotherapy*, 21(8): 940–953.
S125. Rothwell PM, Slattery J & Warlow CP (1996). A systematic comparison of the risks of stroke and death due to carotid endarterectomy for symptomatic and asymptomatic stenosis. *Stroke*, 27(2): 266–269.
S126. Rothwell PM, Slattery J & Warlow CP (1996). A systematic review of the risks of stroke and death due to endarterectomy for symptomatic carotid stenosis. *Stroke*, 27(2): 260–265.
S127. Ruesch S, Walder B & Tramer MR (2002). Complications of central venous catheters. Internal jugular versus subclavian access: a systematic review. *Critical Care Medicine*, 30(2): 454–460.
S128. Rutschmann OT, McCrory DC & Matchar DB (2002). Immunization and MS: a summary of published evidence and recommendations. *Neurology*, 59(12): 1837–1843.
S129. Sachdev M, Miller WC, Ryan T & Jolis JG (2002). Effect of fenfluramine-derivative diet pills on cardiac valves: a meta-analysis of observational studies.
American Heart Journal, 144(6): 1065–1073.
S130. Safdar N, Said A, Gangnon RE & Maki DG (2002). Risk of hemolytic uremic syndrome after antibiotic treatment of Escherichia coli 0157:H7 enteritis: a meta-analysis. *JAMA*, 288(8): 996–1000.
S131. Sagsveen M, Farmer JE, Prentice A & Breeze A (2003). Gonadotrophin-releasing hormone analogues for endometriosis: bone mineral density. *Cochrane Database of Systematic Reviews*, Issue 4, Art. No.: CD001297. DOI: 10.1002/14651858.CD001297.
S132. Sakai H, Hayashi K, Origasa H & Kusunoki T (1999). An application of meta-analysis techniques in the evaluation of adverse experiences with antihypertensive agents. *Pharmacoepidemiology and Drug Safety*, 8(3): 169–177.
S133. Salpeter SR, Ormiston TM, Salpeter EE & Wood-Baker R (2002). Cardioselective beta-blockers for reversible airway disease. *Cochrane Database of Systematic Reviews*, Issue 4, Art. No.: CD002992. DOI: 10.1002/14651858.CD002992.
S134. Salpeter SR, Greyber E, Pasternak GA & Salpeter EE (2003). Risk of fatal and nonfatal lactic acidosis with metformin use in type 2 diabetes mellitus. *Archives of Internal Medicine*, 163(21): 2594–2602.

S135. Salpeter SR, Walsh JM, Greyber E, Ormiston TM & Salpeter EE (2004). Mortality associated with hormone replacement therapy in younger and older women: a meta-analysis. *Journal of General Internal Medicine*, 19(7): 791–804.
S136. Salpeter SR, Ormiston TM & Salpeter EE (2005). Cardioselective beta-blockers for chronic obstructive pulmonary disease. *Cochrane Database of Systematic Reviews*, Issue 4, Art. No.: CD003566. DOI: 10.1002/14651858.CD003566.pub2.
S137. Salpeter SR, Greyber E, Pasternak GA & Salpeter EE (2006). Risk of fatal and nonfatal lactic acidosis with metformin use in type 2 diabetes mellitus. *Cochrane Database of Systematic Reviews*, Issue 1. Art. No.: CD002967.
DOI: 10.1002/14651858.CD002967.pub2.
S138. Schaumberg DA, Dana MR, Christen WG & Glynn RJ (1998). A systematic overview of the incidence of posterior capsule opacification. *Ophthalmology*, 105(7): 1213–1221.
S139. Schoenfeld P (1999). Gastrointestinal safety profile of meloxicam: a meta-analysis and systematic review of randomized controlled trials. *American Journal of Medicine*, 107(6A): 48S–54S.
S140. Serebruany VL, Malinin AI, Eisert RM & Sane DC (2004). Risk of bleeding complications with antiplatelet agents: meta-analysis of 338,191 patients enrolled in 50 randomized controlled trials. *American Journal of Hematology*, 75(1): 40–47.
S141. Seto A, Einarson T & Koren G (1997). Pregnancy outcome following first trimester exposure to antihistamines: meta-analysis. *American Journal of Perinatology*, 14(3): 119–124.
S142. Smith JS, Green J, Berrington de Gonzalez A, Appleby P, Peto J, Plummer M, Franceschi S & Beral V (2003). Cervical cancer and use of hormonal contraceptives: a systematic review. *Lancet*, 361(9364): 1159–1167.
S143. Sneyd JR, Carr A, Byrom WD & Bilski AJ (1998). A meta-analysis of nausea and vomiting following maintenance of anaesthesia with propofol or inhalational agents. *European Journal of Anaesthesiology*, 15(4): 433–445.
S144. Stason WB, Schmid CH, Niedzwiecki D, Whiting GW, Caubet JF, Luo D, Ross SD & Chalmers TC (1997). Safety of nifedipine in patients with hypertension: a meta-analysis. *Hypertension*, 30(1 Part 1): 7–14.
S145. Stason WB, Schmid CH, Niedzwiecki D, Whiting GW, Caubet JF, Cory D, Luo D, Ross SD & Chalmers TC (1999). Safety of nifedipine in angina pectoris: a meta-analysis. *Hypertension*, 33(1): 24–31.
S146. Torgerson DJ & Bell-Syer SEM (2001). Hormone replacement therapy and prevention of nonvertebral fractures: a meta-analysis of randomized trials. *JAMA*, 285(22): 2891–2897.
S147. Torgerson DJ & Bell-Syer SEM (2001). Hormone replacement therapy and prevention of vertebral fractures: a meta-analysis of randomised trials. *BMC Musculoskeletal Disorders*, 2(7).

S148. Torvaldsen S, Roberts CL, Bell JC & Raynes-Greenow CH (2004). Discontinuation of epidural analgesia late in labour for reducing the adverse delivery outcomes associated with epidural analgesia. *Cochrane Database of Systematic Reviews*, Issue 4, Art. No.: CD004457. DOI: 10.1002/14651858.CD004457.pub2.

S149. Tramer MR, Moore RA, Reynolds DJ & McQuay HJ (2000). Quantitative estimation of rare adverse events which follow a biological progression: a new model applied to chronic NSAID use. *Pain*, 85(1-2): 169–182.

S150. Trindade E, Menon D, Topfer LA & Coloma C (1998). Adverse effects associated with selective serotonin reuptake inhibitors and tricyclic antidepressants: a meta-analysis. *Canadian Medical Association Journal*, 159(10): 1245–1252.

S151. Uzzan B, Campos J, Cucherat M, Nony P, Boissel JP & Perret GY (1996). Effects on bone mass of long term treatment with thyroid hormones: a meta-analysis. *Journal of Clinical Endocrinology and Metabolism*, 81(12): 4278–4289.

S152. Vamvakas EC (1995). Perioperative blood transfusion and cancer recurrence: meta-analysis for explanation. *Transfusion*. 35(9): 760–768.

S153. Vamvakas EC (2002). Meta-analysis of randomized controlled trials investigating the risk of postoperative infection in association with white blood cell-containing allogeneic blood transfusion: the effects of the type of transfused red blood cell product and surgical setting. *Transfusion Medicine Reviews*, 16(4): 304–314.

S154. Vamvakas EC (2003). WBC-containing allogeneic blood transfusion and mortality: a meta-analysis of randomized controlled trials. *Transfusion*, 43(7): 963–973.

S155. Vercauteren SB, Bosmans JL, Elseviers MM, Verpooten GA & De Broe ME (1998). A meta-analysis and morphological review of cyclosporine-induced nephrotoxicity in auto-immune diseases. *Kidney International*, 54(2): 536–545.

S156. Verhoeven AC & Boers M (1997). Limited bone loss due to corticosteroids: a systematic review of prospective studies in rheumatoid arthritis and other diseases. *Journal of Rheumatology*, 24(8): 1495–1503.

S157. von Dadelszen P, Ornstein MP, Bull SB, Logan AG, Koren G & Magee LA (2000). Fall in mean arterial pressure and fetal growth restriction in pregnancy hypertension: a meta-analysis. *Lancet*, 355(9198): 87–92.

S158. Vorperian VR, Havighurst TC, Miller S & January CT (1997). Adverse effects of low dose amiodarone: a meta-analysis. *Journal of the American College of Cardiology*, 30(3): 791–798.

S159. Warltier DCE, Marret E, Flahault A, Samama CM & Bonnet F (2003). Effects of postoperative, nonsteroidal, antiinflammatory drugs on bleeding risk after tonsillectomy: meta-analysis of randomized, controlled trials. *Anesthesiology*, 98(6): 1497–1502.

S160. Waterman EJ, Magee LA, Lim KI, Skoll A, Rurak D & von Dadelszen P (2004). Do commonly used oral antihypertensives alter fetal or neonatal heart rate characteristics: a systematic review. *Hypertension in Pregnancy*, 23(2): 155–169.

S161. Wehling M (2002). Meta-analysis of flecainide safety in patients with supraventricular arrhythmias. *Arzneimittel-Forschung Drug Research*, 52(7): 507–514.
S162. Wiffen P, Gill M, Edwards J & Moore A (2002). Adverse drug reactions in hospital patients: a systematic review of the prospective and retrospective studies. *Bandolier Extra*: 1–14.
S163. Wilson K & Mottram P (2004). A comparison of side effects of selective serotonin reuptake inhibitors and tricyclic antidepressants in older depressed patients: a meta-analysis. *International Journal of Geriatric Psychiatry*, 19(8): 754–762.
S164. Xiong X, Buekens P & Wollast E (1995). IUD use and the risk of ectopic pregnancy: a meta-analysis of case-control studies. *Contraception*, 52(1): 23–34.
S165. Yaffe K, Sawaya G, Lieberburg I & Grady D (1998). Estrogen therapy in postmenopausal women: effects on cognitive function and dementia. *JAMA*, 279(9): 688–695.
S166. Zaric DDZ, Christiansen CCC, Pace NL & Punjasawadwong Y (2005). Transient neurologic symptoms (TNS) following spinal anaesthesia with lidocaine versus other local anaesthetics. *Cochrane Database of Systematic Reviews*, Issue 4, Art. No.: CD003006. DOI: 10.1002/14651858.CD003006.pub2.

Appendix B: References associated with specific features in adverse events meta-analyses reviewed in Chapter 6

Meta-analysis feature	**References (numbers refer to references in Appendix A)**
Drug intervention(s)	S2; S3; S6; S7; S8; S9; S11; S12; S19; S20; S21; S22; S23; S26; S27; S28; S29; S31; S34; S35; S36; S37; S38; S40; S41; S42; S43; S44; S45; S46; S47; S48; S49; S50; S51; S52; S53; S54; S55; S56; S57; S59; S60; S61; S62; S63; S64; S66; S67; S68; S69; S70; S72; S73; S76; S77; S81; S82; S83; S84; S85; S86; S87; S88; S89; S91; S92; S93; S94; S95; S97; S98; S99; S102; S103; S104; S108; S110; S111; S112; S113; S115; S116; S117; S123; S124; S129; S130; S131; S132; S133; S134; S135; S136; S137; S139; S140; S141; S142; S144; S145; S146; S147; S149; S150; S151; S155; S156; S157; S158; S159; S160; S161; S162; S163; S165
Surgical intervention(s)	S10; S13; S14; S15; S24; S33; S58; S78; S90; S107; S120; S125; S126; S138
More than 10 meta-analyses	S1; S3; S5; S12; S13; S18; S22; S23; S24; S25; S27; S30; S33; S37; S38; S41; S45; S47; S49; S50; S51; S53; S55; S57; S62; S63; S64; S67; S70; S75; S76; S77; S79; S80; S81; S82; S83; S84; S85; S86; S87; S90; S91; S92; S93; S94; S96; S101; S104; S107; S110; S124; S125; S129; S131; S132; S133; S134; S135; S140; S143; S144; S145; S146; S148; S149; S150; S151; S152; S154; S157; S161; S162; S163
2–5 meta-analyses	S2; S4; S6; S8; S10; S14; S16; S19; S26; S28; S31; S32; S34; S35; S36; S42; S46; S48; S52; S54; S59; S60; S65; S72; S73; S78; S89; S97; S98; S103; S106; S109; S111; S113; S114; S115; S120; S121; S122; S128; S138; S139; S147; S155; S159; S160; S164; S166
One meta-analysis	S9; S11; S17; S21; S39; S58; S74; S116; S130; S141
Study type: trials only	S1; S4; S6; S7; S9; S10; S12; S16; S20; S23; S24; S28; S31; S35; S36; S38; S40; S41; S42; S43; S46; S47; S52; S55; S57; S59; S60; S61; S62; S64; S65; S69; S74; S76; S78; S81; S84; S85; S86; S91; S94; S96; S99; S10; S10; S104; S106; S110; S112; S123; S124; S127; S128; S131; S132; S133; S135; S136; S139; S140; S143; S144; S145; S146; S147; S148; S150; S153; S154; S155; S157; S158; S159; S160; S161; S163; S166
Study type: observational	S8; S11; S14; S19; S21; S22; S26; S27; S29; S33; S34; S44; S45; S48; S49; S50; S51; S54; S56; S58; S66; S67; S70; S71; S72; S73; S75; S77; S79; S80; S82; S83; S88; S89; S90; S92; S93; S97; S105; S109; S113; S114; S115; S116; S117; S118; S119; S120; S125; S126; S142; S151; S152; S162; S164; S165
Study type: mixed trials and observational studies	S13; S15; S17; S25; S30; S32; S37; S39; S53; S63; S87; S95; S98; S101; S103; S107; S108; S111; S121; S129; S130; S134; S137; S138; S149; S156

Continued on next page

– continued from previous page	
Meta-analysis feature	**References (numbers refer to references in Appendix A)**
Forest plots	S1; S6; S10; S14; S16; S18; S19; S20; S23; S24; S25; S26; S27; S29; S30; S31; S32; S34; S39; S41; S42; S44; S45; S47; S49; S51; S54; S59; S61; S62; S63; S66; S70; S71; S74; S78; S79; S82; S83; S86; S87; S89; S91; S92; S93; S94; S96; S97; S98; S100; S101; S102; S103; S104; S105; S110; S114; S115; S116; S119; S120; S121; S125; S126; S127; S128; S129; S130; S131; S132; S133; S134; S135; S136; S137; S139; S143; S146; S147; S148; S151; S153; S154; S155; S158; S159; S163; S166
Other plots	S2; S3; S4; S5; S8; S11; S12; S17; S28; S35; S40; S46; S48; S57; S68; S69; S75; S76; S80; S106; S113; S117; S123; S124; S138; S140; S144; S145; S149; S165
No graphical results	S7; S9; S13; S15; S21; S22; S33; S36; S37; S38; S43; S50; S52; S53; S55; S56; S58; S60; S65; S72; S73; S77; S81; S84; S85; S88; S90; S99; S107; S108; S109; S111; S112; S122; S141; S150; S152; S156; S157; S160; S161; S164
Academic sponsorship	S1; S5; S6; S10; S13; S16; S18; S20; S21; S22; S23; S30; S31; S32; S34; S36; S38; S39; S42; S46; S48; S50; S52; S57; S59; S65; S69; S71; S73; S74; S77; S78; S79; S82; S83; S84; S85; S90; S91; S92; S93; S95; S96; S97; S104; S108; S109; S113; S115; S118; S119; S120; S121; S122; S128; S130; S131; S132; S133; S134; S135; S137; S140; S141; S145; S146; S152; S153; S154; S155; S157; S158; S159; S163; S165; S166
Commercial sponsorship	S2; S3; S7; S11; S26; S35; S49; S53; S55; S60; S64; S66; S67; S72; S81; S87; S99; S107; S111; S112; S123; S124; S139; S143; S144; S147; S161
Government sponsorship	S8; S9; S14; S15; S17; S19; S28; S29; S33; S37; S40; S41; S43; S44; S47; S51; S54; S56; S58; S61; S62; S63; S70; S75; S76; S88; S89; S94; S100; S101; S102; S103; S105; S106; S110; S114; S116; S117; S125; S126; S127; S129; S138; S148; S149; S150; S151; S156; S160; S164
Other funding sources	S12; S24; S25; S27; S45; S68; S80; S86; S98; S136; S142; S162
Outcome metric: odds ratio	S7; S10; S11; S15; S20; S21; S22; S31; S32; S34; S35; S39; S40; S41; S52; S64; S65; S67; S71; S81; S82; S83; S87; S92; S93; S94; S96; S99; S100; S159; S102; S104; S105; S106; S109; S114; S115; S116; S120; S123; S125; S129; S130; S135; S139; S141; S143; S144; S145; S153; S154; S157; S158; S164; S165
Outcome metric: relative risk	S1; S6; S14; S16; S17; S24; S26; S27; S29; S30; S37; S42; S43; S44; S45; S46; S49; S50; S51; S53; S55; S56; S59; S60; S63; S66; S70; S74; S77; S78; S79; S84; S85; S89; S97; S98; S101; S103; S108; S112; S113; S117; S119; S127; S142; S146; S147; S148; S152; S163; S166
Outcome metric: risk difference	S23; S38; S61; S124; S128; S132; S150; S155
Outcome metric: mean difference (comparative)	S76; S86; S91; S121; S134; S137

Continued on next page

– continued from previous page	
Meta-analysis feature	**References (numbers refer to references in Appendix A)**
Outcome metric: standardised mean difference	S5; S9; S12; S19; S90; S110; S151
Outcome metric: percent difference (comparative)	S36; S156
Outcome metric: correlation	S2; S33
Outcome metric: probability or percent	S25; S28; S54; S58; S72; S80; S88; S118; S122; S126; S138; S140; S162
Outcome metric: mean difference (non-comparative)	S3; S68
Outcome metric: percent difference (non-comparative)	S4; S57; S95
Outcome metric: multiple	S13; S18; S47; S48; S62; S69; S73; S75; S107; S111; S131; S133; S136; S149; S160; S161
Meta-analysis method: Standard fixed effect	S5; S7; S10; S11; S21; S27; S28; S31; S39; S41; S47; S50; S52; S53; S56; S57; S60; S65; S67; S76; S78; S80; S81; S86; S87; S92; S93; S94; S97; S99; S102; S104; S105; S106; S113; S114; S115; S125; S126; S129; S130; S131; S133; S134; S136; S137; S141; S142; S143; S148; S151; S157; S159; S164
Meta-analysis method: other fixed effect	S2; S9; S18; S40; S54; S68; S122; S140; S162
Meta-analysis method: standard random effect	S3; S4; S12; S15; S16; S19; S23; S25; S26; S29; S30; S32; S33; S34; S35; S36; S43; S44; S51; S58; S59; S64; S72; S74; S77; S83; S84; S85; S88; S91; S96; S98; S107; S110; S112; S119; S121; S123; S124; S132; S135; S138; S146; S147; S152; S153; S154; S155; S163; S165
Meta-analysis method: marginal analysis	S118; S144; S145
Meta-analysis method: Bayesian methods	S17; S70; S89; S103; S108; S150
Meta-analysis method: multiple	S1; S6; S14; S20; S22; S24; S37; S42; S49; S55; S62; S63; S66; S69; S71; S73; S75; S79; S82; S100; S101; S109; S111; S116; S117; S120; S127; S149; S156; S158; S160; S161; S166

Continued on next page

– continued from previous page	
Meta-analysis feature	**References (numbers refer to references in Appendix A)**
Fixed effect model	S2; S5; S7; S9; S10; S11; S13; S17; S18; S21; S27; S28; S31; S37; S39; S40; S41; S43; S45; S47; S48; S50; S52; S53; S54; S56; S57; S60; S61; S62; S65; S67; S68; S69; S75; S76; S78; S80; S81; S86; S92; S93; S94; S97; S99; S102; S104; S105; S106; S113; S114; S115; S120; S122; S125; S126; S129; S130; S131; S133; S134; S136; S137; S140; S141; S142; S143; S148; S151; S157; S158; S159; S162; S164
Random effects model	S3; S4; S12; S14; S15; S16; S19; S23; S25; S26; S30; S32; S33; S34; S35; S36; S44; S51; S58; S59; S64; S70; S72; S74; S77; S83; S84; S85; S88; S89; S91; S96; S98; S107; S108; S110; S112; S119; S121; S123; S124; S132; S135; S138; S146; S147; S150; S152; S153; S154; S155; S163; S165
Both fixed and random effect(s)	S1; S6; S20; S22; S24; S29; S42; S49; S55; S63; S66; S71; S73; S79; S82; S100; S101; S103; S109; S111; S116; S117; S127; S149; S156; S160; S161; S166
Reason for chosen model: related to heterogeneity	S1; S3; S4; S6; S10; S14; S15; S19; S21; S22; S24; S25; S26; S29; S32; S33; S42; S43; S44; S45; S47; S49; S62; S63; S72; S73; S74; S78; S79; S83; S84; S87; S88; S91; S100; S112; S117; S127; S133; S135; S143; S148; S152; S155; S158; S166
Studies investigating dose–response	S2; S6; S9; S22; S23; S29; S31; S38; S44; S45; S46; S48; S49; S51; S53; S55; S56; S62; S67; S68; S79; S86; S94; S95; S110; S112; S123
Heterogeneity considered	S1; S2; S3; S4; S6; S8; S10; S14; S15; S16; S18; S19; S20; S21; S22; S23; S24; S26; S27; S29; S30; S31; S32; S33; S34; S35; S37; S39; S40; S41; S42; S43; S44; S45; S46; S47; S49; S50; S51; S52; S53; S55; S56; S58; S59; S60; S61; S62; S63; S64; S65; S66; S67; S69; S71; S72; S73; S74; S75; S77; S78; S79; S81; S82; S83; S84; S85; S86; S87; S88; S89; S91; S92; S93; S94; S95; S96; S97; S98; S99; S100; S101; S102; S103; S105; S106; S107; S108; S110; S111; S112; S113; S114; S115; S116; S117; S119; S120; S121; S123; S124; S125; S126; S127; S128; S129; S130; S131; S132; S133; S134; S135; S136; S137; S138; S139; S141; S142; S143; S144; S145; S146; S147; S148; S151; S152; S153; S154; S155; S156; S157; S158; S159; S160; S163; S164; S165; S166
Individual patient data included	S7; S27
Sparse data considered	S1; S11; S15; S16; S17; S20; S24; S25; S26; S30; S31; S32; S34; S35; S38; S39; S40; S41; S42; S46; S47; S52; S56; S59; S60; S61; S62; S64; S65; S74; S75; S78; S81; S84; S87; S94; S96; S97; S99; S100; S101; S103; S104; S106; S112; S114; S115; S119; S120; S125; S129; S133; S134; S135; S136; S137; S144; S145; S146; S147; S148; S159; S160; S163; S166

Continued on next page

– continued from previous page	
Meta-analysis feature	**References (numbers refer to references in Appendix A)**
Multiple outcomes considered	S1; S4; S5; S7; S8; S10; S12; S13; S16; S18; S20; S24; S25; S30; S32; S33; S34; S35; S38; S40; S41; S43; S45; S47; S50; S52; S53; S54; S55; S56; S57; S59; S61; S62; S63; S64; S65; S69; S70; S71; S75; S76; S78; S79; S80; S81; S84; S85; S86; S87; S88; S90; S91; S94; S95; S96; S97; S99; S100; S101; S102; S104; S107; S108; S109; S110; S114; S115; S119; S120; S121; S123; S125; S126; S127; S128; S129; S131; S132; S133; S134; S135; S136; S137; S139; S140; S143; S144; S145; S146; S148; S149; S150; S151; S152; S154; S156; S157; S158; S159; S160; S161; S163; S165
Subgroups included	S15; S22; S27; S28; S38; S47; S49; S51; S62; S64; S76; S79; S82; S83; S92; S93; S99; S107; S120; S125; S133; S135; S136; S142; S143; S146; S147; S151; S153; S154; S156; S157; S162
Dissemination bias considered	S3; S5; S6; S14; S17; S20; S21; S22; S26; S27; S28; S30; S32; S34; S35; S36; S40; S41; S43; S44; S45; S49; S51; S52; S53; S54; S55; S56; S59; S62; S63; S66; S67; S71; S72; S73; S75; S76; S79; S82; S83; S86; S87; S88; S92; S93; S94; S98; S99; S101; S102; S103; S108; S109; S111; S112; S114; S117; S119; S120; S121; S123; S124; S125; S126; S129; S130; S131; S132; S133; S134; S135; S136; S137; S139; S143; S146; S147; S150; S151; S152; S154; S155; S157; S159; S161; S163; S164; S166
Primary study quality considered	S1; S10; S14; S16; S19; S20; S23; S24; S25; S32; S35; S37; S38; S40; S41; S47; S48; S49; S58; S59; S64; S67; S70; S74; S76; S78; S79; S86; S89; S90; S91; S92; S93; S94; S97; S100; S101; S103; S104; S108; S109; S110; S111; S112; S113; S114; S115; S117; S118; S123; S124; S128; S129; S131; S133; S134; S135; S136; S137; S144; S145; S146; S147; S148; S156; S157; S159; S163; S164; S166
Time-course aspects considered	S4; S23; S33; S35; S57; S64; S66; S89; S91; S94; S103; S104; S107; S114; S128; S131; S138; S154
Dose–response considered	S2; S6; S9; S22; S23; S29; S31; S38; S44; S45; S46; S48; S49; S51; S53; S55; S56; S62; S67; S68; S79; S86; S94; S95; S110; S112; S123
Class effects considered	S2; S12; S20; S28; S34; S36; S37; S38; S41; S42; S43; S44; S45; S47; S50; S51; S53; S56; S57; S63; S66; S67; S68; S70; S76; S77; S82; S83; S84; S85; S86; S89; S91; S92; S93; S94; S95; S97; S98; S100; S102; S103; S104; S105; S106; S108; S110; S111; S112; S113; S115; S116; S117; S124; S129; S130; S131; S132; S133; S135; S136; S139; S140; S141; S142; S146; S147; S149; S150; S151; S156; S159; S160; S163; S165
Publication bias: discussion only	S3; S5; S17; S26; S27; S28; S35; S36; S40; S43; S52; S55; S56; S59; S63; S71; S72; S82; S87; S88; S92; S93; S98; S99; S101; S108; S114; S119; S120; S124; S125; S126; S131; S132; S136; S139; S143; S147; S152; S154; S155; S157; S159; S161
Publication bias: quantitative analysis	S14; S22; S30; S34; S41; S44; S45; S49; S51; S53; S62; S66; S75; S79; S83; S103; S111; S117; S121; S129; S130; S133; S134; S135; S137; S146; S150; S151; S163; S164; S166

Continued on next page

– continued from previous page	
Meta-analysis feature	**References (numbers refer to references in Appendix A)**
Publication bias: test with *p*-value	S22; S30; S41; S45; S51; S53; S62; S66; S83; S111; S121; S129
Publication bias: funnel plot	S14; S22; S30; S31; S34; S41; S44; S45; S49; S51; S62; S75; S79; S83; S111; S117; S121; S129; S130; S133; S134; S135; S137; S146; S150; S151; S163; S164; S166
Published studies only included	S2; S4; S5; S8; S9; S10; S11; S12; S13; S14; S15; S16; S17; S18; S19; S20; S22; S23; S24; S25; S26; S29; S30; S31; S32; S33; S36; S37; S38; S39; S41; S42; S43; S44; S45; S46; S47; S48; S49; S51; S53; S54; S56; S57; S58; S59; S60; S62; S63; S64; S65; S66; S68; S69; S70; S71; S72; S73; S74; S75; S76; S77; S78; S79; S80; S81; S82; S83; S85; S87; S88; S89; S90; S91; S92; S93; S95; S97; S98; S101; S103; S104; S105; S107; S109; S113; S114; S115; S116; S117; S118; S119; S120; S122; S124; S125; S126; S127; S128; S129; S130; S131; S132; S133; S135; S136; S138; S139; S140; S141; S142; S143; S144; S145; S147; S148; S149; S150; S152; S153; S154; S156; S158; S160; S161; S162; S163; S164; S165
Published studies with unpublished data	S35; S40; S50; S61; S84; S86; S94; S96; S99; S100; S102; S106; S108; S110; S121; S151; S155; S157; S159; S166
Published and unpublished studies	S1; S6; S21; S27; S28; S34; S52; S55; S111; S112; S123; S134; S137; S146
Heterogeneity: quantitative assessment only	S1; S3; S4; S6; S8; S10; S15; S16; S20; S21; S22; S23; S24; S26; S27; S29; S30; S31; S32; S33; S34; S35; S37; S39; S40; S41; S42; S43; S46; S47; S49; S50; S51; S52; S53; S55; S56; S58; S59; S60; S62; S63; S64; S65; S66; S67; S69; S71; S73; S75; S77; S78; S79; S81; S82; S83; S84; S85; S87; S89; S91; S92; S93; S94; S96; S97; S98; S99; S100; S101; S102; S103; S105; S106; S108; S111; S112; S113; S114; S115; S116; S117; S119; S120; S121; S124; S125; S126; S127; S128; S130; S131; S132; S133; S134; S135; S136; S137; S138; S139; S141; S142; S143; S146; S147; S148; S151; S152; S153; S154; S155; S158; S159; S160; S163; S164; S165; S166
Heterogeneity: statistical test	S1; S3; S4; S6; S8; S10; S14; S15; S16; S19; S20; S21; S22; S23; S24; S26; S27; S29; S30; S32; S33; S34; S35; S37; S39; S40; S41; S42; S43; S44; S45; S46; S47; S49; S50; S51; S52; S53; S55; S56; S58; S59; S60; S62; S63; S64; S65; S66; S67; S69; S71; S73; S75; S77; S78; S79; S81; S82; S83; S84; S85; S86; S87; S89; S91; S92; S94; S97; S98; S99; S100; S101; S102; S103; S105; S106; S108; S111; S112; S113; S114; S115; S116; S117; S119; S120; S121; S124; S125; S126; S127; S128; S129; S130; S131; S132; S133; S134; S135; S136; S137; S138; S139; S141; S142; S143; S146; S147; S148; S151; S152; S153; S154; S155; S158; S159; S160; S163; S164; S165; S166
Heterogeneity test: *p*-value cut-off 0.05	S1; S6; S10; S24; S30; S33; S43; S45; S53; S55; S60; S71; S98; S105; S106; S108; S111; S112; S115; S120; S139; S143; S151; S152; S153; S158; S159; S164
Heterogeneity test: *p*-value cut-off 0.10	S15; S29; S40; S47; S50; S51; S63; S75; S77; S81; S82; S89; S91; S100; S101; S103; S117; S127; S137; S154; S160; S165; S166

Continued on next page

– continued from previous page	
Meta-analysis feature	**References (numbers refer to references in Appendix A)**
Heterogeneity test: *p*-value stated	S3; S8; S14; S16; S19; S20; S21; S22; S23; S27; S32; S44; S49; S56; S59; S62; S64; S65; S66; S67; S69; S73; S78; S79; S83; S86; S87; S92; S94; S97; S116; S121; S125; S126; S128; S129; S130; S131; S132; S133; S134; S135; S136; S138; S141; S142; S146; S147; S148; S155; S163
Heterogeneity estimate	S1; S16; S20; S26; S44; S47; S78; S91; S94; S131; S133; S136; S137; S148; S155; S166
Heterogeneity present	S1; S3; S4; S8; S14; S15; S16; S19; S20; S22; S23; S27; S29; S30; S32; S33; S34; S37; S40; S41; S42; S44; S45; S49; S50; S51; S53; S55; S56; S63; S65; S66; S67; S71; S72; S73; S74; S75; S77; S79; S81; S83; S84; S85; S87; S91; S92; S94; S96; S97; S98; S101; S105; S108; S111; S112; S114; S117; S120; S121; S126; S127; S129; S130; S131; S133; S134; S137; S138; S141; S142; S143; S151; S152; S153; S154; S155; S156; S158; S164; S165; S166
Subgroup analysis	S15; S22; S27; S47; S49; S51; S62; S64; S79; S82; S83; S92; S93; S99; S107; S120; S125; S133; S135; S136; S142; S143; S146; s147; S151; S153; S154; S156; S157
Meta-regression	S2; S3; S18; S30; S31; S40; S46; S51; S62; S72; S74; S79; S88; S95; S106; S107; S112; S117; S123; S124; S126; S132; S135; S138; S144; S145; S157
Qualitative investigation of heterogeneity	S8; S14; S15; S16; S19; S33; S58; S61; S65; S66; S71; S72; S92; S101; S110; S143; S156
Quality: two assessors	S14; S20; S24; S25; S37; S38; S58; S67; S74; S76; S78; S79; S89; S90; S92; S93; S103; S110; S111; S112; S113; S115; S117; S123; S124; S131; S133; S134; S135; S137; S144; S145; S146; S148; S159; S163; S166
Quality: assessment disagreements resolved by consensus	S14; S16; S19; S20; S37; S64; S78; S79; S89; S97; S101; S103; S104; S111; S112; S113; S123; S124; S133; S134; S137; S144; S145; S146; S159; S163
Quality tool used	S1; S10; S14; S16; S20; S23; S24; S25; S31; S35; S37; S40; S41; S47; S58; S59; S64; S67; S70; S74; S76; S78; S79; S89; S90; S91; S100; S101; S103; S104; S108; S109; S110; S111; S112; S113; S114; S115; S117; S118; S123; S124; S128; S131; S133; S134; S135; S137; S144; S145; S146; S148; S159; S163; S164; S166
Poorer quality studies excluded	S31; S48; S49; S58; S70; S89; S90; S108; S114; S118; S133; S159; S166
Subgroup analysis by study quality	S14; S24; S37; S40; S41; S67; S86; S91; S92; S93; S94; S101; S113; S135; S156; S163
Quality information used by other method	S19; S79; S110; S112; S115; S117; S123; S124; S144; S145; S157; S164
No use of quality information	S1; S10; S16; S20; S23; S25; S35; S38; S47; S59; S64; S74; S76; S78; S100; S103; S104; S109; S111; S128; S129; S131; S134; S136; S137; S146; S147; S148

Continued on next page

– continued from previous page	
Meta-analysis feature	**References (numbers refer to references in Appendix A)**
Two-arm studies with zero events	S1; S11; S16; S20; S24; S31; S32; S34; S40; S41; S46; S47; S52; S61; S62; S64; S74; S78; S81; S84; S94; S97; S99; S100; S104; S106; S115; S119; S125; S129; S133; S134; S135; S136; S137; S144; S145; S148; S159; S160; S166
Sparse data studies with odds ratio outcome	S11; S15; S20; S31; S32; S34; S35; S39; S40; S41; S52; S64; S65; S81; S87; S94; S96; S99; S100; S104; S106; S114; S115; S120; S125; S129; S135; S144; S145; S159
Sparse data studies with relative risk outcome	S1; S16; S17; S24; S26; S30; S42; S46; S56; S59; S60; S74; S78; S84; S97; S101; S103; S112; S119; S146; S147; S148; S163; S166
Continuity correction used	S11; S15; S30; S35; S39; S40; S52; S59; S62; S74; S100; S104; S106; S112; S114
Continuity correction not used (in conjunction with studies with sparse data)	S17; S20; S24; S31; S38; S60; S65; S81; S103; S120; S125; S133; S134; S136; S137; S144; S145;
Primary continuity correction: 0.5	S11; S15; S30; S35; S39; S40; S52; S62; S74; S100; S104; S106; S112; S114
Peto method in conjunction with sparse data	S20; S24; S31; S40; S41; S60; S94; S100; S104; S106; S120; S129
Studies with zero events in total included	S11; S34; S41; S61; S62; S74; S104; S106; S115; S119; S133; S134; S136; S137; S144; S145; S160
Studies with zero events in total excluded	S1; S16; S20; S24; S31; S32; S47; S64; S78; S81; S94; S97; S100; S125; S135; S148; S159; S166

Appendix C: Publications based on thesis

Peer-reviewed journal articles

Warren FC, Abrams KR, Sutton AJ & Bongartz T (In preparation). Network meta-analysis for adverse event data with applications to anti-TNF interventions in rheumatoid arthritis.

Warren FC, Abrams KR, Sutton AJ & Golder S (In preparation). A systematic review of meta-analyses where the primary outcome is an adverse or unintended event.

Bongartz T, Warren FC, Mines D, Matteson EL, Abrams KR & Sutton AJ (2009). Etanercept therapy in rheumatoid arthritis and the risk of malignancies: a systematic review and individual patient data meta-analysis of randomised controlled trials. Annals of the Rheumatic Diseases, 68(7), 1177-1183.

Poster presentations

Warren FC, Abrams, KR, Sutton AJ, Bongartz T & Matteson EL (2008). Development of evidence synthesis methods using hierarchical models to investigate influence of class effects and doseresponse: application to anti-TNF drugs for rheumatoid arthritis. Presented at RSS Conference, Nottingham, UK.

Warren FC, Abrams KR, Sutton AJ, Golder S & Ashby D (2007). Review of statistical methods used in the meta-analysis of adverse drugs reactions data. Presented at RSS Conference, York, UK.

Appendix D: Bongartz *et al*. 2009: publication in *Annals of the Rheumatic Diseases* based on Chapter 8

The publication cited below is set out in full in this Appendix, with permission from BMJ Publishing Group Ltd.

Bongartz T, Warren, FC, Mines D, Matteson EL, Abrams KR & Sutton AJ (2009). Etanercept therapy in rheumatoid arthritis and the risk of malignancies. A systematic review and individual patient data meta-analysis of Randomized Controlled Trials. *Annals of the Rheumatic Diseases*, 68(7): 1177–1183.

The website of the journal, *Annals of the Rheumatic Diseases*, is: http://ard.bmj.com/.

Etanercept therapy in rheumatoid arthritis and the risk of malignancies: a systematic review and individual patient data meta-analysis of randomised controlled trials

T Bongartz, F C Warren, D Mines, E L Matteson, K R Abrams and A J Sutton

Ann Rheum Dis 2009;68;1177-1183; originally published online 19 Nov 2008; doi:10.1136/ard.2008.094904

Etanercept therapy in rheumatoid arthritis and the risk of malignancies: a systematic review and individual patient data meta-analysis of randomised controlled trials

T Bongartz,[1] F C Warren,[2] D Mines,[3] E L Matteson,[1] K R Abrams,[2] A J Sutton[2]

[1] Division of Rheumatology and Department of Internal Medicine, Mayo Clinic College of Medicine, Rochester, Minnesota, USA; [2] Department of Health Sciences, University of Leicester, Leicester, UK; [3] Global Safety Surveillance and Epidemiology, Wyeth Research, Collegeville, Pennsylvania, USA

Correspondence to: Dr T Bongartz, Division of Rheumatology, Mayo Clinic, 200 1st Street SW, Rochester, MN 55905, USA; bongartz.tim@mayo.edu

Accepted 29 October 2008
Published Online First 18 November 2008

ABSTRACT

Purpose: Tumour necrosis factor (TNF) plays an important role in inflammation and may affect tumour growth control. To assess the risk of malignancy with etanercept, a fusion protein that inhibits TNF action, a meta-analysis was performed using individual patient data from randomised controlled trials (RCT) in patients with rheumatoid arthritis (RA).

Methods: A search was conducted of bibliographic databases, abstracts from annual meetings and any unpublished studies on file with manufacturers of etanercept to December 2006. Only RCT of etanercept used for 12 weeks or more in patients with RA were included. Nine trials met the inclusion criteria. To adjudicate endpoints, the case narratives of potential cases were reviewed. Patient-level data were extracted from the clinical trials databases.

Results: The nine trials included 3316 patients, 2244 who received etanercept (contributing 2484 person-years of follow-up) and 1072 who received control therapy (1051 person-years). Malignancies were diagnosed in 26 patients in the etanercept group (incidence rate (IR) 10.47/1000 person-years) and seven patients in the control group (IR 6.66/1000 person-years). A Cox's proportional hazards, fixed-effect model stratified by trial yielded a hazard ratio of 1.84 (95% CI 0.79 to 4.28) for the etanercept group compared with the control group.

Conclusions: In this analysis, the point estimate of malignancy risk was higher in etanercept-treated patients, although the results were not statistically significant. The approach of obtaining individual patient data of RCT in cooperation with trial sponsors allowed important insights into the methodological advantages and challenges of sparse adverse event data meta-analysis.

The question of whether the inhibition of tumour necrosis factor (TNF) alpha may increase the risk of malignant disease is still a matter of controversy.[1] Our previous aggregate data meta-analysis of randomised controlled trials (RCT) using anti-TNF antibodies for the treatment of patients with rheumatoid arthritis (RA) showed a significantly increased risk of malignancy in anti-TNF antibody-treated patients compared with control patients.[2 3] Etanercept, a fusion protein that is able to bind TNF and is also used in the treatment of RA, was deliberately excluded from this analysis due to differences in its molecular structure and mechanism of action within the anti-TNF class.[4] Etanercept is an anti-TNF receptor fusion protein with unique properties that distinguish it from the anti-TNF antibodies infliximab and adalimumab. In contrast to anti-TNF antibodies, etanercept also neutralises lymphotoxin alpha, which has been associated with tumour growth control independent of TNF activity.[5 6] The observation that etanercept is not beneficial in Crohn's disease[7] while anti-TNF antibodies are,[8 9] suggests distinct biological properties of the two classes of anti-TNF treatment.

The potential for assessing the safety of etanercept based on single RCT in RA is limited. These trials are valuable tools to assess a drug's efficacy but are limited in their assessment of safety. The sample size chosen on the basis of expected efficacy is usually insufficient to detect potential differences in sparse adverse events between treatment arms. Although observational studies offer a valuable approach to assess the risks of approved drugs, widespread use of a drug after approval for a significant amount of time is required to generate data that can be used for analysis. In addition, selection bias may be a limitation to safety assessments based on observational data.[10] Meta-analyses of RCT, in contrast, may reveal important safety signals early and mitigate the effect of selection bias.[11 12]

We sought to explore further a potential association between anti-TNF therapy and malignancies by performing a systematic review and individual patient data (IPD) meta-analysis of RCT using etanercept in patients with RA.

METHODS

This study was performed according to a protocol that prespecified study selection, eligibility criteria, data extraction and statistical analysis. The methodology was developed according to Cochrane collaboration guidelines (www.cochrane.org) and the manuscript was prepared in accordance with the QUOROM[13] statement.

Search strategy

Our search strategy was divided into two major steps: the first step included an electronic database review performed by a librarian who was blinded to the study hypothesis. EMBASE, Medline, Cochrane Library and Web of Science were searched from inception to December 2006 for RCT of etanercept in patients with RA, using the keyword terms: "arthritis rheumatoid"; "etanercept"; "Enbrel"; "tumour necrosis factor fusion protein"; "randomised controlled trial"; "random allocation"; "clinical trials phase II"; "clinical trials phase III"; "clinical trials phase IV".

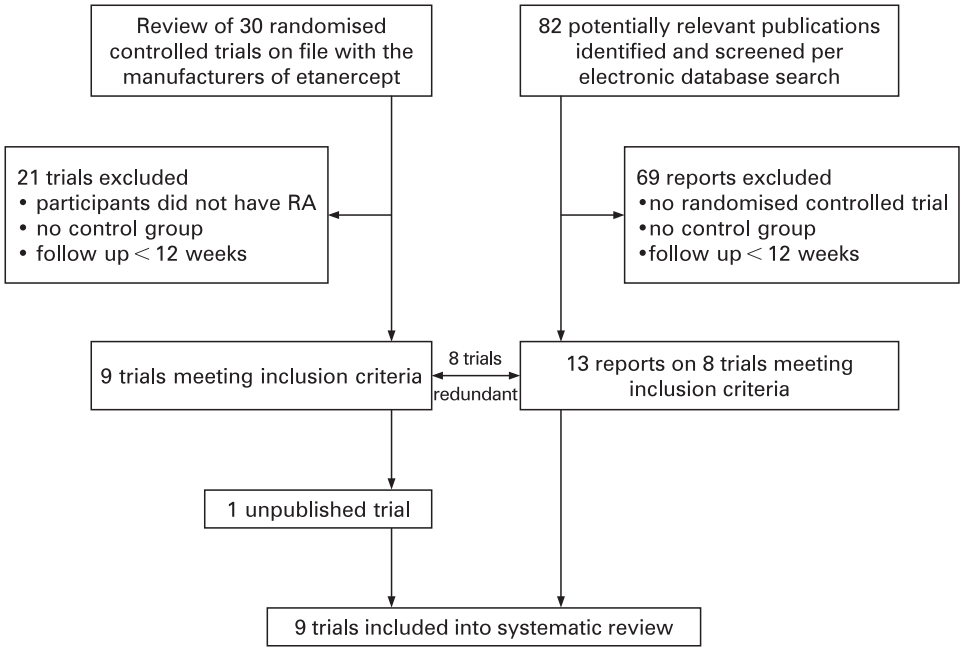


Figure 1 QUOROM-style flow diagram indicating selection of studies for this meta-analysis. RA, rheumatoid arthritis.

The second step encompassed direct communication with the manufacturers of etanercept, Wyeth and Amgen, in order to locate unpublished trials and/or published trials that were missed with the electronic database search.

Trial selection

Trials were included in our analysis if they met the following criteria: study participants were diagnosed with RA according to American College of Rheumatology criteria,[14] patients were randomly assigned to etanercept or control treatment and the study duration was at least 12 weeks. Assessment of eligibility criteria was performed independently by two investigators. Abstracts of all citations retrieved through the electronic database search were reviewed and potential candidates were further evaluated based on final publications. In addition, sponsors were contacted to obtain original trial protocols for in-depth review.

Study quality assessment

All included trials were reviewed for methodological features most relevant to issues of bias. Two independent reviewers assessed randomisation, random allocation concealment, masking of allocation, intent-to-treat analysis, completeness of follow-up, outcome assessment and attrition using original study protocols provided by sponsors as well as published reports. Disagreements were resolved by consensus-forming discussions.

Data extraction

The primary outcome of our analysis was (first) incident cancer, defined as a disease characterised by abnormal cells that divide without control and have the ability to invade other tissues. This definition did not include carcinoma in situ. Three investigators independently adjudicated potential malignancies based on a review of adverse event case narratives from which information about treatment assignment had been removed. Disagreements were resolved by consensus-forming discussions.

After the assessment of case narratives was completed, study sponsors provided data for every patient who participated in trials selected for the meta-analysis: demographic information; treatment assignment; dose of study drug; date of first and last dose of study drug; time point and reason of premature study discontinuation; date of last follow-up and concomitant disease-modifying antirheumatic drug therapy.

Data synthesis

All patients from eligible trials who were randomly assigned and received at least one dose of the study drug were included in the analysis (one patient who was lost to follow-up on the day of the first dose was excluded). The risk window for incident malignancies began with the date of the first dosing of study drug to the date of last follow-up in the respective RCT. A survival analysis of time-to-first-event using a Cox's proportional hazards model stratified by trial and assuming a fixed treatment effect was performed. In addition, a meta-analysis of study-level hazard ratios (HR) based on a random effects model (an approximation of a Cox's proportional hazards model using a Poisson generalised linear model) was conducted.

Sensitivity analyses entailed omitting cancers diagnosed within 6 weeks of trial entry and omitting all non-melanoma skin cancers (NMSC) from case definition. To evaluate any potential duration response, we conducted separate analyses for three non-overlapping periods of follow-up time (<6 months, 6–12 months, >24 months). As a secondary analysis, we performed an aggregate data-based meta-analyses using study-level odds ratios (OR). In contrast to the primary analysis, which uses a time-to-event approach, this analysis used the number of randomly assigned patients as the denominator of the incidence measure. For this secondary data synthesis, Mantel–Haenszel methods were used with a continuity correction inversely proportional to the relative size of the other treatment arm for that study.[15]

All analyses were performed using Stata version 9.2, with the exception of the random effects survival model, which was performed using R version 2.5.0.

Role of the sponsor

This study was sponsored by Wyeth, who together with Amgen markets etanercept in North America. Wyeth and Amgen provided data for the analysis, and Wyeth provided payment to

Table 1 Characteristics of randomised controlled trials included in the meta-analysis

Trial/reference	Randomly assigned patients	Disease characteristics	Active treatment groups (n)	Control group (n)	Duration of trial
TNR 00102 Unpublished	158 (153)*	Active RA with inadequate response to MTX	Etanercept 10 mg biw (52)	Placebo (50)	12 Weeks
			Etanercept 25 mg biw (52)		
160004 Moreland *et al*, 1997[20]	180 (180) *	Active RA with inadequate response to ⩾1 DMARD	Etanercept 0.25 mg/m^2 biw (46)	Placebo (44)	12 Weeks
			Etanercept 2.0 mg/m^2 biw (46)		
			Etanercept 16.0 mg/m^2 biw (44)		
160009 Moreland *et al*, 1999[21]	246 (234) *	Active RA with inadequate response to ⩾1 DMARD	Etanercept 10 mg biw (76)	Placebo (80)	26 Weeks
			Etanercept 25 mg biw (78)		With extension up to 52 weeks
160012 Bathon *et al*, 2000[6]	654 (632)*	Active early RA <3 years (no previous MTX)	Etanercept 10 mg biw (208)	Placebo + MTX (217)	104 Weeks
			Etanercept 25 mg biw (207)		
Genovese *et al*, 2002[18]					
160014	89 (89)*	Active RA with inadequate response to MTX	Etanercept 25 mg biw + MTX (59)	Placebo + MTX (30)	24 Weeks
Weinblatt *et al*, 1999[23]					
160029	564 (534)*	Active RA and at least one comorbidity that increases the likelihood of infection	Etanercept 25 mg biw (266)	Placebo (269)	16 Weeks
Baumgartner *et al*, 2004[24] (abstract)					
0881300 Ericson and Wadjula, 1999[26] (abstract)	559 (558)*	Active RA with inadequate response to ⩾1 DMARD	Etanercept 10 mg qw (122)	Placebo (105)	12 Weeks
			Etanercept 25 mg qw (111)		
			Etanercept 10 mg biw (110)		
			Etanercept 25 mg biw (111)		
0881308 Klareskog *et al*, 2004[19]	686 (682)*	Active RA with inadequate response to DMARD other than MTX	Etanercept 25 mg biw (223)	MTX + placebo (228)	Approx 180 weeks
			Etanercept 25 mg biw + MTX (231)		
van der Heijde *et al*, 2006[22]					
van der Heijde *et al*, 2006[28] (abstract)					
Mola *et al*, 2006[27] (abstract)					
0881309 Combe *et al*, 2006[17]	260 (254)*	Active RA with inadequate response to sulfasalazine	Etanercept 25 mg biw + sufasalazine (101)	Sulfasalazine + placebo (50)	104 Weeks
			Etanercept 25 mg biw + placebo (103)		
Combe *et al*, 2005[25] (abstract)					

*Values in parentheses refer to the number of randomly assigned patients who received at least one allocated treatment dose. biw, twice weekly; DMARD, disease-modifying antirheumatic drug; MTX, methotrexate; qw, every week; RA, rheumatoid arthritis.

support the costs of study preparation, data analysis and manuscript preparation. The current meta-analysis, designed to evaluate the risk of malignancy only, arose in the context of a request to Wyeth from a regulatory agency regarding cancer risk. It was in this context that DM, a Wyeth employee and one of the study co-authors, approached TB and ELM in January 2006.

Both companies had the opportunity to comment on the study design and manuscript. However, all final decisions

Table 2 Summary of malignancies in the randomised controlled trials

	Etanercept-treated participants (N = 2244)			Control patients (N = 1072)	
Trial reference	Patients with ⩾1 malignancy/type of malignancy	Etanercept dose	Time point of diagnosis	Patients with ⩾1 malignancy/type of malignancy	Time point of diagnosis
TNR 00102 Unpublished	0	–	–	0	–
160004 Moreland *et al*, 1997[20]	0	–	–	0	–
160009 Moreland *et al*, 1999[21]	1 Lung adenocarcinoma	25 mg biw	Week 40	0	
160012 Bathon *et al*, 2000[16]	1 Ductal breast cancer	10 mg biw	Week 45	1 Non-melanoma skin cancer	Week 12
	1 Prostate adenocarcinoma	25 mg biw	Week 39	1 Urethral carcinoma	Week 56
Genovese *et al*, 2002[18]	1 Basal cell carcinoma	10 mg biw	Week 91	1 Colon adenocarcinoma	Week 40
	1 Lobular breast carcinoma	10 mg biw	Week 102	1 Basal cell carcinoma	Week 86
	1 Basal cell carcinoma	25 mg biw	Week 47		
	1 Basal cell carcinoma	25 mg biw	Week 44		
	1 Lung carcinoid	25 mg biw	Week 37		
	1 Basal cell carcinoma	20 mg biw	Week 40		
	1 Lung NSCC	10 mg biw	Week 9		
	1 Hodgkin lymphoma	25 mg biw	Week 37		
160014 Weinblatt *et al*, 1999[23]	0	–	–	0	–
160029 Baumgartner et al, 2004[24] (abstract)	1 Squamous cell carcinoma skin	25 mg biw	Week 2	1 Basal cell carcinoma	Week 3
	1 Basal cell carcinoma	25 mg biw	Week 9	1 Lung squamous cell carcinoma	Week 11
0881300				0	
Ericson and Wadjula, 1999[26] (abstract)	1 LGL syndrome (monoclonal)	10 mg biw	Week 1		
	1 Ductal breast carcinoma	10 mg qw	Week 4		
0881308 Klareskog *et al*, 2004[19]	1 Ductal breast carcinoma	25 mg biw	Week 34	1 Ductal breast carcinoma	Week 64
	1 Adenocarcinoma of the rectum	25 mg biw	Week 24		
van der Heijde *et al*, 2006[22]	1 Oesophageal adenocarcinoma	25 mg biw	Week 58		
van der Heijde *et al*, 2006[28] (abstract)	1 Basal cell carcinoma	25 mg biw	Week 36		
	1 Colon carcinoma	25 mg biw	Week 84		
Mola *et al*, 2006[27] (abstract)	1 Lung adenocarcinoma	25 mg biw	Week 64		
	1 Malignant melanoma	25 mg biw	Week 9		
	1 Basal cell carcinoma	25 mg biw	Week 45		
	1 Basal cell carcinoma	25 mg biw	Week 30		
	1 Endometrium carcinoma	25 mg biw	Week 106		
0881309 Combe *et al*, 2006[17] Combe *et al*, 2005[25] (abstract)	1 Acute myelogenous leukemia	25 mg biw	Week 12	0	–

biw, twice weekly; LGL, large granular lymphocyte; NSCC, non-small-cell carcinoma; qw, every week.

regarding study design, analysis, and reporting and interpretation of results rested with the academic investigators.

RESULTS

Trials included

A total of 82 publications was initially considered, from which 69 articles were excluded based on abstracts and content. Of the remaining 13 citations, eight full-text publications[16–23] and five poster abstracts[24–28] reported on eight RCT, which met our inclusion criteria. Figure 1 summarises the flow of eligible clinical trials into our analysis.

To ensure complete data acquisition, 30 RCT of etanercept on file with the manufacturers were assessed for eligibility based on a review of original study protocols. In addition to the eight trials already identified through the electronic database search, one unpublished study (study TNR 00102) that was eligible for analysis was identified. Therefore, nine RCT were finally included in our analysis.

Table 3 Effect of etanercept therapy on the occurrence of malignancies in randomised controlled trials

Dataset	Model	Events in etanercept group	Events in control group	HR (95% CI)	p Value
Full	Fixed effects survival model stratified by trial	26	7	1.84 (0.79 to 4.28)	0.16
Full	Random effects survival model stratified by trial	26	7	1.82 (0.78 to 4.22)	0.17
NMSC excluded	Fixed effects survival model stratified by trial	17	4	1.86 (0.62 to 5.59)	0.27
Cancers diagnosed within first 42 days excluded	Fixed effects survival model stratified by trial	23	6	1.87 (0.75 to 4.62)	0.18
<6 Months	Fixed effects survival model stratified by trial	8	3	1.52 (0.35 to 6.55)	0.99
6–12 Months	Fixed effects survival model stratified by trial	12	1	5.81 (0.73 to 46.16)	0.17
>12 Months	Fixed effects survival model stratified by trial	6	3	0.88 (0.21 to 3.66)	0.86
Full	Fixed effects survival model, treatment varying with ln(time)	26	7	0.97 (0.47 to 2.01)	0.93
Aggregate data	Fixed effects Mantel–Haenszel model	26	7	1.93 (0.85 to 4.38)	0.12
Aggregate data	Random effects DerSimonian and Laird model	26	7	1.71 (0.73 to 4.01)	0.21

HR, hazard ratio; NMSC, non-melanoma skin cancer.

Trial characteristics

The characteristics of included trials are displayed in table 1. Trial duration ranged from 12 to 180 weeks. Four trials extended beyond the observation period reported in the initial publication. For three of these trials (studies 0881309, 0881308, 160012), follow-up data were published in subsequent reports after initial publication. For one trial (study 160009), extension data were not published but were obtained from the sponsor.

All but one trial excluded patients with a history of cancer with less than a 5-year disease-free state (except NMSC). Trial 1881300 excluded patients who had a history of cancer at any time (except NMSC).

Based on the review of original study protocols, all trials were judged to be of high quality with appropriate randomisation, random allocation concealment and intent-to-treat analysis. Completeness of follow-up and attrition were assessed using IPD: 574 of 2244 (25.6%) in the anti-TNF treatment arms discontinued study treatment early compared with 455 of 1072 patients (42.4%) in the control arms. Common reasons for early discontinuation in etanercept-treated patients were adverse events (31.5%) and lack of efficacy (32.6%). Similarly, common reasons for early discontinuation in control patients were adverse events (25.1%) and lack of efficacy (46.6%). For the majority of patients who discontinued study treatment prematurely, follow-up beyond the date of treatment discontinuation was available: 90.8% in the etanercept arms versus 92.7% in the control arms.

All trials were sponsored by Wyeth or Amgen.

Patients

A total of 3396 participants was randomly assigned in the nine trials we assessed. Eighty individuals were excluded from further analysis, 79 of whom had never received the allocated treatment and one who was lost to follow-up immediately after the first dose of study drug. Our data for analysis comprised 2244 participants who received etanercept (contributing 2484 person-years of follow-up) and 1072 participants who received control therapy (contributing 1051 person-years of follow-up). Dataset validation revealed eight patients who participated in two trials. A total of 3308 separate individuals thus generated a denominator of 3316 participants. None of these patients who transferred between trials had an incident malignancy.

Malignancies

Twenty-six patients with incident malignancies were identified in the treatment groups (incidence rate (IR) 10.47/1000 person-years) and seven patients in the control groups (IR 6.66/1000 person-years). A detailed summary of all incident malignancies is given in table 2. In three trials (TNR 00102, 160004 and 160014), no incident malignancies were observed.

For one trial (study 160012), three additional incident NMSC were identified based on individual case narratives when compared with the original publication, which did not report on this type of malignancy.

Data synthesis

Combined analysis according to our primary model (IPD survival analysis) yielded an HR of 1.84 (95% CI 0.79 to 4.28) for malignancies in patients using etanercept compared with control treatment. Using a random effects model resulted in a similar estimate, with an HR of 1.82 (95% CI 0.78 to 4.22).

For methodological comparison, an aggregate data meta-analysis was performed. When applying Mantel–Haenszel methods, the OR for malignancies in patients using etanercept compared with patients receiving control treatment was 1.93 (95% CI 0.85 to 4.38), using a continuity correction according to Sweeting *et al*.[15] The results of using a random effects DerSimonian and Laird model were very similar (HR 1.71; 95% CI 0.73 to 4.01), using the same continuity correction, reflecting the observation that between-trial heterogeneity using I-squared was 0.0%.

Additional analyses

Four malignancies were diagnosed during the first 6 weeks after the first treatment dose. As these cancers were likely to be present yet undetected when patients began the trial, we excluded these four patients as part of our sensitivity analysis.

With these exclusions, the HR for malignancies in patients treated with etanercept compared with the non-etanercept group was 1.87 (95% CI 0.75 to 4.62).

In the light of recent observational data,[29–31] which suggest a significantly increased risk of NMSC (but not other solid malignancies in patients treated with anti-TNF therapy), we decided to exclude as events all NMSC from our analysis. Using this approach, the results were essentially unchanged (HR 1.86; 95% CI 0.62 to 5.59).

To investigate whether there are any particular time periods in which etanercept treatment is associated with an increased incidence of cancer, the dataset was stratified according to three different time points: 0–6 months; 6–12 months and more than 12 months. This analysis did not reveal a time period in which the risk of cancer was significantly increased. We also performed an exploratory analysis (stratified fixed effects model) of the effect of dose, categorising etanercept dosing regimens into two groups, less than 50 mg/week and 50 mg/week or greater (or 25 mg twice weekly). The lower dosing range accounted for only 21.2% of etanercept follow-up time. Compared with the control arms, the relative risk estimates were similar for each dosing range (HR for the higher dose group was 1.92; 95% CI 0.80 to 4.62), and for the lower dose group it was 1.59 (95% CI 0.49 to 5.09), both using comparator as reference (table 3).

Statistical power

We used a traditional sample size formula (log rank test, Freedman method)[32] for a single study to obtain a statistical power approximation (ie, this ignores stratification by study and differential follow-up times between studies) in this meta-analysis context. With a probability of seven malignancies per 1072 patients in the control group, it would require at least 9305 participants to detect a HR of 2.0 (statistical significance level of 5% and a power of 80%) in a large RCT, assuming 32% of patients are allocated to control—reflecting the situation in the existing studies. The number of individuals in our dataset (3316) was substantially lower.

Based on the numbers derived from the existing studies, the probability of detecting a doubling in the risk of malignancy (HR of 2.0) between the two groups, should such a difference exist, was 39%.

DISCUSSION

Our analysis found a higher incidence estimate of malignancies in etanercept as compared with placebo-treated patients, although the results are statistically not significant. Therefore, this study does not provide sufficient evidence to establish an association of malignancies and etanercept treatment. However, given the wide confidence interval of the effect measure (HR 1.84; 95% CI 0.79 to 4.28), it also cannot exclude a clinically meaningful association.

This meta-analysis provides important insight into methodological issues of an IPD meta-analysis of sparse adverse event data. Publication bias is usually viewed as one of the more prominent threats to the validity of a systematic review and meta-analysis. Our approach of obtaining IPD in cooperation with primary investigators and trial sponsors allowed us to review the complete collection of manufacturer-sponsored RCT for etanercept, making publication bias very unlikely. A review of inclusion and exclusion criteria of candidate studies, based on original study protocol review, resulted in a more reliable eligibility assessment. Furthermore, the IPD approach allowed us to include follow-up data that extended beyond the published period of follow-up.

Clinical trials of biological use in RA often show imbalances in the percentage of withdrawals between treatment and control groups, as a result of a higher rate of treatment failure in placebo-treated patients. This carries the theoretical risk of false estimates due to a higher loss to follow-up in the control groups and a longer exposure to the study treatment in the active treatment arm. A major benefit of our IPD approach was the ability to perform a time-to-event analysis. This allowed the censoring of patients who discontinued treatment early or were lost to follow-up, thereby removing them from the denominator. Of note, our aggregate data analysis yielded similar results. This strengthens the validity of an aggregate data approach of analysing sparse events in clinical trials of RA. Our analysis was robust to the application of a wide variety of statistical methods for data synthesis, reflecting that our results did not depend upon the assumptions of any one particular method.

The major weakness of our study is the lack of statistical power. Assessment of statistical power in a meta-analysis has been suggested but is rarely performed.[33–35] Our experience emphasises the importance of such an analysis to estimate the likelihood of missing an association given that it was true. The combination of several underpowered studies can still produce a meta-analysis that is inadequate to detect a clinically important effect size. Nonetheless, our meta-analysis does provide a more precise assessment of cancer risk than those available from the individual RA trials.

Comparing our results with published clinical data, the observed excess of malignancies in patients who received the TNF receptor fusion protein etanercept is not inconsistent with the results of a meta-analysis of anti-TNF antibody treatment in RA patients.[3] An updated version of this analysis[2] including 5788 patients yielded an OR of 2.4 (95% CI 1.2 to 4.8). The risk of malignancies in this study appeared to be more pronounced in patients who received higher doses of anti-TNF antibodies according to a subanalysis that stratified by dose. In the etanercept meta-analysis we did not observe clear evidence of a dose response, although the proportion of patients treated at doses lower than 50 mg per week was small and does not allow definite conclusions.

Of note, an RCT of etanercept in Wegener's granulomatosis[36] found a significantly increased risk of malignancy in the etanercept arm.

The results of two large observational studies[29 31] that included a greater number of patients and had longer follow-up did not replicate the overall increase of malignancies seen with the synthesis of RCT data. However, they did reveal a significantly increased risk of NMSC in anti-TNF-treated patients.

Contrasting results of trial data and observational data provides a valuable stimulus to explore further not only the central clinical question of a potential association of anti-TNF therapy and malignancies in particular, but also the methodological strengths and weaknesses of both methods for drug safety assessment in general. Even the best-designed observational studies may produce inaccurate answers, as differences in patient characteristics across treatment groups can seldom be perfectly controlled. In contrast, with successful randomisation in RCT, the baseline risk of a subsequent adverse event should be similar in all treatment groups, and whereas meta-analysis combines data across studies, stratification by study preserves the benefit of randomisation. Given sufficient trial evidence, a meta-analysis of RCT data may be able to detect potential drug hazards early, before observational data become available after a

drug is marketed. However, the ability of meta-analysis to assess sparse event data in randomised trials is often constrained by relatively short follow-up periods and finite cumulative trial enrolment, which translate to limited statistical power to detect differences between treatment groups for rare events.

Different strategies may be considered to improve statistical power in this context. Several RCT of anti-TNF treatment for indications such as ankylosing spondylitis, psoriatic arthritis, cardiomyopathy, inflammatory bowel disease and a variety of connective tissue disorders have been performed. Including all these trials in a comprehensive, IPD meta-analysis will have a higher chance of approaching an adequate sample size and delivering more precise estimates. As an additional or alternative step, different agents with a similar mode of action may be combined to improve statistical power, yet this gain may come at the price of validity, if the effects of study drugs are not homogeneous.

A large, comprehensive meta-analysis utilising these two steps to improve statistical power by including all three approved anti-TNF agents over a wide range of different indications has been requested by the European Medicine Evaluation Agency and is currently in progress.

Funding: This study was sponsored by Wyeth, who together with Amgen, markets etanercept in North America. Wyeth and Amgen provided data for the analysis, and Wyeth provided payment to support the costs of study preparation, data analysis and manuscript preparation.

Competing interests: Declared.
TB, FCW and AJS received grant support from Wyeth. ELM served as an investigator for Amgen, Biogen-IDEC, Centocor, Genentech, Hoffmann-LaRoche, Human Genome Sciences, Wyeth. He received grant support from Amgen, Centocor/Johnson & Johnson, Genentech, Mayo Foundation and Wyeth. He served as a consultant/on scientific advisory boards for Abbott, Amgen, BiogenIDEC and Centocor. DM is employed by Wyeth and owns stock options in the company. KRA received grant support from Wyeth. He served as a consultant to United BioSource Corporation (UBC) regarding a ''mixed treatment comparison'' project, which UBC has conducted for Bristol-Meyers Squibb in relation to rheumatoid arthritis.

REFERENCES

1. **Askling J.** Malignancy and rheumatoid arthritis. *Curr Rheumatol Rep* 2007;**9**:421–6.
2. **Bongartz T,** Matteson EL, Montori VM, Sutton AJ, Sweeting M, Buchan I. Risk of serious infections and malignancies with anti-TNF antibody therapy in rheumatoid arthritis—in reply. *JAMA* 2006;**296**:2203–4.
3. **Bongartz T,** Sutton AJ, Sweeting MJ, Buchan I, Matteson EL, Montori V. Anti-TNF antibody therapy in rheumatoid arthritis and the risk of serious infections and malignancies: systematic review and meta-analysis of rare harmful effects in randomized controlled trials. *JAMA* 2006;**295**:2275–85.
4. **Scallon B,** Cai A, Solowski N, Rosenberg A, Song XY, Shealy D, *et al*. Binding and functional comparisons of two types of tumor necrosis factor antagonists. *J Pharmacol Exp Ther* 2002;**301**:418–26.
5. **Dobrzanski MJ,** Reome JB, Hollenbaugh JA, Hylind JC, Dutton RW. Effector cell-derived lymphotoxin alpha and Fas ligand, but not perforin, promote Tc1 and Tc2 effector cell-mediated tumor therapy in established pulmonary metastases. *Cancer Res* 2004;**64**:406–14.
6. **Qin Z,** van Tits LJ, Buurman WA, Blankenstein T. Human lymphotoxin has at least equal antitumor activity in comparison to human tumor necrosis factor but is less toxic in mice. *Blood* 1995;**85**:2779–85.
7. **Sandborn WJ,** Hanauer SB, Katz S, Safdi M, Wolf DG, Baerg RD, *et al*. Etanercept for active Crohn's disease: a randomized, double-blind, placebo-controlled trial. *Gastroenterology* 2001;**121**:1088–94.
8. **Hanauer SB,** Sandborn WJ, Rutgeerts P, Fedorak RN, Lukas M, MacIntosh D, *et al*. Human anti-tumor necrosis factor monoclonal antibody (adalimumab) in Crohn's disease: the CLASSIC-I trial. *Gastroenterology* 2006;**130**:323–33; quiz 591.
9. **Hanauer SB,** Feagan BG, Lichtenstein GR, Mayer LF, Schreiber S, Colombel JF, *et al*. Maintenance infliximab for Crohn's disease: the ACCENT I randomised trial. *Lancet* 2002;**359**:1541–9.
10. **McPherson K,** Hemminki E. Synthesising licensing data to assess drug safety. *BMJ* 2004;**328**:518–20.
11. **Hemminki E,** McPherson K. Impact of postmenopausal hormone therapy on cardiovascular events and cancer: pooled data from clinical trials. *BMJ* 1997;**315**:149–53.
12. **Juni P,** Nartey L, Reichenbach S, Sterchi R, Dieppe PA, Egger M. Risk of cardiovascular events and rofecoxib: cumulative meta-analysis. *Lancet* 2004;**364**:2021–9.
13. **Moher D,** Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. *Onkologie* 2000;**23**:597–602.
14. **Altman R,** Alarcon G, Appelrouth D, Bloch D, Borenstein D, Brandt K, *et al*. The American College of Rheumatology criteria for the classification and reporting of osteoarthritis of the hip. *Arthritis Rheum* 1991;**34**:505–14.
15. **Sweeting MJ,** Sutton AJ, Lambert PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Stat Med* 2004;**23**:1351–75.
16. **Bathon JM,** Martin RW, Fleischmann RM, Tesser JR, Schiff MH, Keystone EC, *et al*. A comparison of etanercept and methotrexate in patients with early rheumatoid arthritis. *N Engl J Med* 2000;**343**:1586–93.
17. **Combe B,** Codreanu C, Fiocco U, Gaubitz M, Geusens PP, Kvien TK, *et al*. Etanercept and sulfasalazine, alone and combined, in patients with active rheumatoid arthritis despite receiving sulfasalazine: a double-blind comparison. *Ann Rheum Dis* 2006;**65**:1357–62.
18. **Genovese MC,** Bathon JM, Martin RW, Fleischmann RM, Tesser JR, Schiff MH, *et al*. Etanercept versus methotrexate in patients with early rheumatoid arthritis: two-year radiographic and clinical outcomes. *Arthritis Rheum* 2002;**46**:1443–50.
19. **Klareskog L,** van der Heijde D, de Jager JP, Gough A, Kalden J, Malaise M, *et al*. Therapeutic effect of the combination of etanercept and methotrexate compared with each treatment alone in patients with rheumatoid arthritis: double-blind randomised controlled trial. *Lancet* 2004;**363**:675–81.
20. **Moreland LW,** Baumgartner SW, Schiff MH, Tindall EA, Fleischmann RM, Weaver AL, *et al*. Treatment of rheumatoid arthritis with a recombinant human tumor necrosis factor receptor (p75)-Fc fusion protein. *N Engl J Med* 1997;**337**:141–7.
21. **Moreland LW,** Schiff MH, Baumgartner SW, Tindall EA, Fleischmann RM, Bulpitt KJ, *et al*. Etanercept therapy in rheumatoid arthritis. A randomized, controlled trial. *Ann Intern Med* 1999;**130**:478–86.
22. **van der Heijde D,** Klareskog L, Rodriguez-Valverde V, Codreanu C, Bolosiu H, Melo-Gomes J, *et al*. Comparison of etanercept and methotrexate, alone and combined, in the treatment of rheumatoid arthritis: two-year clinical and radiographic results from the TEMPO study, a double-blind, randomized trial. *Arthritis Rheum* 2006;**54**:1063–74.
23. **Weinblatt ME,** Kremer JM, Bankhurst AD, Bulpitt KJ, Fleischmann RM, Fox RI, *et al*. A trial of etanercept, a recombinant tumor necrosis factor receptor:Fc fusion protein, in patients with rheumatoid arthritis receiving methotrexate. *N Engl J Med* 1999;**340**:253–9.
24. **Baumgartner S,** Paulus H, Burch F, Kivitz A, Dunn M, Kerr D, *et al*. A study to determine the safety of etanercept (ENBREL) in patients with rheumatoid arthritis who have concomitant comorbid conditions. *Arthritis Rheum* 2004;**50**(Suppl 5):S660–1.
25. **Combe B,** Kvien TK, Fatenejad S, Wajdula J. A 2-year double-blind comparison of etanercept (enbrel (R)) and sulfasalazine, alone and combined in patients with active rhuematoid arthritis. *Arthritis Rheum* 2005;**52**(Suppl 5):S142.
26. **Ericson ML,** Wajdula J, European Etanercept Investigators. A double-blind, placebo-controlled study of the efficacy and safety of four different doses of etanercept in patients with rheumatoid arthritis. *Arthritis Rheum* 1999;**42**(Suppl 5):S82.
27. **Mola EM,** van der Heijde D, Melo Gomes J, Codreaunu C, Fatenejad S, Macpeek D. Sustained clinical efficacy and safety of combination therapy with etnaercept plus methotrexate in ra patients through 4 years: TEMPO trial extension results. *Ann Rheum Dis* 2006;**56**(Suppl 2):331.
28. **van der Heijde D,** Klareskog L, Baker P, Wajdula J, Fatenejad S. Etanercept combined with methotrexate is well tolerated for 3 years: results from the Trial of Etanercept and Methotrexate with Radiographic and Patient Outcomes (TEMPO). *Ann Rheum Dis* 2006;**65**:510.
29. **Wolfe F,** Michaud K. Biologic treatment of rheumatoid arthritis and the risk of malignancy: analyses from a large US observational study. *Arthritis Rheum* 2007;**56**:2886–95.
30. **Chakravarty EF,** Michaud K, Wolfe F. Skin cancer, rheumatoid arthritis, and tumor necrosis factor inhibitors. *J Rheumatol* 2005;**32**:2130–5.
31. **Askling J,** Fored CM, Brandt L, Baecklund E, Bertilsson L, Feltelius N, *et al*. Risks of solid cancers in patients with rheumatoid arthritis and after treatment with tumour necrosis factor antagonists. *Ann Rheum Dis* 2005;**64**:1421–6.
32. **Freedman LS.** Tables of the number of patients required in clinical trials using the logrank test. *Stat Med* 1982;**1**:121–9.
33. **Muncer S,** Taylor S, Craigie M. Power dressing and meta-analysis: incorporating power analysis into meta-analysis. *J Adv Nurs* 2002;**38**:274–80.
34. **Daya S.** Optimal information size. *Evidence-based Obstet Gynecol* 2002;**4**:53–5.
35. **Pogue J,** Yusuf S. Overcoming the limitations of current meta-analysis of randomised controlled trials. *Lancet* 1998;**351**:47–52.
36. **The Wegener's Granulomatosis Etanercept Trial (WGET) Research Group**. Etanercept plus standard therapy for Wegener's granulomatosis. *N Engl J Med* 2005;**352**:351–61.

Bibliography

Agnelli G, Eriksson BI, Cohen AT, Bergqvist D, Dahl OE, Lassen MR, Mouret P, Rosencher N, Andersson M, Bylock A, Jensen E & Boberg B (2009). Safety assessment of new antithrombotic agents: Lessons from the EXTEND study on ximelagatran. *Thrombosis Research*, 123(3): 488–497.

Als-Nielsen B, Chen W, Gluud C & Kjaeregard LL (2003). Association of funding and conclusions in randomized drug trials: A reflection of treatment effect or adverse events? *JAMA*, 290(7): 921–928.

Altman DG & Bland JM (2003). Statistics notes: Interaction revisited: the difference between two estimates. *BMJ*, 326(7382): 219.

Andersson M, Kamby C, Jensen M-B, Mouridsen H, Ejlertsen B, Dombernowsky P, Rose C, Cold S, Overgaard M, Andersen J & Kjær M (1999). Tamoxifen in high-risk premenopausal women with primary breast cancer receiving adjuvant chemotherapy. Report from the Danish Breast Cancer Co-operative Group DBCG 82B trial. *European Journal of Cancer*, 35(12): 1659–1666.

Ashby D & Hutton JL (1996). Bayesian epidemiology. In: *Bayesian biostatistics*. Berry DA & Stangl DK eds. New York, NY, USA: Marcel Dekker: pp 109–138.

Ashby D, Hutton JL & McGee MA (1993). Simple Bayesian analyses for case-control studies in cancer epidemiology. *Statistician*, 42(4): 385–397.

Begg CB & Pilote L (1991). A model for incorporating historical controls into a meta-analysis. *Biometrics*, 47(3): 899–906.

Bergman L, Beelen MLR, Gallee MPW, Hollema H, Benraadt J, van Leeuwen FE & Group CCCA (2000). Risk and prognosis of endometrial cancer after tamoxifen for breast cancer. *Lancet*, 356(9233): 881–887.

Berkey CS, Hoaglin DC, Antczak-Bouckoms A, Mosteller F & Colditz GA (1998). Meta-analysis of multiple outcomes by regression with random effects. *Statistics in Medicine*, 17(22): 2537–2550.

Berlin JA, Laird NM, Sacks HS & Chalmers TC (1989). A comparison of statistical methods for combining event rates from clinical trials. *Statistics in Medicine*, 8(2): 141–151.

Bongartz T, Warren, FC, Mines D, Matteson EL, Abrams KR & Sutton AJ (2009). Etanercept therapy in rheumatoid arthritis and the risk of malignancies. A systematic review and individual patient data meta-analysis of Randomized Controlled Trials. *Annals of the Rheumatic Diseases*, 68(7): 1177–1183.

Bongartz T, Sutton AJ, Sweeting MJ, Buchan I, Matteson EL & Montori V (2006). Anti-TNF antibody therapy in rheumatoid arthritis and the risk of serious infections and malignancies. Systematic review and meta-analysis of rare harmful effects in randomized controlled trials. *JAMA*, 295(19): 2275–2285.

Bonovas S, Filioussi K & Sitaras NM (2008). Statin use and the risk of prostate cancer: A metaanalysis of 6 randomized clinical trials and 13 observational trials. *International Journal of Cancer*, 123(4): 899–904.

Bonovas S, Filioussi K, Tsavaris N & Sitaras NM (2005). Use of statins and breast cancer: A meta-analysis of seven randomized clinical trials and nine observational studies. *Journal of Clinicial Oncology*, 23(34): 8606–8612.

Borenstein M, Hedges LV, Higgins JPT & Rothstein HR (2009). *Introduction to Meta-analysis*. Chichester, UK: John Wiley & Sons, Ltd.

Bradburn MJ, Deeks JJ, Berlin JA & Localio AR (2007). Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. *Statistics in Medicine*, 26(1): 53–77.

Braithwaite RS, Chlebowski RT, Lau J, George S, Hess R & Col NF (2003). Meta-analysis of vascular and neoplastic events associated with tamoxifen. *Journal of General Internal Medicine*, 18(11): 937–947.

Breedveld FC, Weisman MH, Kavanaugh AF, Cohen SB, Pavelka K, van Vollenhoven R, Sharp J, Perez JL & Spencer-Green GT (2006). The PREMIER study: A multicenter, randomized, double-blind clinical trial of combination therapy with adalimumab plus methotrexate versus methotrexate alone or adalimumab alone in patients with early, aggressive rheumatoid arthritis who had not had previous methotrexate treatment. *Arthritis & Rheumatism*, 54(1): 26–37.

Breslow N (1981). Odds ratio estimator when data are sparse. *Biometrika*, 68(1): 73–84.

Brewster AM, Hortobagyi GN, Broglio KR, Kau S-W, Santa-Maria CA, Arun B, Buzdar AU, Booser DJ, Valero V, Bondy M & Esteva FJ (2008). Residual risk of breast cancer recurrence 5 years after adjuvant therapy. *Journal of the National Cancer Institute*, 100(16): 1179–1183.

Brookes ST, Whitely E, Egger M, Davey Smith G, Mulheran PA & Peters TJ (2004). Subgroup analyses in randomized trials: risks of subgroup-specific. *Journal of Clinical Epidemiology*, 57(3): 229–236.

Brooks SP & Gelman A (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4): 434–455.

Caldwell DM, Ades AE & Higgins JP (2005). Simultaneous comparison of multiple treatments: combining direct and indirect evidence. *BMJ*, 331(7521): 897–900.

Carlin JB (1992). Meta-analysis for 2 x 2 tables - a Bayesian approach. *Statistics in Medicine*, 11(2): 141–158.

Chalmers TC, Smith HJ, Blackburn B, Silverman B, Schroeder B, Retiman D & Ambroz A (1981). A method for assessing the quality of a randomized controlled trial. *Controlled Clinical Trials*, 2(1): 31–49.

Chou R & Helfand M (2005). Challenges in systematic reviews that assess treatment harms. *Annals of Internal Medicine*, 142(12 Pt 2): 1090–1099.

Clayton D & Hills M (1993). *Statistical models in epidemiology*. Oxford, UK: Oxford University Press.

Cochran WG (1954). The combination of estimates from different experiments. *Biometrics*, 10(1): 101–129.

Combe B, Codreanu C, Fiocco U, Gaubitz M, Geusens PP, Kvien TK, Pavelka K, Sambrook PN, Smolen JS, Wajdula J & Fatenejad S (2006). Etanercept and sulfasalazine, alone and combined, in patients with active rheumatoid arthritis despite receiving sulfasalazine: a double-blind comparison. *Annals of the Rheumatic Diseases*, 65(10): 1357–1362.

Congdon P (2006). *Bayesian Statistical Modelling*. Chichester, UK: John Wiley & Sons, Ltd.

Cooper NJ, Sutton AJ, Lu G & Khunti K (2006). Mixed comparison of stroke prevention treatments in individuals with nonrheumatic atrial fibrillation. *Archives of Internal Medicine*, 166(12): 1269–1275.

Cornelius VR, Perrio MJ, Shakir SAW & Smith LA (2009). Systematic reviews of adverse effects of drug interventions: a survey of their conduct and reporting quality. *Pharmacoepidemiology and Drug Safety*, 18(12): 1223-1231.

Deeks JJ, Higgins JPT & Altman DG (2008). Analysing data and undertaking meta-analyses. In: *Cochrane Handbook for Systematic Reviews of Interventions*. Higgins JPT & Green S eds. Chichester, UK: John Wiley & Sons, Ltd: pp 243–296.

Derry S, Kong Loke Y, Aronson JK (2001). Incomplete evidence: the inadequacy of databases in tracing published adverse drug reactions in clinical trials. *BMC Medical Research Methodology*, 1: 7.

DerSimonian R & Laird N (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3): 177–188.

Dominici F, Parmigiani G, Wolpert RL & Hasselblad V (1999). Meta-analysis of migraine headache treatments:combining information from heterogeneous designs. *Journal of the American Statistical Association*, 94(445): 16–28.

DuMouchel W (1995). Meta-analysis for dose-response data. *Statistics in Medicine.*, 14(5–7): 679–685.

Dunlop MG, Farrington SM, Carothers AD, Wyllie AH, Sharp L, Burn J, Liu B, Kinzler KW & Vogelstein B (1997). Cancer risk associated with germline DNA mismatch repair gene mutations. *Human molecular genetics*, 6(1): 105–110.

Early Breast Cancer Trialists' Collaborative Group (1998). Tamoxifen for early breast cancer: an overview of the randomised trials. *Lancet*, 351(9114): 1451–1467.

Eddy DM, Hasselblad V & R. S (1990). An introduction to a Bayesian method for meta-analysis: The confidence profile method. *Medical Decision Making*, 10(1): 15–23.

Edwards IR & Aronson JK (2000). Adverse drug reactions: definitions, diagnosis, and management. *Lancet*, 356(9237): 1255–1259.

Egger M, Davey Smith G & Altman DGE eds. (2001). *Systematic reviews in health care: meta-analysis in context, 2nd ed*. London, UK: BMJ Publishing Group.

Emerson JD (1994). Combining estimates of the odds ratio: the state of the art. *Statistical Methods in Medical Research*, 3(2): 157–178.

Fisher B, Anderson S, Tan-Chiu E, Wolmark N, Wickerham DL, Fisher ER, Dimitrov NV, Atkins JN, Abramson N, Merajver S, Romond EH, Kardinal CG, Shibata HR, Margolese RG & Farrar WB (2001). Tamoxifen and chemotherapy for axillary node-negative estrogen receptor-negative breast cancer: findings from National Surgical Adjuvant Breast and Bowel Project B-23. *Journal of Clinical Oncology*, 19(4): 931–942.

Fleiss JL (1993). The statistical basis of meta-analysis. *Statistical Methods in Medical Research*, 2(2): 121–145.

Friedman M (1982). Piecewise exponential models for survival data with covariates. *Annals of Statistics*, 10(1): 101–113.

Gart JJ (1970). Point and interval estimation of the common odds ratio in the combination of 2 x 2 tables with fixed marginals. *Biometrika*, 57(3): 471–475.

Gelman A, Carlin JB, Stern HS & Rubin DB (2004). *Bayesian Data Analysis*. Boca Raton, FL, USA: Chapman & Hall/CRC.

Gerber S, Tallon D, Trelle S, Schneider M, Jni P & Egger M (2007). Bibliographic study showed improving methodology of meta-analyses published in leading journals 1993-2002. *Journal of Clinical Epidemiology*, 60(8): 773–780.

Gibbons RD, Hur K, Bhaumik DK & Mann JJ (2006). The relationship between antidepressant prescription rates and rate of early adolescent suicide. *American Journal of Psychiatry*, 163(11): 1898–1904.

Glasziou PP & Irwig LM (1995). An evidence based approach to individualising treatment. *BMJ*, 311(7016): 1356–1359.

Golder S, Loke Y & McIntosh HM (2008). Poor reporting and inadequate searches were apparent in systematic reviews of adverse effects. *Journal of Clinical Epidemiology*, 61(5): 440–448.

Golder S, McIntosh HM, Duffy S, Glanville J & Centre for Reviews and Dissemination and UKCCSFDG (2006a). Developing efficient search strategies to identify reports of adverse effects in MEDLINE and EMBASE. *Health Information & Libraries Journal*, 23(1): 3–12.

Golder S, McIntosh HM & Loke Y (2006b). Identifying systematic reviews of the adverse effects of health care interventions. *BMC Medical Research Methodology*, 6: 22.

Golder S, Loke Y & McIntosh HM (2006c). Room for improvement? A survey of the methods used in systematic reviews of adverse effects. *BMC Medical Research Methodology*, 6: 3.

Greenland S & Salvan A (1990). Bias in the one-step method for pooling study results. *Statistics in Medicine*, 9(3): 247–252.

Gunnell D, Saperia J & Ashby D (2005). Selective serotonin reuptake inhibitors (SSRIs) and suicide in adults: a meta-analysis of drug company data from placebo controlled, randomised controlled trials submitted to the MHRA's safety review *BMJ*, 330(7488): 385.

Hallas J, Harvald B, Gram LF, Grodum E, Brosen K, Haghfelt T & Damsbo N (1990). Drug related hospital admissions: the role of definitions and intensity of data collection, and the possibility of prevention. *Journal of Internal Medicine*, 228(2): 83–90.

Hammad TA, Laughren T & Racoosin J (2006). Suicidality in pediatric patients treated with antidepressant drugs. *Archives of General Psychiatry*, 63(3): 332–339.

Hardy RJ & Thompson SG (1996). A likelihood approach to meta-analysis with random effects. *Statistics in Medicine*, 15(6): 619–629.

Hauck WW (1984). A comparative study of conditional maximum likelihood estimation of a common odds ratio. *Biometrics*, 40(4): 1117–1123.

Herbison P, Hay-Smith J & Gillespie WJ (2006). Adjustment of meta-analyses on the basis of quality scores should be abandoned. *Journal of Clinical Epidemiology*, 59(12): 1249–1256.

Hernández-Díaz S & García Rodríguez LA (2000). Association between non-steroidal anti-inflammatory drugs and upper gastrointestinal tract bleeding/perforation: an overview of epidemiologic studies published in the 1990s. *Archives of Internal Medicine*, 160(14): 2093–2099.

Higgins J & Whitehead A (1996). Borrowing strength from external trials in a meta-analysis. *Statistics in Medicine*, 15(24): 2733–2749.

Higgins JPT, Deeks JJ, Altman DG & eds. (2008). Special topics in statistics. In: *Cochrane handbook for systematic reviews of interventions*. Higgins JPT & Green S eds. Chichester, UK: John Wiley & Sons, Ltd: pp 481–529.

Higgins JPT & Thompson SG (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21(11): 1539–1558.

Higgins JPT, Thompson SG, Deeks JJ & Altman DG (2003). Measuring inconsistency in meta-analyses. *BMJ*, 327(7414): 557–560.

Higgins JPT, Thompson SG & Spiegelhalter DJ (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society Series A*, 172(1): 137–159.

Hill AB (1965). The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine*, 58(5): 295–300.

Hind D, Ward S, De Nigris E, Simpson E, Carroll C & Wyld L (2007). Hormonal therapies for early breast cancer: systematic review and economic evaluation, *Health Technology Assessment*, 11(26).

Ioannidis JP, Evans SJ, Gøtzsche PC, O'Neill RT, Altman DG, Schulz K, & Moher D. (2004). Better reporting of harms in randomized trials: an extension of the CONSORT statement. *Annals of Internal Medicine*, 141(10): 781–788.

Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJM, Gavaghan DJ & McQuay HJ (1996). Assessing the quality of reports of randomized clinical trials: Is blinding necessary? *Controlled Clinical Trials*, 17(1): 1–12.

Jha AK, Kuperman GJ, Teich JM, Leape L, Shea B, Rittenberg E, Burdick E, Seger DL, Vander Vliet M & Bates DW (1998). Identifying adverse drug events: development of a computer-based monitor and comparison with chart review and stimulated voluntary report. *Journal of the American Medical Informatics Association*, 5(3): 305–314.

Kaizar EE, Greenhouse JB, Seltman H & Kelleher K (2006). Do antidepressants cause suicidality in children? A Bayesian meta-analysis. *Clinical Trials*, 3(2): 73–98.

Keystone EC, Schiff MH, Kremer JM, Kafka S, Lovy M, De Vries T & Burge DJ (2004a). Once-weekly administration of 50 mg etanercept in patients with active rheumatoid arthritis. *Arthritis & Rheumatism*, 50(2): 353–363.

Klijn JGM, Beex LVAM, Mauriac L, van Zijl JA, Veyret C, Wildiers J, Jassem J, Piccart M, Burghouts J, Becqaurt D, Seynaeve C, Mignolet F & Duchateau L (2000). Combined treatment with buserelin and tamoxifen in premenopausal metastatic breast cancer: a randomized study. *Journal of the National Cancer Institute*, 92(11): 903–911.

Koopman L, van der Heijden GJMG, Glasziou PP, Grobbee DE & Rovers MM (2007). A systematic review of analytical methods used to study subgroups in (individual patient data) meta-analyses. *Journal of Clinical Epidemiology*, 60(10): 1002–1009.

Kvasz M, Allen IE, Gordon MJ, Ro EY, Estok R, Olkin I & Ross SD (2000). Adverse drug reactions in hospitalized patients: A critique of a meta-analysis. *Medscape General Medicine*, 2(2): E3.

Lambert PC, Sutton AJ, Abrams KR & Jones DR (2002). A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. *Journal of Clinical Epidemiology*, 55(1): 86–94.

Lambert PC, Sutton AJ, Burton PR, Abrams KR & Jones DR (2005). How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statistics in Medicine*, 24(15): 2401–2428.

Lazarou J, Pomeranz BH & Corey PN (1998). Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *JAMA*, 279(15): 1200–1205.

Leape LL, Kabcenell A, Berwick DM & Roessner J (1998). *Reducing Adverse Drug Events*. Boston, MA, USA: Institute of Health Improvement.

Leombruno JP, Einarson TR & Keystone EC (2008). The safety of anti-Tumor Necrosis Factor treatments in rheumatoid arthritis: meta and exposure adjusted pooled analyses of serious adverse events. *Annals of the Rheumatic Diseases*.

Li Z & Begg CB (1994). Random effects models for combining results from controlled and uncontrolled studies in a meta-analysis. *Journal of the American Statistical Association*, 89(428): 1523–1527.

Lindsey JK (1995). Fitting parametric counting processes by using log-linear models. *Journal of the Royal Statistical Society Series C*, 44(2): 201–212.

Loke YK, Derry S & Aronson JK (2004). A comparison of three different sources of data in assessing the frequencies of adverse reactions to amiodarone. *British Journal of Clinical Pharmacology*, 57(5): 616–621.

Loke YK, Price D & Herxheimer A (2008). Adverse effects. In: *Cochrane Handbook for Systematic Reviews of Interventions*. Higgins JPT & Green S eds. Chichester, UK: John Wiley & Sons, Ltd: pp 433–449.

Loke YK, Price D, Herxheimer A & Cochrane Adverse Effects Methods Group (2007). Systematic reviews of adverse effects: framework for a structured approach. *BMC Medical Research Methodology*, 7.

Lu G & Ades AE (2004). Combination of direct and indirect evidence in mixed treatment comparisons. *Statistics in Medicine*, 23(20): 3105–3124.

Lu G, Ades AE, Sutton AJ, Cooper NJ, Briggs AH & Caldwell DM (2007). Meta-analysis of mixed treatment comparisons at multiple follow-up times. *Statistics in Medicine*, 26(20): 3681–3699.

Ma R, Krewski D & Burnett RT (2003). Random effects Cox models: A Poisson modelling approach. *Biometrika*, 90(1): 157–169.

Maini RN, Breedveld FC, Kalden JR, Smolen JS, Furst D, Weisman MH, St. Clair EW, Keenan GF, van der Heijde D, Marsters PA & Lipsky PE (2004). Sustained improvement over two years in physical function, structural damage, and signs and symptoms among patients with rheumatoid arthritis treated with infliximab and methotrexate. *Arthritis & Rheumatism*, 50(4): 1051–1065.

Mantel N (1963). Chi-square tests with one degree of freedom; extensions of the Mantel–Haenszel procedure. *Journal of the American Statistical Association*, 58(303): 690–700.

Mantel N & Haenszel W (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst*, 22: 719–748.

McCullagh P & Nelder JA (1989). *Generalized Linear Models*. Boca Raton, FL, USA: Chapman & Hall/CRC.

McIntosh HM, Woolacott NF & Bagnall A-M (2004). Assessing harmful effects in systematic reviews. *BMC Medical Research Methodology*, 4.

McLeod RS (1999). Issues in surgical randomized controlled trials. *World Journal of Surgery*, 23(12): 1210–1214.

Mehta CR, Patel NR & Gray R (1985). Computing an exact confidence interval for the common odds ratio in several 2 x 2 contingency tables. *Journal of the American Statistical Association*, 80(392): 969–973.

Michel P, Quenon JL, de Sarasqueta AM & Scemama O (2004). Comparison of three methods for estimating rates of adverse events and rates of preventable adverse events in acute care hospitals. *BMJ*, 328(7433): 199.

Miller J, Carr A, Smith D, Emery S, Law MG, Grey P & Cooper DA (2000). Lipodystrophy following antiretroviral therapy for of primary HIV infection. *AIDS*, 14(15): 2406–2407.

Million Women Study Collaborators (2003). Breast cancer and hormone-replacement therapy in the Million Women Study. *Lancet*, 362(9382): 419–427.

Minelli C, Abrams KR, Sutton AJ & Cooper NJ (2004). Benefits and harms associated with hormone replacement therapy: clinical decision analysis. *BMJ*, 328(7436): 371.

Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D & Stroup DF (1999). Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. *Lancet*, 354(9193): 1896-1900.

Nebeker JR, Barach P, Samore MH, Nebeker JR, Barach P & Samore MH (2004). Clarifying adverse drug events: a clinician's guide to terminology, documentation, and reporting. *Annals of Internal Medicine*, 140(10): 795–801.

O'Neil AC, Petersen LA, Cook EF, Bates DW, Lee TH & Brennan TA (1993). Physician reporting compared with medical-record review to identify adverse events. *Annals of Internal Medicine*, 119(5): 370–376.

Peters JL, Sutton AJ, Jones DR, Rushton L, & Abrams KR (2006). A systematic review of systematic reviews and meta-analyses of animal experiments with guidelines for reporting. *Journal of Environmental Science & Health - Part B: Pesticides, Food Contaminants, & Agricultural Wastes*, 41(7): 1245–1258.

Pirmohamed M, James S, Meakin S, Green C, Scott AK, Walley TJ, Farrar K, Park BK & Breckenridge AM (2004). Adverse drug reactions as cause of admission to hospital: prospective analysis of 18 820 patients. *BMJ*, 329(7456): 15–19.

Plackett RL (1964). The continuity correction in 2 x 2 tables. *Biometrika*, 51(3/4): 327–337.

Pocock SJ (1997). Clinical trials with multiple outcomes: A statistical perspective on their design, analysis and interpretation. *Controlled Clinical Trials*, 18(6): 530–545.

Prevost TC, Abrams KR & Jones DR (2000). Hierarchical models in generalized synthesis of evidence: an example based on studies of breast cancer screening. *Statistics in Medicine*, 19(24): 3359–3376.

Riley RD, Lambert PC, Staessen JA, Wang J, Gueyffier F, Thijs L & Boutitie F (2008). Meta-analysis of continuous outcomes combining individual patient data and aggregate data. *Statistics in Medicine*, 27(11): 1870–1893.

Riley RD, Simmonds MC & Look MP (2007). Evidence synthesis combining individual patient data and aggregate data: a systematic review identified current practice and possible methods. *Journal of Clinical Epidemiology*, 60(5): 431–439.

Roberts KA, Dixon-Woods M, Fitzpatrick R, Abrams KR, & Jones DR (2002). Factors affecting uptake of childhood immunisation: a Bayesian synthesis of qualitative and quantitative evidence. *Lancet*, 360(9345): 1596–1599.

Rücker G, Schwarzer G, Carpenter J & Olkin I (2009). Why add anything to nothing? The arcsine difference as a measure of treatment effect in meta-analysis with zero cells. *Statistics in Medicine*, 28(5): 721–738.

Rutqvist LE, Johansson H, Signomklao T, Johannsson U, Fornander T & Wilking N (1995). Adjuvant tamoxifen therapy for early stage breast cancer and second primary malignancies. *Journal of the National Cancer Institute*, 87(9): 645–651.

Salanti G, Kavvoura FK & Ioannidis JPA (2008). Exploring the goemetry of treatment networks. *Annals of Internal Medicine*, 148(7): 544–553.

Schairer C, Mink PJ, Carroll L & Devesa SS (2004). Probabilities of death from breast cancer and other causes among female breast cancer patients. *Journal of the National Cancer Institute*, 96(17): 1311–1321.

Schulz KF, Chalmers I, Hayes RG & Altman DG (1995). Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA*, 273(5): 408–412.

Simpson EH (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society Series B*, 13(2): 238–241.

Simmonds MC, Higgins JPT, Stewart LA, Tierney JF, Clarke MJ & Thompson SG (2005). Meta-analysis of individual patient data from randomized trials: a review of methods used in practice. *Clinical Trials*, 2(3): 209–217.

Spiegelhalter DJ, Abrams KR & Myles JP (2004). *Bayesian Approaches to Clinical Trials and Healthcare Evaluation*. Chichester, UK: John Wiley & Sons, Ltd.

Spiegelhalter DJ & Best NG (2003). Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effectiveness modelling. *Statistics in Medicine*, 22(23): 3687–3709.

Spiegelhalter DJ, Best NG, Carlin BP & Van der Linde A (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B*, 64(4): 583–616.

Stewart HJ (1992). The Scottish trial of adjuvant tamoxifen in node-negative breast cancer. *Journal of the National Cancer Institute Monographs*, 11: 117–120.

Stewart HJ, Prescott RJ & Forrest PM (2001). Scottish Adjuvant Tamoxifen Trial: a randomized study updated to 15 years. *Journal of the National Cancer Institute*, 93(6): 456–462.

Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, Moher D, Becker BJ, Sipe TA & Thacker SB (2000). Meta-analysis of observational studies in epidemiology: A proposal for reporting. *JAMA*, 283(15): 2008–2012.

Sutton AJ & Abrams KR (2001). Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research*, 10(4): 277–303.

Sutton AJ, Abrams KR, Jones DR, Sheldon TA & Song F (2000). *Methods for meta-analysis in medical research*. Chichester, UK: John Wiley & Sons, Ltd.

Sutton AJ, Cooper NJ, Abrams KR, Lambert PC & Jones DR (2005). A Bayesian approach to evaluating net clinical benefit allowed for parameter uncertainty. *Journal of Clinical Epidemiology*, 58(1): 26–40.

Sutton AJ, Cooper NJ, Lambert PC, Jones DR, Abrams KR & Sweeting MJ (2002). Meta-analysis of rare and adverse event data. *Expert Review of Pharmacoeconomics & Outcomes Research*, 2(4): 367–379.

Sutton AJ, Kendrick D & Coupland CAC (2008). Meta-analysis of individual- and aggregate-level data. *Statistics in Medicine*, 27(5): 651–669.

Sweeting MJ, Sutton AJ & Lambert PC (2004). What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in Medicine*, 23(9): 1351–1375.

Takkouche B, Cadarso-Suárez C & Spiegelman D (1999). Evaluation of old and new tests of heterogeneity in epidemiologic meta-analysis. *American Journal of Epidemiology*, 150(2): 206–215.

Teicher MH, Glod C & Cole JO (1990). Emergence of intense suicidal preoccupation during fluoxetine treatment. *American Journal of Psychiatry*.

Tengs TO & Wallace A (2000). One thousand health-related quality-of-life estimates. *Medical Care*, 38(6): 583–637.

The ATAC (Arimidex, Tamoxifen Alone or in Combination) Trialists' Group (2002). Anastrozole alone or in combination with tamoxifen versus tamoxifen alone for adjuvant treatment of postmenopausal women with early breast cancer: first results of the ATAC randomised trial. *Lancet*, 359(9324): 2131–2139.

The EuroQoL Group (1990). EuroQoL - a new facility for the measurement of health-related quality of life. *Health Policy*, 16(3): 199–208.

the Uppsala Monitoring Centre (2000). Safety Monitoring of Medicinal Products: Guidelines for Setting Up and Running a Pharmacovigilance Centre, Uppsala: *the* Uppsala Monitoring Centre.

Tian L, Cai T, Pfeffer MA, Piankov N, Cremieux P-Y & Wei LJ (2009). Exact and efficient inference procedure for meta-analysis and its application to the analysis of independent 2 x 2 tables with all available data but without artificial continuity correction. *Biostatistics*, 10(2): 275–281.

Torrance GW (1986). Measurement of health state utilities for economic appraisal: A review. *Journal of Health Economics*, 5(1): 1–30.

Tudur Smith C, Williamson PR & Marson AG (2005a). Investigating heterogeneity in an individual patient data meta-analysis of time to event outcomes. *Statistics in Medicine*, 24(9): 1307–1319.

Tudur Smith C, Williamson PR & Marson AG (2005b). An overview of methods and empirical comparison of aggregate data and individual patient data results for investigating heterogeneity in meta-analysis of time-to-event outcomes. *Journal of Evaluation in Clinical Practice*, 11(5): 468–478.

Turner RM, Spiegelhalter DJ, Smith GCS & Thompson SG (2009). Bias modelling in evidence synthesis. *Journal of the Royal Statistical Society Series A*, 172(1): 21–47.

Tweedie RL & Mengersen KL (1995). Meta-analytic approaches to dose-response relationships, with application in studies of lung cancer and exposure to environmental tobacco smoke. *Statistics in Medicine*, 14(5–7): 545–569.

van der Heijde D, Klareskog L, Rodriguez-Valverde V, Codreanu C, Bolosiu H, Melo-Gomes J, Tornero-Molina J, Wajdula J, Pedersen R & Fatenejad S (2006). Comparison of etanercept and methotrexate, alone and combined, in the treatment of rheumatoid arthritis: two-year clinical and radiographic results from the TEMPO study, a double-blind randomized trial. *Arthritis & Rheumatism*, 54(4): 1063–1074.

Vincent C, Neale G & Woloshynowych M (2001). Adverse events in British hospitals: preliminary retrospective record review. *BMJ*, 322(7285): 517-519.

Walsh C & Mengersen K (2007). Model specification in hierarchical meta analysis. Dublin, Ireland, Department of Statistics, Trinity College.

Weingart SN, Wilson RM, Gibberd RW & Harrison B (2000). Epidemiology of medical error. *BMJ*, 320(7237): 774-777.

Wen J, Ren Y, Wang L, Li Y, Liu Y, Zhou M, Liu P, Ye L, Li Y & Tian W (2008). The reporting quality of meta-analyses improves: a random sampling study. *Journal of Clinical Epidemiology*, 61(8): 770–775.

Westhovens R, Yocum D, Han J, Berman A, Strusberg I, Geusens P & Rahman MU (2006). The safety of infliximab, combined with background treatments, among patients with rheumatoid arthritis and various comorbidities: a large randomized, placebo-controlled trial. *Arthritis & Rheumatism*, 54(4): 1075–1086.

Whitehead A (2002). *Meta-analysis of Controlled Clinical Trials*. Chichester, UK: John Wiley & Sons, Ltd.

Whitehead J (1980). Fitting Cox's regression model to survival data using GLIM. *Journal of the Royal Statistical Society Series C*, 29(3): 268–275.

Whitehead A & Whitehead J (1991). A general parametric approach to the meta-analysis of randomised clinical trials. *Statistics in Medicine*, 10(11): 1665–1677.

Wieland S & Dickersin K (2005). Selective exposure reporting and Medline indexing limited the search sensitivity for observational studies of the adverse effects of oral contraceptives. *Journal of Clinical Epidemiology*, 58(6): 560–567.

Wolfe F & Michaud K (2004). Lymphoma in rheumatoid arthritis. The effect of methotrexate and anti-tumor necrosis factor therapy in 18,572 patients. *Arthritis & Rheumatism*, 50(6): 1740–1751.

World Health Organization (1969). International drug monitoring: the role of the hospital, *World Health Organization Technical Report Series*, 425. Geneva: WHO.

Yates F (1934). Contingency tables involving small numbers and the χ^2 test. *Supplement to the Journal of the Royal Statistical Society*, 1(2): 217–235.

Zelen M (1971). The analysis of several 2 x 2 contingency tables. *Biometrika*, 58(1): 129–137.