

**GENERALISED SYNTHESIS METHODS IN HUMAN
HEALTH RISK ASSESSMENT**

Thesis submitted for the degree of
Doctor of Philosophy
At the University of Leicester

By

Jaime Louise Peters B.Sc., M.Sc.

Department of Health Sciences

University of Leicester

April 2006

UMI Number: U495644

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI U495644

Published by ProQuest LLC 2013. Copyright in the Dissertation held by the Author.
Microform Edition © ProQuest LLC.

All rights reserved. This work is protected against
unauthorized copying under Title 17, United States Code.



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

Generalised synthesis methods in human health risk assessment

Jaime Louise Peters B.Sc. M.Sc.

Abstract

This thesis critically explores the potential of systematic review and generalised synthesis methods to assist in human health risk assessments of exposure to chemicals in the environment. Current methods used to review and evaluate diverse, yet relevant, human and animal evidence for such risk assessments are described and shown to be lacking some degree of quantification, transparency and structure. Systematic review and generalised evidence synthesis methods demonstrate potential in overcoming some of these limitations.

In this thesis the use of systematic review and meta-analysis methods to evaluate evidence from animal experiments to inform human health related decisions is reviewed and results suggest that the quality of reporting of these reviews needs to be improved in order to make more efficient use of the animal evidence. To investigate the potential of systematic review and generalised synthesis methods in human health risk assessments, the methods are applied to two different examples where relevant evidence comes from human studies and animal experiments. In both examples, the use of systematic review methods was not only feasible, but provided structure and transparency in the identification and review process above that found in current risk assessments. In the first example hierarchical Bayesian models are used to synthesise evidence across species, leading to the derivation of a safe exposure limit. However, in the second example, the relevant data were much more diverse and sparse and so synthesis across species was of limited value. Nevertheless, synthesis of evidence within species assisted in evaluating the totality of relevant evidence. The quantitative framework of generalised evidence synthesis methods allows greater ease in assessments of between-study heterogeneity and publication bias. In this thesis the performance of methods to detect and adjust for publication bias has been assessed in scenarios likely to occur in a human health risk assessment context, and an alternative test has been proposed. However the findings suggest that caution is needed when carrying out and reporting results of an evidence synthesis especially when between-study heterogeneity and/or publication bias are suspected.

This thesis illustrates the potential of systematic review and generalised evidence synthesis methods in assisting human health risk assessments of environmental exposures and their ability to overcome some of the limitations of current methods. Although improvements in the quality of reporting of systematic reviews and meta-analyses of animal experiments need to be made, such an approach far outweighs current alternatives.

Acknowledgements

I would like to thank my supervisor Alex Sutton for his help, guidance and patience throughout my PhD. But, in particular, I am grateful for his continued enthusiasm. Thanks to Lesley Rushton for her help and encouragement, both during and since my time at IEH. I am also incredibly grateful to David Jones and Keith Abrams for their help, support and advice (though I still deny rejecting the idea of ever doing a PhD!). I would like to acknowledge the UK Department of Health for providing me with the opportunity to continue working towards my PhD. But finally, thanks to all my friends in and outside PRW and IEH, who have been so supportive and encouraging, in particular, mum, dad and Em.

Contents

1 Introduction.....	1
1.1 Thesis aims.....	1
1.2 Background.....	1
1.3 Risk assessment methods.....	3
1.4 Systematic reviews and meta-analyses.....	4
1.5 Bayesian methods for meta-analysis.....	5
1.6 Thesis overview.....	7
2 Risk assessment methods.....	11
2.1 Chapter overview.....	11
2.2 Introduction.....	12
2.3 Threshold substances.....	13
2.3.1 The critical effect.....	13
2.3.2 Uncertainty factors (UFs).....	16
2.3.3 The reference dose (RfD).....	18
2.4 Non-threshold substances.....	19
2.5 Risk assessments for occupational exposure to manganese (Mn).....	20
2.5.1 Occupational exposure to Mn.....	21
2.5.2 Reviewing the risk assessment documents.....	21
2.6 Risk assessments, systematic reviews and meta-analyses.....	26
3 Meta-analyses and generalised synthesis of evidence methods.....	29
3.1 Chapter overview.....	29
3.2 Systematic review methods.....	30
3.3 Meta-analysis methods.....	31
3.3.1 Fixed and random effects meta-analysis models.....	31
3.3.2 Between-study heterogeneity.....	33
3.3.3 Publication bias.....	34
3.3.4 Bayesian meta-analysis models.....	34

3.4 Generalised synthesis of evidence on environmental exposures.....	36
3.5 Other relevant research.....	39
3.5.1 Synthesis methods for toxicological data.....	39
3.5.2 Methods for generalised synthesis of human evidence.....	42
3.6 Summary.....	48
4 Systematic reviews and meta-analyses applied to animal experiments.....	52
4.1 Chapter overview.....	52
4.2 Introduction.....	53
4.3 Methods.....	54
4.3.1 Systematic review of the literature.....	54
4.3.2 Current guidelines for reporting systematic reviews and meta-analyses.....	55
4.4 Results.....	56
4.4.1 Articles identified.....	56
4.4.2 Features of the systematic reviews.....	59
4.4.3 Features of the meta-analysis.....	60
4.5 Quality reporting guidelines.....	67
4.5.1 Development of guidelines.....	67
4.5.2 Findings of the quality assessment.....	70
4.6 Summary.....	77
5 Trihalomethanes and low birth weight: example I.....	81
5.1 Chapter overview.....	81
5.2 Trihalomethane exposure and low birth weight.....	82
5.3 A systematic search of the literature.....	83
5.4 Obtaining study-specific estimates.....	86
5.5 Methods of synthesis.....	87
5.5.1 Synthesis of study-specific estimates within study type.....	89

5.5.2	Synthesis of study-specific estimates across study type.....	90
5.6	Results.....	92
5.6.1	Study-specific dose-response slope estimates.....	92
5.6.2	Within study type pooled dose-response slope estimates.....	92
5.6.3	Overall pooled dose-response slope estimates.....	95
5.7	Sensitivity analyses.....	97
5.7.1	Dose-response models.....	97
5.7.2	Sensitivity of assumptions on body weight, water consumption and low birth weight cut off values.....	100
5.7.3	Checking of model assumptions and sensitivity of prior distributions.....	106
5.7.4	Changing relevance of the animal data.....	119
5.8.	Interpretation of dose-response estimates.....	121
5.9.	Summary.	123
6	Manganese and neurobehavioural effects: example II.....	128
6.1	Chapter overview.....	128
6.2	Introduction.....	129
6.3	The systematic review.....	130
6.4	The evidence.....	133
6.4.1	Human epidemiological evidence.....	133
6.4.2	Animal experiment evidence.....	136
6.4.3	Summary.....	137
6.5	Synthesis of the evidence.....	138
6.5.1	Assessing the human evidence.....	139
6.5.2	Assessing the animal evidence.....	145
6.5.3	Summary of species-specific evidence.....	149
6.6	Further application of meta-analysis methods.....	152
6.6.1	Synthesis across species.....	152
6.6.2	Sensitivity analyses.....	158
6.7	Summary.....	162

7 Performance of tests and adjustments for publication bias.....	165
7.1 Chapter overview.....	165
7.2 Introduction.....	166
7.3 Methods for detecting publication bias.....	167
7.4 Review of related simulation studies.....	174
7.4.1 Rank correlation and regression tests.....	174
7.4.2 Trim and fill method.....	178
7.5 Simulations.....	182
7.5.1 Parameters.....	182
7.5.2 Regression models.....	190
7.5.3 Estimate of effect.....	193
7.5.4 The analyses.....	194
7.6 Results.....	196
7.6.1 The rank correlation test.....	196
7.6.2 The regression model tests.....	197
7.6.3 Estimates of effect - the trim and fill method.....	204
7.6.4 Estimates of effect – using the largest or most precise study.....	212
7.6.5 Estimates of effect – summary.....	214
7.7 Further analyses.....	215
7.7.1 Alternative parameters for the simulations.....	215
7.7.2 Relative risk as the measure of effect.....	221
7.8 Summary.....	224
8 Assessment of publication bias in the presence of between-study heterogeneity.....	228
8.1 Chapter overview.....	228
8.2 Motivating example.....	229
8.2.1 The Mapstone et al. (2003) meta-analysis.....	229
8.2.2 Assessment of publication bias.....	233
8.3 Simulation analyses.....	238
8.3.1 Parameters.....	238
8.3.2 Techniques for assessing publication bias.....	242
8.4 Results.....	244

8.4.1	Between-study heterogeneity.....	244
8.4.2	Naïve assessment of publication bias - the rank correlation test.....	245
8.4.3	Naïve assessment of publication bias - the regression tests.....	248
8.4.4	Accounting for predictable between-study heterogeneity – the regression tests.....	253
8.4.5	Summary and application to Mapstone <i>et al.</i> (2003) meta-analysis.....	258
8.5	Summary.....	261
9	Discussion.....	264
9.1	Thesis summary.....	264
9.2	Systematic reviews and meta-analyses of animal experiments...	265
9.3	Synthesis of human and animal evidence.....	267
9.4	Publication bias and between-study heterogeneity.....	270
9.5	Conclusions.....	273
	Bibliography.....	275
Appendix A:	Summary of the derivation of occupational exposure limits for manganese (Mn).....	307
Appendix B:	Search strategies for systematic reviews and meta-analyses of animal experiments.....	311
Appendix C:	List of references reviewed in Chapter 4	316
Appendix D:	Summaries of the meta-analyses of animal experiments.....	325
Appendix E:	Peters <i>et al.</i> (2005) JRSS C paper.....	340
Appendix F:	Search strategy for THMs articles.....	341
Appendix G:	Search strategy and list of included references for Mn articles.....	343
Appendix H:	Peters <i>et al.</i> (2006) JAMA paper.....	353
Appendix I:	Additional simulation results for Chapter 7.....	354
Appendix J:	Additional simulation results for Chapter 8.....	358

Introduction

1.1 Aims of thesis

In assessing risks to human health from exposure to chemical substances in the environment, relevant evidence comes from a variety of sources including human studies and animal experiments. Current risk assessment methods are not ideal due to their over-reliance on expert judgement and the lack of transparency in the methods used and the decisions made. Systematic review and meta-analysis methods have the potential to overcome some of the limitations of the current process. There has been limited exploration of methods for the quantitative synthesis of evidence from human and animal studies (DuMouchel and Harris, 1983; DuMouchel and Groër, 1989; Cox and Piegorsch, 1994). The aim of this thesis is to assess the potential use of systematic review and meta-analysis methods to assist in current human health risk assessments of exposure to chemical substances in the environment.

1.2 Background

In all aspects of life, humans are exposed to thousands of chemical substances. For example, carbon monoxide at home from the gas cooker, environmental tobacco smoke at the local pub and diesel exhaust on the walk to work. In fact there are around 100 000 chemical substances registered in the EU (Commission of the European Communities, 2001). 2 700 of these are categorised as *new* substances (available on the market since September 1981). Before they were marketed the *new* chemicals were subject to testing for human health risks (EEC, 1967).

However the remaining 97 000 or so chemicals are classed as *existing* substances (available before September 1981) and have not been subject to the same testing requirements as the *new* substances (Commission of the European Communities,

2001). Thus there are thousands of chemicals on the market, in our environment, about which little is known. It is therefore imperative that risk assessments are carried out for these chemicals to ensure safe human exposure.

Assessing the human health effects of exposure to chemical substances is a complex task. Although they would be ideal to avoid bias, randomized controlled trials (RCTs), where groups of humans are exposed to different doses of a chemical substance, are not usually carried out because of the obvious ethical issues of deliberately exposing humans to potentially lethal chemicals, inducing adverse health effects. However, some controlled human studies where volunteers are exposed to chemicals (e.g. chamber/booth studies) do exist, but tend to be based on a limited number of volunteers (Ernstgard *et al.*, 2005; Joffres *et al.*, 2005). Instead evidence for a risk assessment may come from controlled animal *in vivo* and *in vitro* experiments and observational human studies. Animal experiments have the advantage of being controlled: the exposure and response can be closely monitored. On the other hand, observational human studies may not be as controlled, but an advantage lies in their results providing information on the human experience of exposure to certain chemicals (unlike the animal experiments). Thus, evidence relevant for a risk assessment can come from diverse sources, including epidemiology and toxicology (WHO, 1994). It is likely that within the disciplines of epidemiology and toxicology, different methodologies will have been used (e.g. observational cohort and case-control studies in epidemiology; *in vitro* and *in vivo* studies in toxicology). There are also likely to be many differences between these studies beyond their design, such as the different routes of exposure (inhalation, oral, dermal), the different levels of exposure and the different species and strains of animals used.

Several methods to evaluate diverse evidence in a risk assessment are available and used throughout the world by governments and regulatory agencies. All of the methods require assumptions and judgements to be made during the risk assessment process. It is important that when setting limits for safe (or acceptable) human exposure to chemical substances (in order to avoid increased risk to human health) there is transparency regarding the various assumptions made. Moreover, those

setting, enforcing and using the limits must understand the nature and degree of uncertainty that exists in the estimation of these exposure limits.

1.3 Risk assessment methods

Methods for human health risk assessment for exposure to chemical substances in the environment are often based on subjective, narrative reviews of the evidence (including that from both human and animal studies) as described in Chapter 2.

When limits for human exposure to these chemicals need to be determined, they are often calculated from an animal experiment. The dose level from a particular study, observed to cause no or little adverse effects in animals is divided by a number of ‘uncertainty’ factors (UFs) to reflect a limit for human exposure (WHO, 1994; Woolley, 2003; Edler *et al.*, 2005). More complex biological processes can be used in risk assessment, particularly for chemicals believed to be carcinogens (RATSC, 1999a; Woolley, 2003; Watanabe, 2005), but these methods often require quite detailed evidence that is rarely available.

The risk assessment methods are not ideal. Current strategies to review and incorporate diverse evidence to inform risk assessments are generally not systematic and lack some degree of quantification (RATSC, 1999b). The nature and degree of uncertainty that exists in the estimation of these exposure limits needs to be clear to both setters and users of these limits. There also needs to be some harmonization of the different approaches used by various agencies and governments throughout the world (RATSC, 1999b). There is a real need for the development of a systematic, transparent methodology to combine human and animal data that could formally incorporate biological/mechanistic data, in so far as it exists (Budtz-Jørgensen *et al.*, 2001) and allow estimation of inherent uncertainty. Methods commonly used in medical research to systematically identify, review and evaluate human evidence on a particular intervention or policy, *systematic review* and *meta-analysis methods*, have the potential to help improve the objectivity and transparency of the current risk assessment process.

1.4 Systematic reviews and meta-analyses

Initially used in the fields of educational and psychological research (Hedges and Olkin, 1985), systematic reviews and meta-analyses are now commonly used in medical research, particularly to combine evidence from human RCTs. As the emphasis on evidence-based medicine increases, *systematic reviews* are used to compile, to assess the extent and quality of, and to summarise the results of research. The aim of a systematic review is to comprehensively evaluate the available evidence keeping potential biases to a minimum. This is achieved through the structured and transparent nature of such a review where the target research questions, methods and results are clearly laid out, making explicit assumptions and decisions made in execution. This also allows reproducibility and ease in updating the review (Sutton *et al.*, 2000; Egger *et al.*, 2001).

Where appropriate, *meta-analyses* extend the systematic review by quantitatively synthesising results from the relevant articles. Advantages of a meta-analysis include greater statistical power than that in a single study, the potential for more precise estimates, a framework for investigation of possible sources of heterogeneity between studies and the potential to be more generalisable (Fleiss and Gross, 1991; Blettner *et al.*, 1999). The pooled estimate from a meta-analysis is a weighted average of the results of the primary studies, where the weighting is dependent upon the precision of the estimate from each study (i.e. studies with large precision are given more weight in the meta-analysis). Fixed and random effects models are used in meta-analysis and are both described in Chapter 3. Meta-analysis is not just about obtaining an estimate of effect from multiple studies. Its quantitative framework allows investigation of *between-study heterogeneity* and *publication bias*. Between-study heterogeneity describes variability in the true underlying effects between studies in a meta-analysis (Higgins and Thompson, 2002). There may be many possible sources of between-study heterogeneity, for instance the way in which the exposure or health outcomes are measured may differ across studies. When it exists, between-study heterogeneity can affect the inference and conclusions of a meta-analysis and its possible sources must be investigated. Publication bias is the tendency for some studies to be less (or more) likely to be published and hence included in a systematic review because of the size and/or

statistical significance of their effect (Song *et al.*, 2000). If publication bias occurs, the subsequent systematic review or meta-analysis of published literature may be misleading. Therefore an assessment of publication bias should be undertaken before conclusions can be drawn from a meta-analysis. In Chapter 3, between-study heterogeneity and publication bias are described and discussed in further detail.

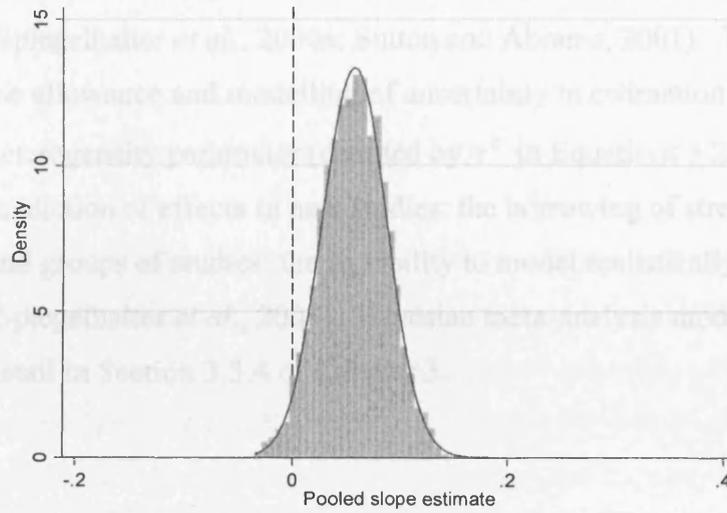
1.5 Bayesian methods for meta-analysis

Bayesian methods of statistical analysis are now increasingly popular, due mainly to advances in computing (Smith and Roberts, 1993). A Bayesian approach to statistical analysis allows subjective beliefs and/or external evidence to be formally incorporated alongside the dataset of interest. This prior information (e.g. from previous reports or expert opinion) must be expressed in terms of probability distributions before it can be included in an analysis. The basis for the Bayesian framework is *Bayes' Theorem*

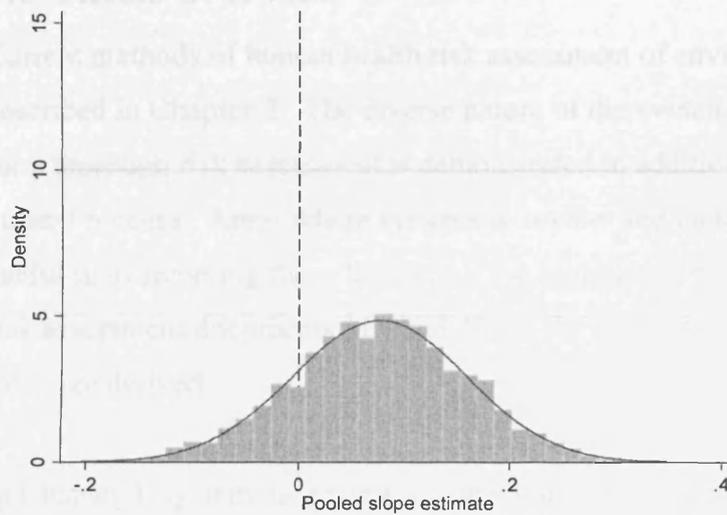
$$P(\theta|Data) \propto P(\theta)P(Data|\theta) \quad (1.1)$$

where $P(\theta)$ is the prior density function (findings from previous studies or expert opinion), $P(Data|\theta)$ is the likelihood function (the dataset of interest) and $P(\theta|Data)$ is the posterior density function which is often summarised by a mean or median (Sutton *et al.*, 2000). The importance given to the prior distribution compared to the likelihood function depends upon the precision of the information forming the prior distribution. The more information available, the more precise the prior distribution, thus the more impact it has in the analysis. In Figure 1.1 an example of a Bayesian analysis is represented showing the prior distribution, the likelihood function and the posterior distribution for the pooled slope estimate from a Bayesian meta-analysis. In this example the prior distribution is more precise than the likelihood function. This is reflected in the posterior distribution since it is more similar to the prior distribution than the likelihood function.

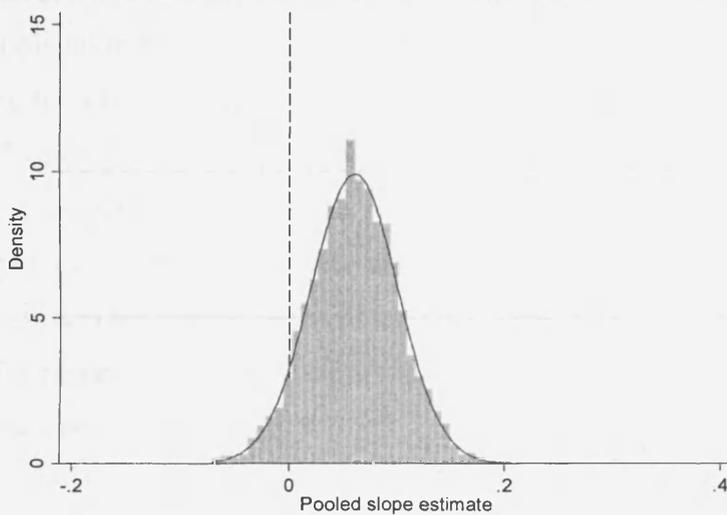
Figure 1.1 Graphical presentation of a Bayesian analysis



Prior distribution



Likelihood distribution



Posterior distribution

Bayesian methods are increasingly used to meta-analyse evidence in human healthcare and have a number of advantages over classical meta-analysis models (Spiegelhalter *et al.*, 2000a; Sutton and Abrams, 2001). These advantages include the allowance and modelling of uncertainty in estimation of the between-study heterogeneity parameter (denoted by τ^2 in Equations 3.2 and 3.4 in Chapter 3); the prediction of effects in new studies; the borrowing of strength from similar studies, and groups of studies; the flexibility to model realistically complex relationships (Spiegelhalter *et al.*, 2004). Bayesian meta-analysis models are discussed in more detail in Section 3.3.4 of Chapter 3.

1.6 Thesis overview

Current methods of human health risk assessment of environmental exposures are described in Chapter 2. The diverse nature of the evidence that must be considered for a thorough risk assessment is demonstrated in addition to the limitations of the current process. Areas where systematic review and meta-analysis methods may be useful in overcoming these limitations are highlighted from a comparative review of risk assessment documents in which limits for occupational exposure to manganese (Mn) are derived.

In Chapter 3, systematic review and meta-analysis methods are described and discussed, including details on assessing between-study heterogeneity and publication bias and the use of Bayesian methods for meta-analysis. Quantitative methods for the synthesis of human and animal evidence are also reviewed in Chapter 3. Since there has been little use and development of methods to combine human and animal evidence, potentially relevant methods for the generalised synthesis of human evidence are reviewed. In an attempt to identify an appropriate approach for exploring the potential of systematic reviews and meta-analyses in the risk assessment context, common features of the models reviewed in Chapter 3 are discussed.

The extent of use and quality of systematic reviews and meta-analyses of animal experiments is previously unreported and in Chapter 4 the details and findings of a

systematic review of the current use of systematic review and meta-analysis methods for the synthesis of animal evidence to inform human health are reported. As a consequence of the findings of this review and an absence of detailed guidance for systematic reviews and meta-analyses of animal studies, guidelines are proposed to aid the conduct and reporting of such articles.

In Chapters 5 and 6 the potential of systematic review and meta-analysis methods to assist in human health risk assessments of environmental chemicals is explored. The two examples used to illustrate the potential of systematic review and meta-analysis methods have been chosen because of their diversity, thus allowing assessment of issues that may be specific to each example and more general issues that may exist in the application of these methods to the human health risk assessment process. The examples are i) the risk of delivering a low birth weight baby associated with exposure to trihalomethanes (THMs) in drinking water (Chapter 5), and ii) possible neurobehavioural effects resulting from occupational exposure to manganese (Mn) (Chapter 6). For each example, the relevant human and animal data are identified and reviewed using systematic review methods and are synthesised using hierarchical Bayesian models that allow complex modelling of the diverse evidence and incorporation of additional evidence using informative prior distributions. As the method and result sections of Chapter 5 demonstrate, systematic reviews and meta-analyses can be used effectively in a risk assessment. However, because of diverse and sparse evidence such an application is not so straightforward for the second example, Mn exposure and neurobehavioural effects (Chapter 6). In the THMs example, the animal data appear consistent with the human data and so methods to synthesise evidence within *and* across different species and strains are investigated. Such consistency between species is, however, lacking in the Mn example, thus although synthesis across species may not be appropriate in this example, the advantages of systematic review and meta-analysis methods to summarise different types of evidence are demonstrated.

Consideration or assessment of the possibility of publication bias in a meta-analysis was an aspect found to be particularly deficient in the reporting of systematic reviews and meta-analyses of animal experiments in Chapter 4. Publication bias is

no less important in meta-analyses of animal experiments than it is in meta-analyses of human studies. Evidence suggests that current methods for assessing and adjusting for publication bias do not perform well (Begg and Mazumdar, 1994; Sterne *et al.*, 2000; Macaskill *et al.*, 2001; Schwarzer *et al.*, 2002; Terrin *et al.*, 2003). Using simulations prompted by the characteristics of the 46 meta-analyses of animal experiments identified in Chapter 4, the performance of these commonly used models and methods for assessing publication bias are assessed in Chapter 7. The rank correlation test (Begg and Mazumdar, 1994), Egger's regression tests (Egger *et al.*, 1997), Macaskill's regression test (Macaskill *et al.*, 2001) and the trim and fill method (Duval and Tweedie, 2000a; Duval and Tweedie, 2000b), are all investigated in the presence and absence of induced publication bias and between-study heterogeneity in addition to a number of alternative regression tests. From these simulation analyses an improved regression test for publication bias, based on sample size, is identified and its performance in likely meta-analyses for human health risk assessments is discussed.

In Chapter 8 investigation of the performance of tests for publication bias is extended to include the scenario where a measured study-level covariate may help explain observed between-study heterogeneity. Following on from a motivating example (Mapstone *et al.*, 2003) and the observation in Chapter 4 that many meta-analyses of animal experiments include different species and strains of animal, the performance of tests to help detect publication bias is assessed when some of the between-study heterogeneity can be explained by a measured covariate. This is likely to be a common feature of meta-analyses of animal experiments where results from experiments using different species and strains are synthesised.

Finally, in Chapter 9, the research contained within this thesis is summarised and the potential of systematic review and meta-analysis methods to assist in human health risk assessments is discussed, with particular attention paid to i) the application of systematic reviews and meta-analyses to evaluate animal experiments, ii) the quantitative synthesis of human *and* animal evidence, and iii) the performance of tests for the detection of publication bias in likely scenarios in human health risk assessments. Limitations of the work presented here are considered, in addition to recommendations for practice and possible areas of future

research. The thesis ends with conclusions on the potential of systematic review and meta-analysis methods to assist in human health risk assessments to environmental exposures.

Risk assessment methods

2.1 Chapter overview

In this chapter current approaches to human health risk assessment for environmental exposures are described and discussed. In Section 2.2 methods for the daunting and important task of identifying, reviewing and evaluating evidence on potential health risks from exposure to environmental chemicals and determination of exposure limits is introduced. Current risk assessment practice generally involves categorising substances into threshold and non-threshold substances. *Threshold substances* are assumed to have some level of exposure below which there are no adverse health effects. Above this threshold, exposure is thought to lead to adverse health effects. *Non-threshold substances* are assumed to be harmful, no matter how small the exposure. The majority of substances having a toxic effect are assumed to be threshold substances; non-threshold substances usually refer to genotoxic carcinogens. The approach taken to the risk assessment typically depends on whether a chemical is believed to be a threshold or non-threshold substance. In Section 2.3 the risk assessment approach for threshold substances is described, including determination of the critical effect for the risk assessment (2.3.1), the uncertainty factors (2.3.2) and the reference dose (2.3.3). General approaches for the risk assessment of non-threshold substances are described in Section 2.4. To highlight the limitations of the current risk assessment process, documents describing a human health risk assessment for occupational exposure to Mn are reviewed in Section 2.5. The advantages of systematic review and meta-analysis methods and their potential to overcome some of the limitations of current risk assessment approaches are discussed in Section 2.6.

2.2 Introduction

A number of chemicals in our environment, natural and man-made may be harmful to human health at certain levels of exposure. The aim of a human health risk assessment is to review and evaluate all the relevant evidence on the potential risks to human health from exposure to these chemicals. These risk assessments are usually carried out by governments or regulatory agencies, such as the World Health Organization (WHO) and the US Environmental Protection Agency (US EPA). The risk assessment process generally involves four steps (RATSC, 1999a; Woolley, 2003; Cogliano, 2005):

1. Hazard identification – identifying what adverse effects result from exposure to the substance (whether the effects are in animals or humans)
2. Hazard characterization – quantitative evaluation of the adverse effect
3. Exposure assessment – measured, estimated or predicted exposure to the substance by humans
4. Risk characterization – combines the three steps above to predict the severity of human effects from exposure to the substance, if there are any, and identifies the population likely to be affected (e.g. elderly, those occupationally exposed)

An informative and useful assessment of the risks to human health from exposure to chemicals should therefore be based on all available data relevant to the area of interest. Often these data are from a variety of sources, including animal experiments and human epidemiology studies (WHO, 1994; Woolley, 2003).

Within these sources it is likely that different methodologies will have been used (e.g. observational cohort and case-control design human epidemiology studies, in vitro and in vivo animal experiments). Hence the data potentially relevant for a human health risk assessment may be quite diverse. Although data from human studies are preferred over that from animal experiments in terms of relevance, the human data are likely to have low power and precision and be particularly prone to biases and confounding since often only data from observational studies are available. On the other hand, data from animal experiments are less likely to suffer from such biases and confounding and often be much more precise (WHO, 1994). However, results of an animal experiment are less relevant to an assessment of risks

to human health than results from a human study. Thus, a balance must be achieved between the merits of the data, such as relevance and precision, from these various sources when considering evidence for a risk assessment.

In many situations, especially in Europe, an *expert committee* will review the available published evidence and derive an exposure limit to an environmental substance below which there is no adverse risk to human health (Rubery *et al.*, 1990; WHO, 1994; RATSC, 1999b). However, for non-threshold substances an exposure limit below which there is no risk to human health cannot be achieved and so a level of 'acceptable' risk is defined. The exposure limit is then derived from this level of 'acceptable' risk (RATSC, 1999b). More quantitative methods are generally used in the US for risk assessment (RATSC, 1999b). These methods can take the form of quite complex mathematical and/or biological models, if and where the data are available, to model the process from exposure to a substance in a particular to species, including the concentration of the chemical in different organs through to any adverse effects likely to be observed (RATSC, 1999a; Watanabe, 2005). The approaches generally taken in the risk assessment of threshold and non-threshold chemical substances are described in Sections 2.3 and 2.4 respectively.

2.3 Threshold substances

2.3.1 *The critical effect*

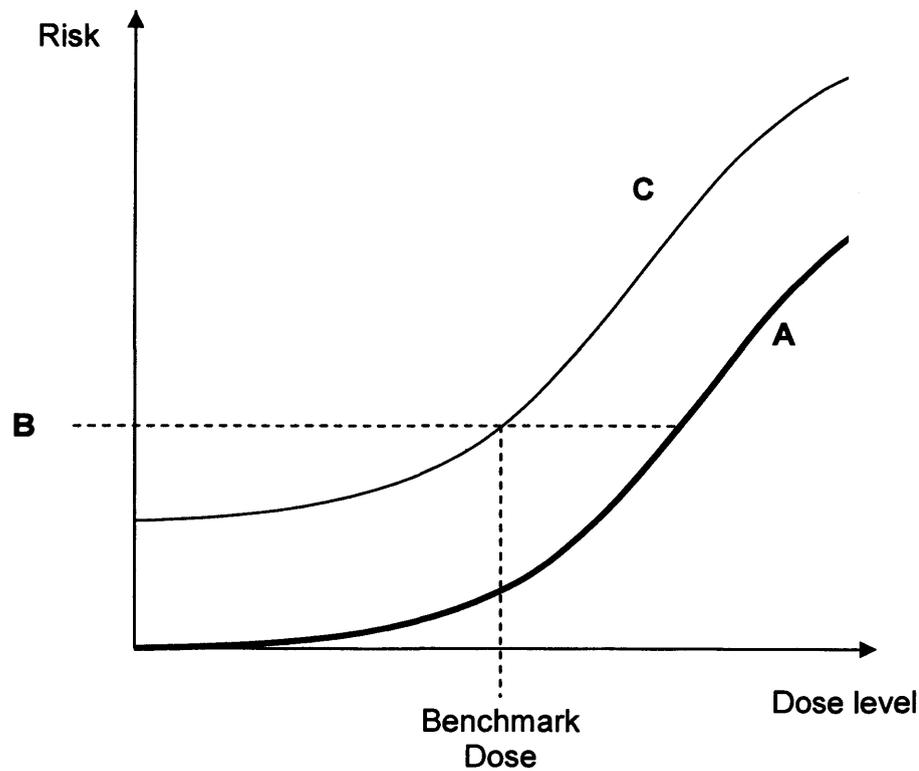
For a risk assessment the relevant evidence is identified and reviewed before a particular study is chosen as the *pivotal study* on which the calculation of the exposure limit is based. This pivotal study is usually chosen to be of high quality, and although human studies are preferred when investigating human health effects, typically only experiments on animals are suitable for this purpose. When a number of experiments are candidates for the pivotal study, the one using the most sensitive species or strain of animal, conducted over the longest study period is often chosen (Woolley, 2003). From this experiment the dose level at which there is *no observed adverse effect (NOAEL)* relating to the most serious health effect, or that which occurs at the lowest dose level is identified. The NOAEL is the highest dose (or exposure) level in the study at which no adverse health effects are seen in the study

subjects. When a NOAEL cannot be obtained from the study (i.e. when an adverse effect is observed at every dose level) the lowest dose at which an adverse health effect is observed in the experimental group, the LOAEL (*lowest observed adverse effect level*), is used (WHO, 1994).

There are a number of limitations of this approach. By its very nature, the N(L)OAEL is restricted by the design of the pivotal study it is taken from. It can only be a dose-level *used* in the pivotal study and so choice of doses and spacing between doses in the experiment becomes an important issue (Woolley, 2003). Identification of a N(L)OAEL also depends upon the statistical power of the experiment. If the experiment does not have sufficient statistical power, no adverse effects may be observed, possibly leading to a conclusion that exposure to a substance is not harmful when in fact the study was just not powerful enough to detect an effect (WHO, 1994). The approach is also subject to the issues of multiple-testing since effects at each dose level are tested separately (Edler *et al.*, 2005). Furthermore selection of a N(L)OAEL from a particular experiment is dependent upon the species, strain, sex and age of the animal, the duration of the exposure and observational period and the sensitivity of the method used to measure the outcome (WHO, 1994). A further criticism of the N(L)OAEL is that it makes no use of the data relating to other non-zero dose levels in this experiment, and so is not particularly efficient (Gaylor *et al.*, 1998).

A number of these criticisms of the N(L)OAEL can be overcome by use of the benchmark dose (BMD) method (Crump, 1984). Based on data from the pivotal study the BMD is calculated from a dose-response curve of this data (curve A in Figure 2.1). It is defined as the dose level corresponding to a specified level of increased risk (point B in Figure 2.1) calculated from the upper limit (often the 95% confidence limit) of the estimated dose-response slope (curve C in Figure 2.1).

Figure 2.1 Calculation of the benchmark dose



Thus, data from all dose levels in the pivotal study are used in the derivation of the BMD. This is in contrast to the N(L)OAEL approach where data from only one dose level are used. In the US, the BMD method is becoming increasingly popular and is replacing use of the N(L)OAEL approach. The popularity of this approach is reflected through its application to human epidemiological data (Budtz-Jørgensen *et al.*, 2000) and the fact that the US Environmental Protection Agency (EPA) is developing software to facilitate its application (EPA, 2003). However use of the BMD approach is not straightforward; a particular issue is the difficulty in finding an appropriate model for estimation of the dose-response curve used to obtain the BMD. This difficulty is demonstrated with the US EPA BMD software having provision for 16 different dose-response models (EPA, 2003), which inherently introduces an extra source of uncertainty into the risk assessment process: model uncertainty. Another criticism of the BMD approach is that in some cases, because of limited data within the pivotal study, it may not be possible to construct a dose-response curve (RATSC, 1999a). If, however, a dose-response curve can be

calculated on such little data, use of the BMD may be restricted by large or incalculable confidence limits (Murrell *et al.*, 1998). Such a situation would call for more data in terms of dose levels and animals within a study (RATSC, 1999b). Finally, as with the N(L)OAEL, the BMD is based on data from just one study. A more efficient approach would consider all available relevant evidence in setting an exposure limit.

2.3.2 Uncertainty factors (UFs)

If the pivotal study is a human study, there may be possible differences in the sensitivity between the study population and the population of interest for the specific risk assessment. For example, the pivotal study may be an occupational epidemiology study (and so fairly healthy individuals make up the study population), but the risk assessment is for exposures in the general population, including children, the elderly and those who may be ill. These differences in sensitivity to the exposure and/or outcome, known as *intraspecies variation*, must be considered. In current risk assessment practice, this is achieved by applying uncertainty factors (UFs) accounting for intraspecies variation to the N(L)OAEL or BMD (WHO, 1994; Woolley, 2003). If the pivotal study is on animal experiments, UFs are applied to account for differences in the sensitivity between responses from the laboratory animals and humans (*interspecies variation*) as well as differences between humans (*intraspecies variation*). Applying UFs to the N(L)OAEL or BMD simply involves dividing the N(L)OAEL or BMD by the product of the UFs. The resulting figure is the exposure limit and has many terms depending on the type of exposure being assessed; they include the tolerable daily intake, the acceptable daily intake, and the reference concentration. For the purpose of this thesis this figure is referred to as the *reference dose* (RfD), defined by

$$RfD = \frac{N(L)OAEL \text{ or } BMD}{\prod UF_s} \quad (2.1)$$

It is an estimate of the amount of contaminant/chemical, expressed on a body weight basis, which can be ingested over a lifetime without appreciable health risks to humans.

The most commonly used maximum values for the intraspecies and interspecies UFs are 10 and 10. Thus, when an animal experiment is the pivotal study, the total UF accounting for intraspecies and interspecies variation is 100 (10 for intraspecies variation x 10 for interspecies variation) (WHO, 1994). There may be further sources of uncertainty between the pivotal study and the exposure and outcomes of interest for the particular risk assessment. These include the severity of the adverse health effect, the use of a LOAEL rather than a NOAEL, different exposure routes (e.g. if the exposure route of interest for the risk assessment is different to that from which the N(L)OAEL is obtained), absence of data on chronic effects and whether sufficient data on similar compounds are available to help inform the risk assessment (WHO, 1994; Woolley, 2003). For each of these possible sources of uncertainty an UF can be applied. Decisions on the sources of uncertainty that exist and the magnitude of any UFs to be applied are made by expert judgement, usually in committees. In situations where great uncertainty exists, UFs of 10,000 have been used (WHO, 1994). Attempts to increase the accuracy of UFs by subdividing them into more specific areas of uncertainty have been proposed. Renwick (1993) has suggested that the interspecies UF of 10 should be divided into two components; one of up to 4 for toxicokinetic data (information on the association and process between the external and internal dose) and one of up to 2.5 for toxicodynamic data (information concerning the internal dose and its effect), so that,

$$\begin{aligned}\text{interspecies UF} &= \text{toxicokinetic uncertainty} \times \text{toxicodynamic uncertainty} \\ &= 4 \times 2.5 \\ &= 10\end{aligned}$$

When data are available for one or both of these factors the estimate of uncertainty is more precise and the interspecies UF is consequently more accurate. Similarly, splitting the UF for the intraspecies variation into two components to account for corresponding toxicokinetics and toxicodynamics data has been advocated (Renwick and Lazarus, 1998), but where each is allocated equal weighting (i.e. $3.2 \times 3.2 = 10$). This assumes that the maximum difference between sensitive subjects and the mean of the general population is a factor of 3.2 for both the toxicodynamics and toxicokinetics of the exposure. If toxicodynamic or toxicokinetic data do not exist, the default UF of 10 should be applied (Renwick, 1993).

The scientific background for the default UF of 100 for interspecies and intraspecies variation is not clear (Renwick and Lazarus, 1998). Renwick's proposal of dividing these two components further to reflect available information on toxicodynamics and toxicokinetics has been adopted by at least one regulatory agency: the WHO International Programme on Chemical Safety (IPCS) (Dourson *et al.*, 1996; Renwick and Lazarus, 1998). Although these UFs are still confined within the default of 100 for intraspecies and interspecies variation, there is evidence to suggest that in practice this default performs well to protect humans (Dourson and Stara, 1983; Dourson *et al.*, 1996). However, we do not know whether this will always be the case. The UK Department of the Environment reported that the interspecies and intraspecies sources of variation can take values anywhere between 10 and 1,000 depending on judgments made by the expert committee (DoE, 1993). Moreover when further sources of uncertainty exist (e.g. route of administration, use of a LOAEL) the magnitude of the total UF may be very large. In such cases the resulting RfD can be imprecise and too conservative. With this in mind the WHO has put forward a guideline maximum value for a total UF of 10,000 (WHO, 1994).

2.3.3 The reference dose (RfD)

There is a danger with the calculation and presentation of a single number (RfD) representing exposure limit for a particular chemical. Firstly, the RfD carries all the limitations of the parameters used in its calculation, the N(L)OAEL or BMD and the UFs. Therefore it can only be interpreted with an understanding of how these parameters are calculated and the assumptions involved. Often such data on the derivation of the N(L)OAEL, BMD and UFs are not easily available, although in recent years there has been a move towards increasing transparency particularly by UK agencies (RATSC, 1999b). Secondly, no measure of the uncertainty or variability of the RfD is given. The size of the UF gives some indication of the uncertainty, according to expert judgement, but more useful is a measure of the variance of this exposure limit. A step towards the evaluation and presentation of a measure of uncertainty for the RfD is in the use of probabilistic UFs. A number of authors have investigated using probabilistic UFs, where a distribution reflects likely values for an UF (Baird *et al.*, 1996; Slob and Pieters, 1998; Swartout *et al.*, 1998), thus accounting and describing the uncertainty in the UFs and leading to a probabilistic RfD.

2.4 Non-threshold substances

It is generally assumed, for risk assessment purposes, that any exposure, however small, to a genotoxic carcinogen or mutagen carries some level of risk, so that there is no threshold below which exposures can be regarded as safe; referred to as *non-threshold* substances. A number of approaches are taken in the risk assessment of these substances (WHO, 1994). For substances found in the environment and in food, methods have been used to obtain a level of exposure corresponding to a very low or 'acceptable' risk to human health (RATSC, 1999b), say a level which leads to one extra cancer per million people exposed for a lifetime (Woolley, 2003). These methods involve dose-response modelling of animal carcinogenicity data followed by the extrapolation of effects to likely exposure levels experienced by humans (Lovell and Thomas, 1997). A number of dose-response models are available for this low dose-level extrapolation which allow for a certain degree of conservatism (Moolenaar, 1994; Lovell and Thomas, 1997; Coglianò, 2005) equivalent to that from the use of UFs in risk assessments of threshold substances (RATSC, 1999b). The linearised, multistage model is the most commonly used extrapolation model (Mendes and Pluygers, 2005).

Physiologically-based pharmacokinetic (PBPK) modelling had been advocated to inform the extrapolation process (RATSC, 1999a; Woolley, 2003). These models allow more informed and precise extrapolation between species and/or routes of exposure where sufficient information on the biological and mechanistic effects of a chemical is available (Watanabe, 2005). However, although advocated, PBPK modelling is generally not used in the UK to inform risk assessment because of a lack of detailed information and expertise in applying the models (RATSC, 1999a).

However, as in the use of PBPK modelling, current understanding of the underlying biological processes is often still too incomplete to permit a confident choice of model, and different models may yield widely differing estimates. Subsequently, risk assessors must be cautious about interpreting precision as accuracy in the dose-response model (RATSC, 1999b). These more quantitative modelling approaches are generally favoured in the US. In particular the US EPA uses the linearised,

multistage models as the default for risk assessment of potentially carcinogenic substances (RATSC, 1999b).

Because of the limited information available for decisions on the choice of dose-response model, and the fact that a number of different models can be applied to the same data, use of dose-response models in the risk assessment of non-threshold substances is rarely carried out in the UK. Instead a qualitative, case-by-case weight of evidence approach is taken by an expert committee. This involves a narrative review of the evidence using subjective judgements to ascertain an exposure limit associated with an 'acceptable' risk to human health (Lovell and Thomas, 1997). Because of this subjectivity, there is uncertainty on the consistency of expert committees to weight various pieces of evidence (RATSC, 1999b).

However, the largest disadvantage of a narrative weight of evidence approach is that a quantitative estimate of risk from a specified level of exposure cannot be obtained. Instead the main finding of a risk assessment of a non-threshold substance from a weight of evidence approach is classification of that substance such as,

- Carcinogenic to humans
- Probably carcinogenic to humans
- Possibly carcinogenic to humans
- Not classifiable as to its carcinogenicity to humans
- Probably not carcinogenic to humans

(IARC, 2000).

2.5 Risk assessments for occupational exposure to manganese (Mn)

Most of the limitations observed in the previous two sections cannot be overcome until a time when more detailed evidence is available (e.g. to help inform PBPK modelling and the choice of UFs and dose-response models). However, there may be areas where the use of systematic review and meta-analysis methods could help to overcome some of these limitations. To illustrate where systematic reviews and meta-analyses may potentially improve current risk assessment processes, risk

assessment documents for occupational exposure to manganese (Mn) are reviewed, highlighting current practice and limitations.

2.5.1 Occupational exposure to Mn

Mn is an essential element; our bodies need Mn for us to survive and stay healthy. Estimates suggest a daily intake of 2-3 mg is required to maintain good health (WHO, 1981). Mn is an abundant element found in rock, water, air and soil, with food being the primary environmental source for humans (WHO, 1981). However, exposure to high levels of Mn can be harmful. For humans the main source of high level exposure is in occupational settings. Miners, in particular, are exposed to very high levels of Mn, while workers in the manufacturing of steel and iron products are also exposed, albeit to substantially lower levels than miners.

The type of health effects commonly observed in humans exposed to high levels of Mn are neurobehavioural effects: adverse effects relating to emotion, learning and behaviour. A combination of symptoms, known as 'manganism', has been reported by miners occupationally exposed to very high levels of Mn (ATSDR, 2000).

These symptoms are similar to those seen in people with Parkinson's disease, and include mental and emotional disturbances as well as difficulty with movement. In non-mining occupations exposure to Mn is more commonly characterised by difficulty with balance and hand and arm steadiness. Mn is considered to be a threshold substance because the human body requires it to maintain good health, yet higher levels of exposure are associated with adverse neurobehavioural health effects.

2.5.2 Reviewing the risk assessment documents

Since risk assessments are often commissioned by regulatory agencies and governments, a systematic review of published literature is likely to be of limited use in identifying documents describing the risk assessment and derivation of an exposure limit for Mn. Instead, citation searches on known Mn reviews and internet searches of regulatory agencies such as the American Conference of Governmental Industrial Hygienists (ACGIH) and the WHO were carried out to find Mn risk assessment documents for occupational exposure.

Five documents describing the derivation of a limit for occupational exposure to Mn were identified. Two are from the US, including the ACGIH and Clewell *et al.* (2003); one from the WHO, one in Germany from the Deutsche Forschungsgemeinschaft DFG, the German Research Foundation), and one published in the UK by the Institute for Environmental Health and Institute of Occupational Medicine (IEH and IOM). The occupational exposure limits derived in the documents range from 0.03 – 0.5 mg/m³ in air, depending on the type of exposure (total vs. respirable dust) (Table 2.1).

Table 2.1 The different types of occupational exposure limits derived for manganese (Mn)

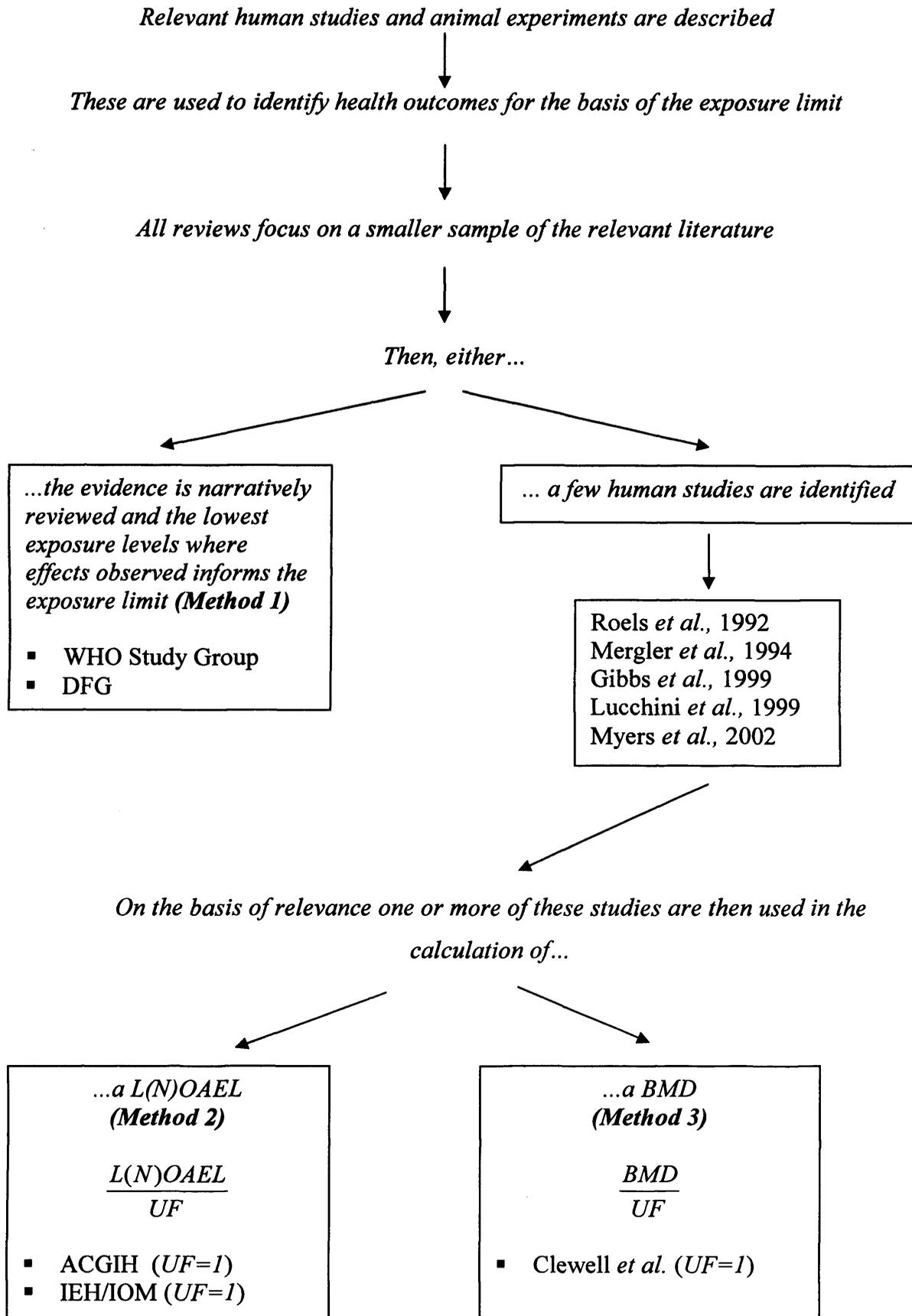
Organisation	Country	Year	Type of limit	Level (mg/m ³)
WHO Study Group	WHO	1980	Occupational Exposure Limit	0.3 mg/m ³ respirable dust (TWA)
DFG	Germany	1999	Maximum Workplace Concentration	0.5 mg/m ³ total dust
ACGIH	US	2002	Threshold Limit Value for Occupational Exposure	0.03 mg/m ³ respirable dust (TWA)
Clewell <i>et al.</i>	US	2003	Occupational Exposure Guideline	0.1-0.3 mg/m ³ respirable dust (8-hr TWA)
IEH and IOM	UK	2004	Occupational Exposure Limit	0.1 mg/m ³ respirable dust

TWA, Time-weighted average (e.g. 8 hours per day for 5 days a week to account for occupational exposure); WHO, World Health Organization; DFG, Deutsche Forschungsgemeinschaft (German Research Foundation); ACGIH, American Conference of Governmental Industrial Hygienists; IEH and IOM, Institute for Environment and Health and Institute for Occupational Medicine

- **Occupational Exposure Limit** (OEL), a limit of the concentration of a hazardous substance in workplace air;
- **Threshold Limit Value** (TLV), reflecting the level of exposure that the typical worker can experience without an unreasonable risk of disease or injury (scientific opinion based solely on health factors);
- **Maximum Workplace Concentration** (MAK), the German equivalent of the OEL

A summary of how the reported exposure limits were derived in each risk assessment document is given in Appendix A. Figure 2.2 details the general approach taken by each of the Mn risk assessments in their derivation of an exposure limit. As this figure illustrates, no UFs are applied to the calculated N(L)OAELs and BMDs (i.e. UF=1). This is because the evidence comes from human occupational epidemiology studies reflecting the population and situation of interest for the exposure limit being derived. However, the different risk assessments use different pivotal studies for the calculation of the N(L)OAEL or BMD which accounts for the different occupational exposure limits obtained by the risk assessments. Different choices of pivotal study can be explained in part by newer studies only being included in the more recent reviews, however this does not explain all of these differences.

In each of the risk assessments in Table 2.1 details on the process through which the evidence was searched and identified to be relevant, in addition to inclusion or exclusion criteria of any kind are not reported. This is a serious omission. Although it is unlikely that an important piece of evidence may have been overlooked (because of the number of experts involved in the risk assessment) the overall evidence base may not be wholly representative if a comprehensive literature search was not undertaken. Therefore if the basis of the evidence cannot be traced, setters and users of the limit cannot be entirely confident that the limit was derived from an unbiased set of evidence. Furthermore, in the reviews following Method 1 in Figure 2.2 (reporting a narrative review of the evidence) the actual data on which the limit is derived are difficult to determine. These limitations regarding transparency of the evidence base could be overcome by the use of systematic review methods where a search strategy is clearly defined, including any restrictions and details of the inclusion and exclusion criteria used. All of which will provide greater transparency than that currently found in human health risk assessments.

Figure 2.2 Flow chart of the approach taken to derive the exposure limit

An additional limitation in the risk assessment approaches described here is that although evidence is available from a number of studies, in Methods 2 and 3 in Figure 2.2 only one study is used for the basis of the exposure limit. Thus decisions on the choice of pivotal study have had to be made and often the reasons for choice of study are not clear. Furthermore there are no sensitivity analyses for this choice of study. For instance, within a particular risk assessment, the following question is not answered: what would the limit be if a different study was the pivotal study? (Although from this review of Mn documents such a question can be addressed.) Use of just one study for the basis of the calculation of the exposure limit when there are other relevant studies available is inefficient. Although these other studies may inform the health outcome used for the derivation of the exposure limit, how the limit is expressed or the UF values applied, the data are not formally incorporated. Moreover statistical power limitations within a study may pose a threat to the validity of the derived exposure limit. If, as mentioned in Section 2.3, an effect could not be seen at a particular exposure level, this may be due to a lack of statistical power, rather than a lack of effect. If, on the other hand, a large effect is seen this could just be due to chance.

2.6 Risk assessments, systematic reviews and meta-analyses

The evidence relevant for a risk assessment can be diverse and often both human and animal data must be considered. Debate on the use of animal experiments to inform human health is on-going, whether for human health risk assessments of environmental chemicals or to investigate the efficacy of medical interventions. Issues concern the relevance of animals for human diseases, the extrapolation of results from animal experiments to human experience, in addition to ethical aspects of using animals for experimentation (Pound, 2001; Smith, 2001; Pound and Ebrahim, 2002). The 3 Rs of animal research were introduced as a means of promoting humane use of animals in experiments and encouraging work towards alternatives to animal experimentation (Russel and Burch, 1959), they are:

- Refinement: minimise suffering and distress
- Reduction: minimise number of animals used
- Replacement: avoid the use of living animals

Many authors have pointed out that systematic review and meta-analysis methods have a role in this debate on animal experimentation (Roberts *et al.*, 2002a; Sandercock and Roberts, 2002; Pound *et al.*, 2004; Khan and Mignini, 2005; Macleod *et al.*, 2005). These methods have the potential to identify and bring together the vast amount of animal, and human, evidence for a risk assessment allowing a rigorous, transparent and explicit evaluation of the current literature. Rather than just deriving an exposure limit, use of these methods in human health risk assessment may lead to a decision that no more animal experiments are required or that more studies in humans are needed. Thus, because of the systematic review and/or meta-analysis it may be clear that further animal experiments are unnecessary. Although not relevant to the context of setting an environmental exposure limit, but relevant to the use of systematic reviews and meta-analyses to review animal experiments, a systematic review of the human and animal evidence on the effects of nimodipine in focal cerebral ischemia demonstrates this important point (Horn *et al.*, 2001). The authors noted that although conducted before and alongside 22 human trials of nimodipine, the animal data were not systematically reviewed until after evidence of an effect could not be found in the human data. The findings from the animal data suggested that there was, in fact, no convincing evidence to begin human clinical trials.

Use of meta-analysis methods to combine data from relevant studies for a human health risk assessment could help to overcome some of the limitations of the current approach described in this chapter. For instance, formal inclusion of evidence from a number of relevant studies rather than reliance on a pivotal study would not only make more efficient use of the evidence, but result in an increase in statistical power over a single study (Fleiss and Gross, 1991; Blettner *et al.*, 1999; Roberts *et al.*, 2002a; Sandercock and Roberts, 2002; Pound *et al.*, 2004). In the case of occupational exposure to Mn, human evidence was found to be sufficient for the basis of the risk assessment. However, sometimes only animal evidence is available and this increase in power from a meta-analysis may have implications for one of the 3Rs of animal research: reducing the number of animals used in experiments.

A further advantage of the use of a meta-analysis of studies rather than a pivotal study is the generalisability of the results from a meta-analysis. The occupational

exposure limit for Mn is intended to cover a number of occupations, and so a meta-analysis of studies in different occupational settings may be particularly beneficial here. Roberts *et al.* (2002a) have also discussed that meta-analysis methods allow consistency to be assessed across species, and this can be extended to assessments across different occupational settings, routes of exposure and so on. Furthermore an understanding of sources of bias that may be apparent in the individual studies, thus leading to improvements in the quality of conducting and reporting human studies and animal experiments, could be identified from a meta-analysis (Sandercock and Roberts, 2002; Macleod *et al.*, 2005). Finally, meta-analysis offers a quantitative framework which allow for the investigation into possible sources of between-study heterogeneity and may help in assessment of possible publication bias (Roberts *et al.*, 2002a; Sandercock and Roberts, 2002; Macleod *et al.*, 2005).

The application of systematic review and meta-analysis methods to data from human studies (RCTs and epidemiology studies) is now commonplace. However, the use of these methods to review and evaluate evidence from animal experiments is less documented. In the following chapter, general systematic review and meta-analysis methods are described, including the use of Bayesian methods of meta-analysis. The application of these methods to the synthesis of human and animal evidence are reviewed, in addition to more recent methods for the generalised synthesis of human evidence in healthcare research.

Meta-analysis and generalised synthesis of evidence methods

3.1 Chapter overview

In the previous chapter, the often quite diverse evidence relevant for human health risk assessments for environmental exposures was described. Current methods for such risk assessments and some of their limitations were presented and discussed, and the use of systematic review and meta-analysis methods in this context was proposed. In this chapter, general systematic review and meta-analysis methods used in human healthcare research are briefly summarised in Sections 3.2 and 3.3, with descriptions of fixed and random effects inverse-variance weighted meta-analysis models and Bayesian methods of meta-analysis. Meta-analysis models used for the assessment of human and animal data to inform human health risk assessments from exposure to environmental chemicals are then reviewed in more detail in Section 3.4. In the last ten years there has been little use and development of these methods to combine human and animal data. However, there has been some application of systematic review and meta-analysis methods to the synthesis of toxicological data only and to the synthesis of human evidence only from different, but related, sources. Methods used in these areas of research are presented in Section 3.5 and common features of all the models reviewed in this chapter are discussed in Section 3.6 with respect to the synthesis of evidence for human health risk assessments of environmental chemicals. Bayesian hierarchical models are identified as providing the most appealing framework for the synthesis of human and animal evidence to inform human health effects. The advantages of such models are discussed and their application to two risk assessment examples (Chapters 5 and 6) is proposed.

3.2 Systematic review methods

As mentioned in Section 1.4, systematic reviews are commonly used in human healthcare research to identify, evaluate and summarise all the available relevant evidence concerning a clearly focused pre-defined research question. Explicit, transparent, repeatable criteria are used to carry out these reviews and minimise all possible sources of bias (Sutton *et al.*, 2000; Egger *et al.*, 2001), including biases in the selection of studies. Research groups have been established to carry out and promote systematic reviews in human healthcare. These include the Cochrane Collaboration and the NHS Centre for Reviews and Dissemination (CRD) based at the University of York. The Cochrane Collaboration and the CRD make procedural documentation available on their websites for carrying out systematic reviews of evidence on medical interventions to inform decision-making. There are a number of important steps in any systematic review, and these are described and discussed in the documentation from the Cochrane Collaboration and CRD. They include:

- Formulating the research question
- Development of a protocol
- Identifying and selecting studies
- Assessing study quality
- Data extraction
- Data synthesis (if appropriate)
- Interpreting results and recommendations

(Deeks *et al.*, 2001; Higgins and Green, 2005).

Advantages of using systematic review methods to review and summarise evidence are well documented in human health research (Cook *et al.*, 1995; Blettner *et al.*, 1999; Moher *et al.*, 1999; Stroup *et al.*, 2000; Bracken *et al.*, 2001; Egger *et al.*, 2001; Chalmers *et al.*, 2002), and beyond. For instance, established in 2000 from the Cochrane Collaboration, the Campbell Collaboration also aims to prepare and maintain systematic reviews of interventions, but in social, behavioural and educational areas, rather than medical (www.campbellcollaboration.org). As with the Cochrane Collaboration and CRD, guidance on carrying out systematic reviews is provided by the Campbell Collaboration.

Where appropriate, a systematic review can be extended to a meta-analysis of the relevant articles: the quantitative synthesis of results from individual studies. The various methods used for meta-analyses and important issues arising from a meta-analysis are described in the following section.

3.3 Meta-analysis methods

Meta-analyses allow quantitative estimation of effects across multiple studies. There are many techniques available for meta-analysis, including the synthesis of p-values from each study, resulting in a pooled p-value for the statistical significance of the effect in question (Sutton *et al.*, 2000; Whitehead, 2002). Although simple, this approach is not particularly informative as the magnitude of an effect is preferable to just an indication of its statistical significance. The most commonly reported pooled estimate is calculated using a weighted average of the results of primary studies, where the weight is dependent upon the precision of the estimate from each individual study - so that studies with large precision are given more weight in the meta-analysis (Sutton *et al.*, 2000). These methods are generally known as *inverse-variance weighted* meta-analysis models. Fixed and random effects models are used in meta-analysis and choice of one over the other depends on the variability of the study specific estimates assumed within the meta-analysis. Excess variability, known as *between-study heterogeneity*, is an important feature of any meta-analysis that must be considered and explored. Another problematical feature of meta-analyses (and systematic reviews) is the possibility of *publication bias*: where an article is more (or less) likely to be submitted and published than another article depending on its findings. Publication bias and between-study heterogeneity are discussed in more detail in Sections 3.3.2 and 3.3.3. First, fixed and random effects meta-analysis models are described.

3.3.1 Fixed and random effects meta-analysis models

The fixed effects model assumes that the summary estimates from the individual studies in a meta-analysis are all estimating the same underlying effect. Any differences between study estimates are assumed to be due to sampling error only (i.e. there is no between-study heterogeneity).

The fixed effects model is given by

$$y_i = \mu + \varepsilon_i \quad \text{var}(y_i) = \sigma_i^2 \quad (3.1)$$

where y_i is the observed effect in study i , μ is the underlying effect, ε_i is the error in the y_i estimate of μ and σ_i^2 is the variance (assumed known) in study i (Sutton *et al.*, 2000). In many meta-analyses however it is unlikely that the only variability between estimates is due to sampling error, often between-study heterogeneity is observed. Between-study heterogeneity may reflect differences between the studies in the meta-analysis, e.g. the way in which exposure was measured or the fact that some studies were carried out in different geographical locations. The presence of between-study heterogeneity means that the assumptions of the fixed effects meta-analysis model are not met. If some proportion of the between-study heterogeneity can be explained by a measured covariate then it is best to use meta-regression or subgroup analyses (see Section 3.3.2). If however, between-study heterogeneity cannot be explained then a random effects model, rather than a fixed effects model, should be used since it assumes that there is greater variability between study estimates than that due to sampling error alone (unlike the fixed effects model). Between-study heterogeneity, denoted by τ^2 , is taken into account and estimated in the random effects meta-analysis model. The model, referred to as a 2-level hierarchical random effects model, is given by

$$y_i = \theta_i + \varepsilon_i \quad \theta_i \sim N(\mu, \tau^2) \quad \text{var}(y_i) = \sigma_i^2 + \tau^2 \quad (3.2)$$

where y_i is the observed effect in study i , θ_i is the true effect in study i and μ is the overall true underlying effect. ε_i is the error in the y_i estimate of μ , σ_i^2 is the variance (assumed known) in study i and τ^2 is between-study variance to be estimated (Sutton *et al.*, 2000).

The decision as to whether a fixed or random effects meta-analysis model should be used depends on the existence of between-study heterogeneity in the meta-analysis, and is discussed further in the next section.

3.3.2 Between-study heterogeneity

Techniques are available for the assessment of between-study heterogeneity in a meta-analysis, including the Q-statistic which tests the hypothesis that the underlying effects are the same in each study (Sutton *et al.*, 2000) and I^2 . I^2 assesses the percentage of total variation across studies that is due to between-study heterogeneity rather than chance (Higgins and Thompson, 2002). It is given as

$$I^2 = \frac{H^2 - 1}{H^2} \quad (3.3)$$

where $H^2 = \frac{Q}{k-1}$, Q is the Q-statistic mentioned above and k is the number of studies in the meta-analysis.

However, when between-study heterogeneity exists it can affect the inference and conclusions of a meta-analysis, so merely accounting for it in a random effects meta-analysis model is not sufficient. Investigation of possible sources of between-study heterogeneity should be carried out and attempts made to explain the extra variability. Subgroup analyses and meta-regression have been advocated to do this (Cook *et al.*, 1995; Moher *et al.*, 1999; Stroup *et al.*, 2000; Higgins and Green, 2005). Meta-regression involves the regression of effect size on study-level covariates believed to explain some of the between-study heterogeneity (e.g. year of publication, country study carried out in). A random effects meta-regression is preferred since it models heterogeneity that is not explained by the study-level covariates (Sutton *et al.*, 2000; Thompson and Higgins, 2002). The model is given by

$$y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + u_i + \varepsilon_i \quad (3.4)$$

where β_m are the regression coefficients to be estimated ($m=0, \dots, k$), X_n are the values for the k study-level covariates ($n=1, \dots, k$), $u_i \sim N(0, \tau^2)$ is the random effect (accounting for extra heterogeneity not explained by the study-level covariates) and ε_i is error (Sutton *et al.*, 2000; Whitehead, 2002).

3.3.3 Publication bias

Publication bias describes the tendency of some studies to be more (or less) likely to be published, hence more (or less) likely to be identified and included in a systematic review or meta-analysis, because of the size and/or statistical significance of the summary estimate from that study (Song *et al.*, 2000). If publication bias occurs, the subsequent systematic review or meta-analysis of published literature may be misleading. A number of techniques to help identify whether a meta-analysis is subject to publication bias have been developed and are commonly used. They include the funnel plot (Light and Pillemar, 1984), a rank correlation test (Begg and Mazumdar, 1994), a regression test (Egger *et al.*, 1997), and a nonparametric method known as trim and fill which adjusts for publication bias (Duval and Tweedie, 2000a; Duval and Tweedie, 2000b). Details of the funnel plot, the rank correlation and regression tests, and the trim and fill method are given in Chapter 7 where their performance is assessed under a number of scenarios using simulation analyses.

3.3.4 Bayesian meta-analysis models

Bayesian methods are often used in meta-analysis (Spiegelhalter *et al.*, 2000a; Sutton and Abrams, 2001) and a number of examples are available (Pasquali *et al.*, 2002; Oberwald *et al.*, 2003; Babapulle *et al.*, 2004). Bayesian methods of meta-analysis are sufficiently flexible to be used to combine evidence from a number of different study designs, as will be seen in Chapter 5. They are also attractive in that they allow borrowing of strength across similar types of studies. Furthermore, they allow, where appropriate, for prior evidence and/or expert judgement to be incorporated into an analysis of evidence from multiple sources (Spiegelhalter *et al.*, 2000a). Because of their flexible nature which allows realistically complex models to be fitted, and their ability to include external evidence in the synthesis, Bayesian methods of meta-analysis are appealing for the synthesis of diverse evidence needed for a human health risk assessment. This will be seen in Sections 3.4 in the review of methods used and will be demonstrated in Chapters 5 and 6 where such models, incorporating informative prior distributions, are applied to two different environmental exposure risk assessment examples. Further advantages of Bayesian meta-analyses include the ability to predict effects in future studies and allowance

of variability in the between-study heterogeneity parameter τ^2 , as mentioned in Section 1.5.

The pooled estimate from a Bayesian meta-analysis does not generally differ to that from a classical meta-analysis. However, since extra variability is accounted for in a Bayesian meta-analysis, e.g. in the estimation of τ^2 in random effects meta-analyses, there is usually greater variability in the pooled estimate from a Bayesian meta-analysis than that from a classical meta-analysis.

A Bayesian random effects meta-analysis model can be specified by

$$\begin{aligned} y_i &\sim N(\theta_i, \sigma_i^2) & \theta_i &\sim N(\mu, \tau^2) & (3.5) \\ \mu &\sim [-, -] & \tau^2 &\sim [-, -] \end{aligned}$$

where y_i is the observed effect in study i , with σ_i^2 denoting the variance in study i . θ_i is the true effect in study i , μ is the overall true underlying effect and τ^2 is the estimate of between-study variance. Prior distributions need to be placed on the overall effect, μ , and the estimate of between-study variance, τ^2 . Two types of prior distributions could be placed on these parameters: vague or informative. A vague prior distribution is intended to provide no additional information on the parameters of interest, allowing the data to dominate in the analysis. Informative prior distributions, on the other hand, can be used to incorporate additional relevant evidence. This additional evidence may come from a number of different sources, including prior beliefs elicited from experts and data from related experiments or studies. It is important that the sensitivity of results from a Bayesian analysis to different prior distributions is assessed. This is especially important when prior distributions are specified that are intended to be vague, since in reality they may be quite informative (Lambert *et al.*, 2005). As shown in Equation 3.5 above, in Bayesian meta-analyses prior distributions are placed on variance parameters (e.g. τ^2); often little additional information is available about these parameters and so vague prior distributions are used. Lambert *et al.* (2005) demonstrated the

importance of sensitivity analyses for vague prior distributions, particularly for variance parameters.

Even when vague prior distributions are placed on parameters in a model, a Bayesian analysis still has a number of advantages over a classical analysis of the evidence (Sutton and Abrams, 2001). These advantages include the ability to make direct probability statements and the calculation of predictive distributions, in addition to the fact that all parameter uncertainty can be accounted for and propagated throughout the modelling process. Hence, a Bayesian approach is taken for the synthesis of evidence in this thesis (see Chapters 5 and 6). The Bayesian analyses reported in this thesis have all been carried out in WinBUGS (Spiegelhalter *et al.*, 2000b). This software package was designed specifically for Bayesian analyses, and is discussed further in Section 5.5 (Chapter 5).

3.4 Generalised synthesis of evidence on environmental exposures

The use of meta-analysis methods to synthesise relevant evidence regarding environmental exposure – response relationships has been noted by a number of authors (DuMouchel and Harris, 1983; DuMouchel and Groër, 1989; Cox and Piegorsch, 1994; Hasselblad, 1995; Tweedie and Mengersen, 1995; Dominici *et al.*, 2000; Simmonds *et al.*, 2003; Dominici *et al.*, 2004; Wolpert and Mengersen, 2004). The majority of articles apply meta-analysis methods to findings from human environmental epidemiology studies, e.g. environmental tobacco smoke and lung cancer (Hasselblad, 1995; Tweedie and Mengersen, 1995; Wolpert and Mengersen, 2004), air pollution and mortality (Dominici *et al.*, 2000 and 2004), health effects from low-level exposure to lead (Hasselblad, 1995) and respiratory effects from exposure to nitrogen oxide (Hasselblad, 1995). However, as seen in Chapter 2, the evidence relevant to the assessment of risks from environmental exposures can be quite diverse, often involving the examination of results from human studies and animal experiments. Few authors (DuMouchel and Harris, 1983; DuMouchel and Groër, 1989; Cox and Piegorsch, 1994) have investigated methods

to combine human and animal evidence relevant to inform human health effects from environmental exposures. The models employed in these articles are now described.

DuMouchel and Harris (1983) applied a *Bayesian hierarchical regression model* with *random effects* to the synthesis of human and animal evidence for an assessment of human lung cancer risk from exposure to a number of environmental emissions, including diesel engine exhaust, benzo(a)pyrene and cigarette smoke. Data from six studies are used (two human epidemiology study, one *in vivo* experiment in mice and three *in vitro* animal experiments – one with Syrian hamster embryo cells and two with L5178Y mouse lymphoma cells). These six studies provide evidence on ten different exposures, where estimates for the effects of a particular exposure are available from at least two different species. The data are synthesised with the aim of informing human health effects from exposure to these ten environmental emissions (the one human study only investigates three of the ten exposures).

A linear regression model is used, where the relative potency of exposures is assumed equal across species, i.e. exposure A is more potent than exposure B which is more potent than exposure C for all species. The model is given by

$$y_{kl} \sim N(\theta_{kl}, c_{kl}^2) \quad \theta_{kl} = \mu + \alpha_k + \gamma_l + \delta_{kl} \quad (3.6)$$

where y_{kl} is the estimated logarithm of the dose-response slope in species k ($k: 1, \dots, 4$) for exposure l ($l: 1, \dots, 9$). μ is the overall mean effect with α_k and γ_l representing the average species and exposure effects, respectively. θ_{kl} is the true log dose-response slope in species k for exposure l , with standard error c_{kl} , which is assumed to be known. $E[\theta_{kl} | \mu, \alpha_k, \gamma_l] = \mu + \alpha_k + \gamma_l$ and δ_{kl} is then defined as $\theta_{kl} - \mu - \alpha_k - \gamma_l$, $\delta_{kl} \sim N(0, \sigma^2)$ (DuMouchel and Harris, 1983).

Informative prior distributions are placed on σ , which reflects the fit of the equal relative potency model and diffuse prior distributions are placed on μ , α_k and γ_l . These prior distributions must be assessed for their sensitivity. DuMouchel and

Harris (1983) also discuss a method for selecting the most relevant evidence from that available to inform investigation of the risk of human lung cancer from exposure to various emissions. This involves finding the set of species and exposure data which minimises σ . The authors conclude that the most efficient use of evidence in this Bayesian hierarchical approach requires reasonably informative data on a number of different exposures and species, rather than very precise estimates for just one exposure or species.

DuMouchel and Groër (1989) apply this model (Equation 3.6) to 13 datasets to inform investigation into the risk of bone cancer in humans from exposure to particular nuclides and isotopes. Data from four biological systems (humans, beagles (via injected or inhaled exposure) and rats) investigating the effects of exposure to two radionuclides and two isotopes, summarised by dose-response slopes, are used. When evidence from multiple studies investigate the same exposure-species interaction, the data are synthesised by the calculation of an inverse-variance weighted mean dose-response slope. Thus, the evidence to be used in the model (Equation 3.6) comes from a 4 x 4 table detailing all possible species-exposure interactions. As in DuMouchel and Harris (1983), specification of prior distributions for the unknown parameters α_k , γ_l and σ is discussed and analysed for sensitivity. DuMouchel and Groër (1989) also conclude that best use of the evidence from such a model requires there to be information across all exposures and species, rather than very precise estimates for one species or exposure.

More recently, Cox and Piegorsch (1994) have described methods for the synthesis of evidence to inform human health effects from environmental exposures. In their report of a meeting organised by the US National Institute of Statistical Sciences and the US EPA methods for the synthesis of p-values, fixed and random effects inverse-variance weighted models and weighted linear regression models are described to combine evidence from human epidemiological studies. The use of Bayesian methods are also discussed, in particular to help model uncertainties in the synthesis of the evidence, especially the between-study heterogeneity parameter. The possible synthesis of relevant human evidence with animal evidence is presented in the form of an *ordinal regression model with random effects*.

Incorporating terms to account for different species and exposure concentrations and durations, the (ranked) severity of health outcomes from the different human and animal studies are modelled. The ordinal regression model is given by

$$P(y_{ij} \geq s | b_i, x_{ij}, z_{ij}, u_{ij}) = \frac{1}{(1 + \exp(-(\alpha + \beta'x_{ij} + \gamma'z_{ij} + b_i'u_{ij})))} \quad (3.7)$$

where y_{ij} is the severity ranking (s describes the severity categories) for exposure group j ($j: 1, \dots, n_i$) in study i ($i: 1, \dots, M$). x_{ij} denotes the dose level/concentration for exposure group j in study i and z_{ij} represents exposure group and study specific covariates. The random effects are modelled by u_{ij} and study-specific parameters b_i . The unknown parameters α , β and γ are estimated by the model. This regression model is not applied to a dataset in Cox and Piegorsch (1994), however articles by Carroll *et al.* (1994) and Guth *et al.* (1997) report the use of this model when applied to evidence on acute inhalation to tetrachoroethylene. These two articles are included in the systematic review of systematic reviews and meta-analyses of animal experiments in Chapter 4.

Both of the approaches described above to synthesise human and animal evidence (Equations 3.6 and 3.7) use random effects regression models which account for species, exposure and outcome differences; one approach uses a Bayesian framework to inform associations between related exposures and species. In the last ten years there has been very little use and development of such methods for the synthesis of human and animal evidence. However, similar models have been used to combine evidence from multiple toxicological experiments and to combine different types of human evidence. The models and their applications are discussed in the next section.

3.5 Other relevant research

3.5.1 Synthesis methods for toxicological data

Following on from the use of *Bayesian hierarchical random effects regression models* to combine human and animal evidence by DuMouchel and Harris (1983)

and DuMouchel and Groër (1989), Wolpert and Warren-Hicks (1992) and Dominici *et al.* (1997) apply similar models to investigate relationships between measurements of substances in lakes. In order to simultaneously analyse field and laboratory data to predict the presence or absence of brook trout in lakes from measurements of water chemistry, Wolpert and Warren-Hicks (1992) use a *hierarchical Bayesian* model. Logistic regression is used to model the presence of trout in lakes on measures of pH, aluminium and calcium ultimately leading to predictions on their presence in lakes where such water chemistry measures are available. Not only does the use of Bayesian models allow evidence from the two different, but relevant, sources of evidence (field and laboratory data) to be synthesised, but an issue of multicollinearity in the example Wolpert and Warren-Hicks use can also be addressed using such a model. The authors conclude that the data dominated in the analysis of the evidence since results from using two different sets of prior distributions were found to be similar.

Dominici *et al.* (1997) apply a similar model to investigate the relationships between measurements of three substances to inform the water quality of 12 lakes in the US. For ten of these 12 lakes, measurements of chlorophyll-*a* (C), total phosphorus (TP) and total nitrogen (TN) have been taken on at least two different occasions. In two lakes, only measurements of C and TP are available. The aim of the analysis is to model C using TP and TN in order to predict water quality from future measurements of TP and TN. The random effects regression model needs to account for multiple measures of a substance at each lake and the fact that measurements of TN are missing for two lakes. Levels of C are modelled in each of the $s = 1, \dots, 12$ lakes by

$$\log(C) = \beta_0^s + \beta_1^s \log(TP) + \beta_2^s \log\left(\frac{TN}{TP}\right) \quad (3.8)$$

Diffuse plausible prior distributions are placed on the unknown parameters for estimation of the relationships between the lake measurements. Dominici *et al.* (1997) conclude that the flexibility of the Bayesian random effects model allows for a comprehensive analysis of the lake water quality problem, to inform future water quality management matters.

More recently, Bayesian hierarchical models have been used to combine estimates of potency from mutagenesis arrays of toxic environmental agents (Simmonds *et al.*, 2003). Findings from multiple Ames/*Salmonella* microsome mutation assays, which use different strains of the bacterium *Salmonella typhimurium* to screen for genetic damage to DNA after exposure to environmental stimuli, are combined to inform possible genotoxic effects from environmental exposures. From each assay the number of mutant colonies observed after a certain time period is recorded as a function of the dose for each strain used in the experiment. The authors state that a better understanding of the potency of the toxic agent is achieved from an evaluation of findings across dose levels and strains. Simmonds *et al.* (2003) use the *Bayesian two-level random effects meta-analysis model* given in Equation 3.5 to synthesise evidence from the Ames assays. A diffuse prior distribution is placed on μ , the true overall potency of the toxic agent, and the following shrinkage prior distribution, $\pi(\tau)$, is used for τ , the estimate of between-study standard deviation, from DuMouchel (1994)

$$\pi(\tau) = \frac{t_0}{(t_0 + \tau)^2} \quad t_0 = \frac{K}{\sum_{k=1}^K \frac{1}{V_k}} \quad (3.9)$$

Although the previous three meta-analysis models (including Equations 3.8 and 3.9) demonstrate the use of Bayesian methods to evaluate toxicological data, only one type of evidence has been combined. As discussed in Chapter 2, evidence from different sources needs to be combined to inform a human health risk assessment of environmental exposures. Over the last 20 years, there has been growing acknowledgement that evidence needed to inform practice and policy in human healthcare is likely to come from a variety of different sources (Ades and Sutton, 2006). With this in mind, an increase in the use and development of frameworks and methods for the generalised synthesis of human evidence has been seen (i.e. the combination of different, but related, evidence needed for a thorough investigation of the effectiveness of medical interventions). In the next section, potentially relevant generalised synthesis frameworks and methods for the combination of evidence for a human health risk assessment of environmental exposures are described.

3.5.2 *Methods for the generalised synthesis of human evidence*

In order to carry out an assessment of health technologies, Eddy (1989) points out that relevant data needed for such an assessment is likely to form a chain of evidence, with types of evidence informing different parts of the chain. With this in mind, Eddy proposed the *confidence profile method* for the synthesis of related evidence on health technologies. Under a Bayesian framework, the method allows different sources of information, including results from RCTs and epidemiological studies, case studies and expert opinion, to be synthesised for an evaluation of the effectiveness of a medical intervention. The method can, and has been, described by four steps (Hasselblad, 1994). Step one is to define the problem, including the intervention to be assessed and the health outcomes of interest. For each outcome, the next step (step two) is to identify and obtain all the available and relevant evidence, accounting for any biases in the estimation of effects from each piece of evidence. In step three the chain of evidence is defined in terms of probability distributions in a Bayesian model. It is solved in step four using a number of methods depending on the type and amount of evidence available and the structure of the models used to synthesise this evidence. Since a Bayesian model is proposed, prior information on particular parameters can be included to further inform the process (Eddy, 1989).

A number of examples exist in which the confidence profile method is used to synthesise evidence on the efficacy of interventions (Critchfield and Eddy, 1987; Gavin *et al.*, 1987; Hasselblad and Critchfield, 1987; Eddy *et al.*, 1988; Adar *et al.*, 1989). However, there appears to have been little use of the confidence profile method in the last 15 years. Medline only identified 5 articles that have explicitly reported use of the confidence profile method since 1990 (Shekelle *et al.*, 1992; Evans *et al.*, 1995; Hurwitz *et al.*, 1996; Klotzbeucher *et al.*, 2000; Lefevre and Aronson, 2000), although Spiegelhalter *et al.* (2004) show how such models can easily be implemented in WinBUGS. This can, perhaps, be explained by the complexity of the approach. Although developed to assess evidence for interventions in health care, Eddy (1989) comments that the applications are far reaching, including the social sciences and environmental sciences.

A few years after Eddy's proposal of the confidence profile method, a US General Accounting Office (GAO) report was published describing the efforts of the GAO to suggest a method for the evaluation of evidence from a number of different study designs on the effectiveness of medical interventions (GAO, 1992). Whereas the confidence profile method looked at evidence in terms of chains (Eddy, 1989; Ades and Sutton, 2006), the *cross design synthesis* approach was proposed by the GAO for the purpose of combining studies that have "...different, complementary designs..." in order to minimise the weaknesses and maximise the strengths of each study design (GAO, 1992). The report describes the cross design synthesis of evidence from RCTs and observational databases, such as clinical databases on cancer patients.

The cross design synthesis approach can be described in four steps. The first involves assessment of the different types of evidence and generalisability of each study's results. The GAO report illustrates this step with assessment of different RCTs, highlighting investigation of patient selection and recruitment, and the likely representativeness of randomised patients. From this assessment decisions on whether combining just the RCTs will provide the relevant information on the population of interest, or whether data from other types of studies also needs to be combined: leading to a cross design synthesis. In the GAO report evidence from observational databases is also assessed, with particular attention paid to the internal validity of this data source. Once this is achieved, the second step involves assessment of biases within each individual study and making appropriate adjustments for these biases. Although the GAO report suggests two possible approaches (either exclusion of the more biased studies, or secondary adjustment to account for those biases within each study), they advocate adjustment rather than exclusion since one of the aims of the cross design synthesis approach is to maximise the strengths of each study. Steps 3 and 4 involve combining the evidence within study type (i.e. RCT or database) and then synthesis across study type. Using a meta-analysis framework, the GAO report (1992) suggests the use of methods for stratification, weighting and extrapolation to account for differences between the study designs being synthesised. The authors stress that this method requires many judgments to be made throughout the process, and further exploration as to how these subjective judgments can be minimised is needed (GAO, 1992).

Both the confidence profile method and the cross design synthesis approach are applicable to the evaluation of the diverse evidence available for an assessment of risks to human health from exposure to environmental chemicals. However, in the examples used in Chapters 5 and 6 (Trihalomethane exposure and Mn exposure), a cross design synthesis approach is more appealing as different sources of evidence are used to inform the same outcome, rather than there being an explicit chain of evidence.

Following the ideas of a cross design synthesis approach, Prevost *et al.* (2000) and Sutton and Abrams (2001) combine evidence from RCTs with evidence from observational studies to investigate the effectiveness of breast cancer screening and electronic foetal heart rate monitoring, respectively. The authors acknowledge that relevant evidence for such investigations comes from both sources (RCTs and observational studies) and that each source has its strengths and weaknesses. For instance, Prevost *et al.* (2000) argue that although data from RCTs may represent the 'gold standard' and provide information on *efficacy* of an intervention, data from observational studies can provide details on the *effectiveness* of an intervention in clinical practice (i.e. outside a RCT). Thus, evidence from both types of study is desirable.

In Prevost *et al.* (2000), the RCT and observational evidence are initially analysed separately by the usual Bayesian random effects meta-analysis model given in Equation 3.5. A number of plausible, but diffuse, prior distributions are placed on the unknown parameters; μ , the true overall relative risk of breast cancer mortality with screening, and τ^2 , the between-study heterogeneity estimate. In Sutton and Abrams (2001), the RCT evidence is meta-analysed using this same Bayesian random effects models, with similarly plausible and diffuse prior distributions placed on the unknown parameters. Their results are presented alongside results from classical fixed and random effects meta-analysis models. To incorporate the observational evidence, Sutton and Abrams (2001) construct informative prior distributions for the true overall difference in the risk of perinatal mortality, μ , using the observational evidence. These prior distributions depend on beliefs about the quality and relevance of the observational data to the research question. Three

'beliefs' are defined and illustrated by Sutton and Abrams. Construction of each informative prior distribution based on the observational evidence requires taking summary estimates from a random effects meta-analysis of the observational data and adjusting (or not adjusting) the variance of this summary estimate. The three 'beliefs' and informative prior distributions are described in Table 3.1.

Table 3.1 Construction of informative prior distributions for μ based on observational evidence in Sutton and Abrams (2001)

Type of belief	Adjustment of variance of pooled observational evidence
Naive	None required
Equivalent	Make equivalent to variance of pooled RCT evidence
Down-weighted	Make larger than variance of pooled RCT evidence

Prevost *et al.* (2000) and Sutton and Abrams (2001) also describe a three-level random effects hierarchical Bayesian model to combine the RCT and observational evidence. The three levels can be thought of as study level, study type level and population level (Sutton and Abrams, 2001). The model is given by

$$y_{ij} \sim N(\psi_{ij}, \sigma_{ij}^2) \quad \psi_{ij} \sim N(\theta_j, \tau_j^2) \quad \theta_j \sim N(\mu, \nu^2) \quad (3.10)$$

where y_{ij} is the observed effect in study i of study type j with σ_{ij}^2 representing the (assumed known) variance of y_{ij} . ψ_{ij} is the true effect in study i , study type j , θ_j is the true effect in study type j and τ_j^2 is the estimate of the variance between studies of type j . μ is the overall effect across study type with ν^2 representing the estimate of the variance between study types. In both articles, plausible diffuse prior distributions are placed on the unknown parameters μ , τ_j^2 and ν^2 . Results from the Bayesian random effects meta-analysis models with diffuse, and informative (Sutton and Abrams, 2001), prior distributions are compared to the findings of the three-level hierarchical Bayesian model (Equation 3.10). All authors caution that the sensitivity of estimates to specifications of the prior distributions should be assessed, especially between study type variance.

Extensions to this three-level hierarchical Bayesian model are discussed by Prevost *et al.* (2000) and Sutton and Abrams (2001). They include placing constraints on the relevance of one study type compared to a second study type (Prevost *et al.*, 2000; Sutton and Abrams, 2001). Using the notation of Equation 3.10 (where μ is the overall effect, θ_1 is the true effect in RCTs and θ_2 is the true effect in observational studies), this involves specifying a constraint of the following form

$$|\mu - \theta_1| < |\mu - \theta_2| \quad (3.11)$$

Prevost *et al.* (2000) also discuss the possible weighting of the observational evidence in terms of its relevance to the RCT evidence in the usual Bayesian random effects model (Equation 3.5). This approach is similar to the weighting of the observational evidence as a prior distribution for the RCT evidence carried out by Sutton and Abrams (2001) and given in Table 3.1.

Further developments in methods for the generalised synthesis of evidence include combining qualitative evidence with quantitative evidence. To investigate factors affecting the uptake of childhood immunisation, Roberts *et al.* (2002b) reviewed the relevant available evidence. This evidence came from qualitative and quantitative studies. The qualitative evidence came from studies using focus groups, postal questionnaires and interviews based on open-ended questions. The quantitative evidence included surveys of patient satisfaction. Using a Bayesian fixed effects meta-analysis model (see Equation 3.1), the quantitative evidence was synthesised, with informative prior distributions placed on μ , the pooled log odds ratio for patient satisfaction with general practice care. The informative prior distributions were initially constructed from the authors prior beliefs of factors relevant to immunisation uptake. These beliefs were revised in light of findings from an analysis of the qualitative evidence. These up-dated beliefs formed the prior distributions for μ in the synthesis of the quantitative evidence. Roberts *et al.* (2002b) conclude that an analysis of purely quantitative or purely qualitative evidence would have omitted potentially important factors affecting the uptake of childhood immunisation.

More recently, Spiegelhalter and Best (2003) describe and discuss the use of Bayesian hierarchical random effects models to synthesise relevant, yet disparate, sources of evidence for cost-effectiveness analyses. In line with methods used by Prevost *et al.* (2000) and Sutton and Abrams (2001) for down-weighting certain types of evidence, methods to down-weight potentially biased evidence are also presented by Spiegelhalter and Best (2003). Each relevant study is intending to estimate the parameter of interest, δ , but because of differences in study population, for instance, between types of studies, the biased parameter, $\delta + \delta_h$, is being estimated in each study, where δ_h is termed ‘external bias’; $\delta_h \sim N(0, \sigma_h^2)$. These studies may also be subject to internal bias, due to their quality. Therefore, in each study the parameter being estimated is $\delta + \delta_h + \delta_b$ (where δ_b is the internal bias parameter, $\delta_b \sim N(0, \sigma_b^2)$). Thus, for study i

$$\begin{aligned} \delta_i &\sim N(\delta, \sigma_h^2 + \sigma_{bi}^2) \\ &\sim N\left(\delta, \frac{\sigma_h^2}{q_i}\right) \end{aligned} \quad (3.12)$$

where $q_i = \frac{\sigma_h^2}{\sigma_{bi}^2 + \sigma_h^2}$. Spiegelhalter and Best refer to q_i as the ‘quality weight’ for each study; this could be equal to 1 for a high-quality human RCT and so taken at face-value, or equal to 0.1 for a non-randomised study, resulting in a down-weighting of this evidence in the synthesis model. A cost-effectiveness analysis for choice of two hip replacement prostheses based on the time to revised replacement (the revision hazard ratio) is used to illustrate the use of a Bayesian hierarchical random effects model for the synthesis, and down-weighting, of relevant evidence. Revision rates from three different studies are used in the cost-effectiveness analysis: i) a registry of hip replacements in Sweden (non-randomised), ii) a RCT, and iii) a case series (non-randomised). The parameter of interest, HR_k , describes the revision hazard ratio (for new compared to usual prosthesis) in study k ($k:1-3$) is given by

$$\begin{aligned}
r_{ik} &\sim \text{Bin}(p_{ik}, N_{ik}) \\
\log(-\log(1 - p_{2k})) &= \log H_{1k} \\
\log(-\log(1 - p_{2k})) &= \log H_{2k} = \log H_{1k} - \log HR_k
\end{aligned} \tag{3.13}$$

Where i ($i: 1, 2$) denotes the prosthesis type, N_{ik} is the total number of patients receiving prosthesis i in study k and r_{ik} is the number of patients in study k with prosthesis i requiring a revision operation (Spiegelhalter and Best, 2003). Using Equation 3.12,

$$\log HR_k \sim N\left(\log \overline{HR}, \frac{\sigma_h^2}{q_k}\right) \tag{3.14}$$

The following informative prior distribution is placed on the between-study standard deviation, $\sigma_h \sim N(0.2, 0.05^2)$. Results are presented assuming 1) all three datasets have q_k (from Equation 3.14) equal to one (i.e. no down-weighting of any evidence) and 2) the non-randomised evidence is down-weighted, using two formations: a) $q_1 = 0.1$ (hip registry), $q_2 = 1$ (RCT) and $q_3 = 0.2$ (case series) and b) $q_1 = 0.1$, $q_2 = 1$ and $q_3 = 0.05$. These results demonstrate the sensitivity of the pooled HR_k to differing weightings of the evidence (Spiegelhalter and Best, 2003).

3.6 Summary

There has been little use and development of methods for the synthesis of human and animal evidence to assess risks to human health from exposure to chemicals in the environment. However, there has been an increase in the use and development of generalised evidence synthesis methods in human health care research, some of which describe models potentially relevant to the environmental risk assessment context. Models for the synthesis of human and animal evidence, multiple toxicological datasets and human evidence from related but different study designs have been reviewed in this chapter. The confidence profile method (Eddy, 1989) and the cross design synthesis approach (GAO, 1992) have also been described as relevant frameworks for the synthesis of human and animal evidence in an environmental risk assessment context. The cross design synthesis approach in

particular looks appealing, because the evidence is parallel and not in sequence, to the synthesis of related human and animal evidence for a risk assessment of exposure to environmental chemicals.

A common theme running through all but one of the approaches described in this chapter is the use of Bayesian hierarchical models. Such models provide flexibility in the modelling of realistically complex relationships. For instance, allowing additional data or beliefs on the form of relationships between types of data to be incorporated into the synthesis of evidence via prior distributions (DuMouchel and Harris, 1983; DuMouchel and Groër, 1989; Dominici *et al.*, 1997; Prevost *et al.*, 2000; Sutton and Abrams, 2001). Bayesian models are also appealing since the extent to which different types of evidence are considered relevant to the main aim of the synthesis (in most cases here that aim is to inform human health) can be controlled by placing constraints on both the parameters and the weights given to various pieces of evidence in the model.

Bayesian hierarchical models have been used in a variety of health and non-health related contexts to model simple and more complex hierarchical evidence from single and multiple studies, some of which have been described in this chapter (see Sections 3.4 and 3.5). In addition to those described in this chapter, Bayesian hierarchical models have been used for the analysis of human epidemiological studies with spatial and/or temporal aspects (Richardson *et al.*, 1995; Mueller *et al.*, 2001; Lopez-Vizcaino *et al.*, 2002; Gemperli *et al.*, 2004; Boyd *et al.*, 2005; Peng *et al.*, 2005), the analysis of financial data (Gallizo *et al.*, 2002; Talih and Hengartner, 2005) and weather forecasting data (Hoar *et al.*, 2003; Fox and Wikle, 2005), in the analysis of veterinary studies (Hirst *et al.*, 2002; Ranta *et al.*, 2005; Stevenson *et al.*, 2005) and genetic studies (Sillanpaa *et al.*, 2001; Kuhnert and Do, 2003; Bae and Mallick, 2004; Waldmann *et al.*, 2005).

The naturally hierarchical structure of the data in these types of studies necessitates a hierarchical model. However, the use of *Bayesian* hierarchical models in these articles provides many advantages, e.g. flexibility in the formation of the model by inclusion of prior information (Bae and Mallick, 2004; Talih and Hengartner, 2005; Ranta *et al.*, 2005; Stevenson *et al.*, 2005; Boyd *et al.*, 2005), the use of Bayes

factors to select the most appropriate model (Gallizo *et al.*, 2002) and the ease in obtaining predictive distributions (Gallizo *et al.*, 2002) and direct probabilities from a Bayesian analysis (Hirst *et al.*, 2002). Such advantages are attractive for the synthesis of human and animal evidence for human health risk assessments.

A second similarity in the synthesis models described here is the use of random effects over fixed effects models. The rationale for one over the other depends on the assumption that all studies are estimating the same effect. In analyses where different types of studies are combined, use of a fixed effects model may not be appropriate because of the assumption that the only variability between estimates is due to chance. A random effects model allows for excess heterogeneity between study estimates that is likely to be present when combining estimates from different study types. Furthermore, this heterogeneity is modelled and provides information on where and how studies may differ in their estimate of the effect. For instance, in the three-level hierarchical model given in Equation 3.10, variation both *between* study estimates within a study type and *across* study types is modelled.

Consideration of a random effects model leads to discussion of a third feature common to some of these models: the use of a hierarchical structure. Such a framework allows explicit description and estimation of the inter-relationships between data sources, which can lead to a more informed understanding of study differences and between-study heterogeneity.

A fourth feature to consider is the modelling of dose-response relationships. The use of these relationships and the subsequent dose-response slope estimate is likely to be required in applications of synthesis methods to risk assessment examples (as in Chapter 5), as they are in all of the articles concerned with environmental effects described in this chapter (DuMouchel and Harris, 1983; DuMouchel and Groër, 1989; Cox and Piegorsch, 1994; Dominici *et al.*, 1997; Simmonds *et al.*, 2003).

In this chapter general meta-analysis and more sophisticated synthesis models have been described. Use of the cross design synthesis approach, Bayesian hierarchical models, random effects and dose-response relationships are all likely to be useful in a risk assessment context and the models described here look promising. A cross design synthesis approach allows close examination of the relevance, quality and

the strengths and weaknesses of one source of evidence over other sources. DuMouchel and colleagues (DuMouchel and Harris, 1983; DuMouchel and Groër, 1989) have already applied similar models to human and animal evidence to inform human health risk assessments, where different species and exposures are modelled using a Bayesian random effects regression (Equation 3.6). In this thesis the generalised evidence synthesis models presented by Prevost *et al.* (2000) and Sutton and Abrams (2001) are applied to the relevant human and animal evidence. So far, these models have only been applied to the synthesis of different types of human evidence, RCTs and observational studies. The generalised evidence synthesis models differ to the models used by DuMouchel *et al.* in that a random effect is placed on the pooled human and pooled animal parameters, rather than the relevance of different species to effects in humans being explicitly modelled (as with DuMouchel's model). Investigation of the application of these models to evidence for human health risk assessment provides a framework for the building of more complex hierarchical models, as is clearly demonstrated in Chapter 5 and also allows assessment and investigation of methods for down-weighting particular sources of evidence (such as those described in Prevost *et al.* (2000), Sutton and Abrams (2001) and Spiegelhalter and Best (2003)). In Chapters 5 and 6, the use of systematic review and meta-analysis methods within a cross design synthesis approach to combine diverse data for human health risk assessments of exposure to environmental chemicals is explored. However, before these models are applied to the relevant human and animal evidence, it is important to consider current use of systematic review and meta-analysis methods to combine animal evidence alone. Although some authors advocate the use of these methods for animal evidence (Roberts *et al.*, 2002a; Sandercock and Roberts, 2002; Pound *et al.*, 2004; Macleod *et al.*, 2005) and provide some guidance for their application in the risk assessment context (McKnight, 1992) the extent to which they are carried out is unknown. In the next chapter, a systematic review of the use of systematic review and meta-analysis methods for the review and evaluation of animal experiments is reported. The aim of this systematic review is to assess the extent, and quality, of these methods when applied to animal evidence.

Systematic reviews and meta-analyses applied to animal experiments

4.1. Chapter overview

Limitations in current methods for human health risk assessment and the potential use of systematic reviews and meta-analyses to overcome some of these limitations have been considered in Chapter 2. The extent to which systematic review and meta-analysis methods are used to review animal experiments to inform human health is unknown. In this chapter details of a systematic review to establish the extent and quality of systematic reviews and meta-analyses of animal experiments are described. In Section 4.2, recent literature discussing the use of systematic reviews and meta-analyses to review animal experiments are presented. Details of the systematic review are given in Section 4.3, alongside a brief review of current guidelines for reporting systematic reviews and meta-analyses. In Section 4.4 the methods reported in the included articles are reviewed with respect to established guidelines and recommendations for conducting and reporting systematic reviews and meta-analyses. As a consequence of this critique, guidelines for good quality reporting of systematic reviews and meta-analyses of animal experiments are developed and presented in Section 4.5. A quality assessment of each meta-analysis is then made using these guidelines, although the question still remains as to whether the quality of the reporting or the quality of the actual meta-analysis is being assessed, or a combination of the two. In Section 4.6 the findings of the systematic review and the quality assessment are discussed.

4.2 Introduction

With the increase in evidence-based medicine, systematic reviews and meta-analyses are now commonly used to review human evidence from RCTs and observational epidemiology studies on matters of health. The advantages of this approach to identify, review and evaluate the available evidence have been noted by many researchers (Sackett *et al.*, 1996; Egger and Davey Smith, 1997; Sutton *et al.*, 2000; Egger *et al.*, 2001; Cochrane Collaboration, 2005). Sometimes, relevant evidence may also come from animal experiments in addition to human studies and so this information should be reviewed and evaluated as well. Recently there has been some debate about the usefulness of using systematic review and meta-analysis methods to review animal experiments to inform matters related to human health care (Pound *et al.*, 2004; Roberts *et al.*, 2002a; Sandercock and Roberts, 2002).

In a risk assessment context, animal experiments are often the basis for an assessment of the human health effects from exposure to chemical substances in the environment (e.g. cadmium, dioxins (CDC, 2005)). In these cases the available human evidence is usually limited and so animal experiments are the main source of evidence. The potential of meta-analyses to help in an evaluation of evidence for such risk assessments has been reported (e.g. McKnight, 1992), and although it has been stated that 1 in 10000 animal records in Medline were tagged 'meta-analysis' (Roberts *et al.*, 2002a), the extent to which these methods are used to evaluate animal evidence is unknown. It is important to assess the quality of these systematic reviews and meta-analyses, particularly as many authors are advocating their use. In this chapter the extent to which systematic reviews and meta-analyses have been used to review and evaluate data from animal experiments, the methods used and the quality of reporting of these methods are investigated.

4.3 Methods

4.3.1 Systematic review of the literature

The electronic databases MEDLINE (1966 – July 2005), EMBASE (1980 – July 2005), TOXLINE (1945 – July 2005) and ScienceDirect (1900 – July 2005) were searched to identify reports of the application of systematic reviews and/or meta-analyses to animal experiments. The search strategy is given in Appendix B (Table B.1). As some systematic reviews and/or meta-analyses may not find their way to being published as journal articles indexed in electronic databases, an electronic search of the grey literature was also undertaken. The grey literature covers reports and papers that may be prepared for funding bodies, governments and committees that are unlikely to be published as articles in peer-reviewed journals, for example, or book chapters, theses, conference and meeting abstracts (Hartling *et al.*, 2005). The websites, databases and keywords used to search the grey literature are also given in Appendix B (Table B.2).

The following criteria were used to identify relevant systematic reviews and meta-analyses of animal experiments. For systematic reviews, details on the source(s) of evidence searched and some information on at least one of the following were sought,

- Search terms used
- Inclusion and exclusion criteria
- Any limitations placed on the search.

For meta-analyses, a report of some quantitative synthesis of results of more than one experiment was required. Unlike meta-analyses in medicine, some of the meta-analyses identified in this chapter did not use systematic review methods to identify the data used in the meta-analysis. Regardless of whether a systematic review was used or not, all meta-analyses of animal experiment data were sought.

The reference lists of all relevant articles identified from these searches were assessed and my own files were searched to identify further pertinent studies. There were a number of criteria for the inclusion of articles into this systematic review. The systematic reviews and/or meta-analyses had to involve *in vivo* animal experiments; where the purpose of reviewing animal evidence was to inform human

health regarding 1) a medical intervention, 2) an epidemiological association (e.g. physical activity and cancer) or 3) the effect of an exposure to a chemical substance (e.g. diesel exhaust). Articles were considered relevant even if human evidence was also sought in addition to the animal evidence in the systematic review and/or meta-analysis.

4.3.2 Current guidelines for reporting systematic reviews and meta-analyses

Considerable attention has been paid to the quality of reported systematic reviews and meta-analyses including many articles that have critically appraised these types of studies (e.g. Jadad *et al.*, 1998; Christensen, 2001; Shea *et al.*, 2002; Hemels *et al.*, 2004; Dixon *et al.*, 2005). In addition to these, reports from conferences and meetings have developed into guidelines for the reporting of systematic reviews and meta-analyses. The most recognised of these have resulted in the QUOROM (Quality Reporting of Meta-analyses: Moher *et al.*, 1999) and MOOSE (Meta-analyses of Observational Studies in Epidemiology: Stroup *et al.*, 2000) statements, although other guidelines existed prior to these (e.g. Cook *et al.*, 1995). As a demonstration of the common use of the QUOROM and MOOSE statements, Web of Science (<http://wok.mimas.ac.uk>) holds 465 and 288 citations, respectively, for these two articles (as of 31st January, 2006). Moreover, a number of prominent journals now encourage submitted systematic reviews and meta-analyses to follow guidelines set out in the QUOROM and MOOSE statements (e.g. BMJ, JAMA, Annals of Internal Medicine).

Further guidance can be found in the Cochrane Reviewer's Handbook (Higgins and Green, 2005) and from the NHS Centre for Reviews and Dissemination (Deeks *et al.*, 2001). Indeed, there is a plethora of information available to guide researchers undertaking systematic reviews and meta-analyses in many areas of interest. For example, meta-analyses of diagnostic studies (Irwig *et al.*, 1994), environmental epidemiology studies (Blair *et al.*, 1995), clinical trials (Moher *et al.*, 1999), observational studies (Stroup *et al.*, 2000), studies of prognostic markers (Riley *et al.*, 2003), genetic association studies (Munafò and Flint, 2004). However, there do not appear to be any guidelines for the conduct and reporting of systematic reviews and meta-analyses of animal experiments. In the next section, features of the methods reported in the relevant systematic reviews and meta-analyses of animal

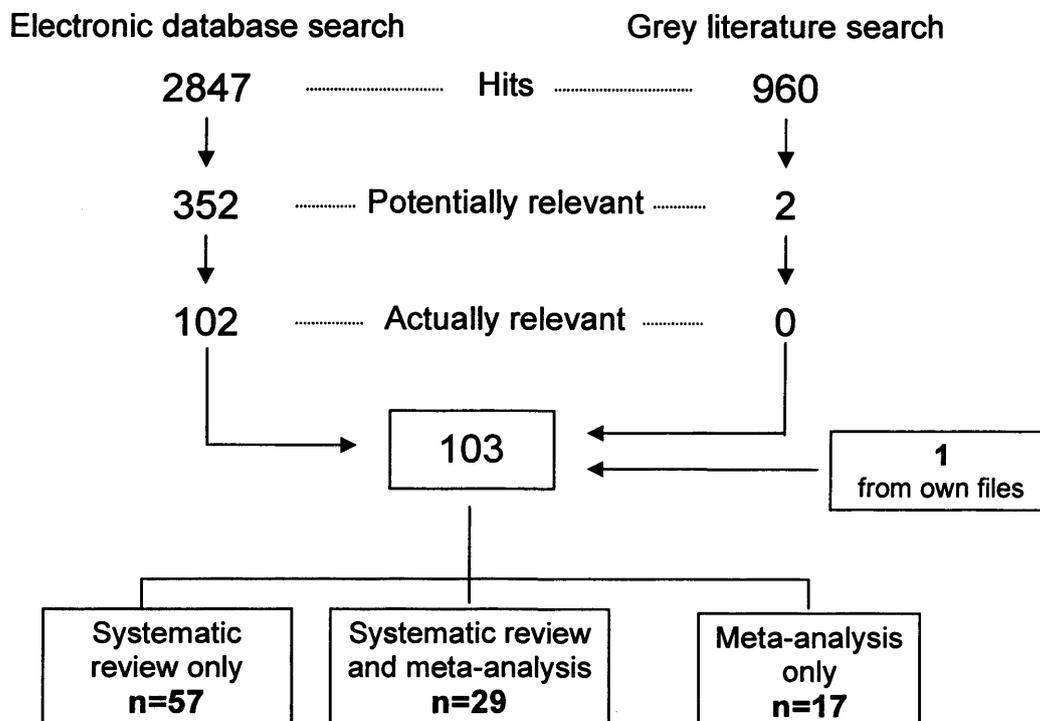
experiments are critiqued with respect to the guidelines mentioned in this section. This will then lead on to the development of guidelines specifically for systematic reviews and meta-analyses of animal experiments.

4.4 Results

4.4.1 Articles identified

The total number of relevant articles identified from the search strategies is given in Figure 4.1. The three articles reviewed in Section 3.4 (DuMouchel and Harris, 1983; DuMouchel and Groër, 1989; Cox and Piegorsch, 1994) are not included in this systematic review since their interest lies in the methodological aspects of synthesising multiple human and animal studies (this no doubt explains why they were not identified from the systematic review described in Section. 4.3.1).

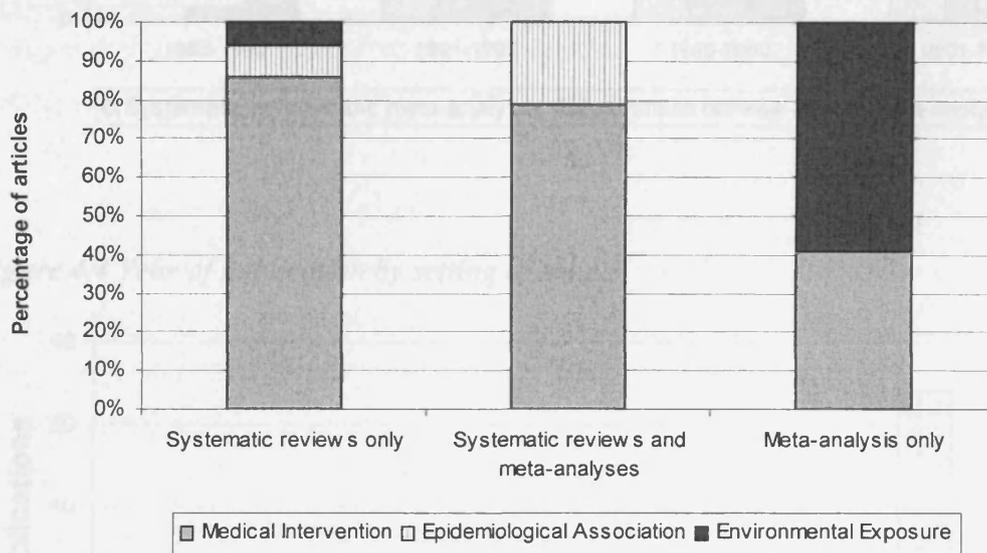
Figure 4.1 Relevant articles identified from the systematic review



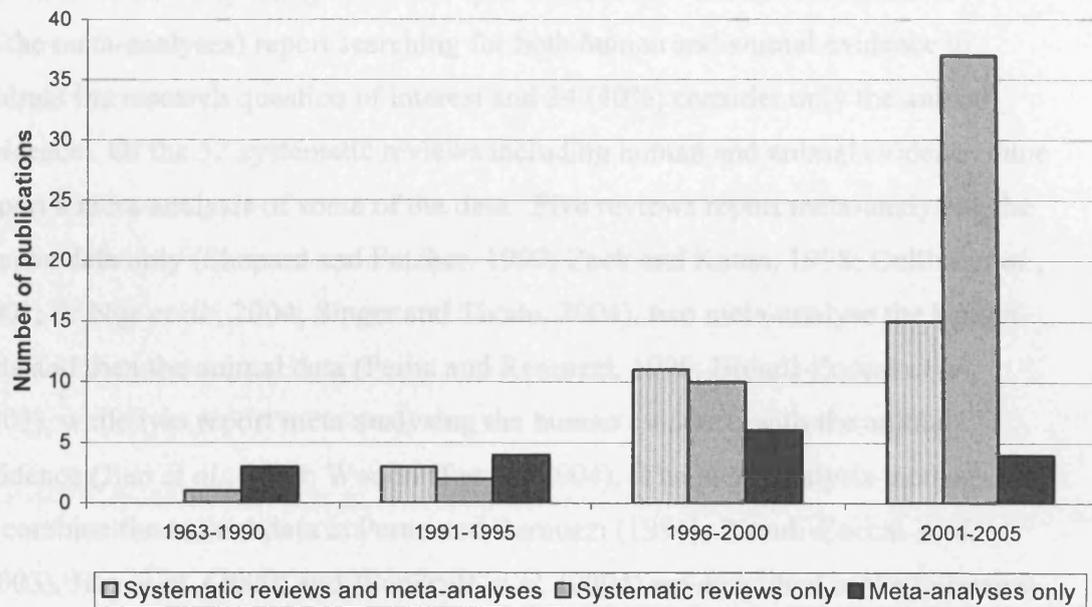
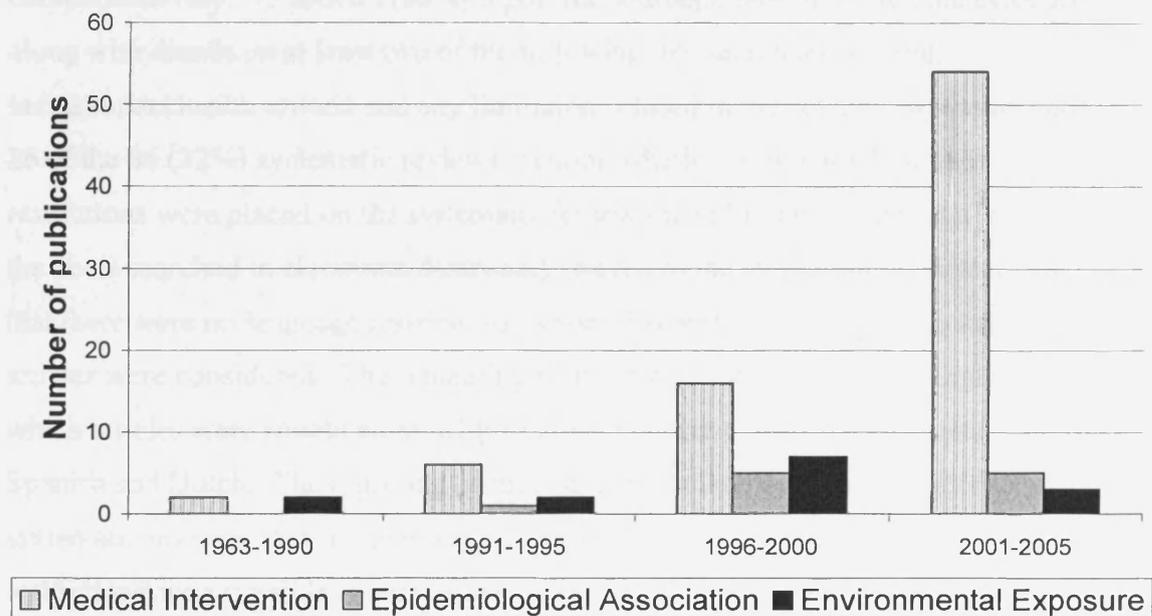
As in Figure 4.1, the 103 relevant articles can be split into three categories: i) those reporting details of a systematic review only ($n=57$), ii) those reporting a systematic review followed by a meta-analysis ($n=29$) and iii) articles only reporting details of a meta-analysis ($n=17$). A full list of these references is given in Appendix C.

Whether an article is a systematic review only, a systematic review and meta-analysis, or a meta-analysis only appears to be related to the setting of the article (i.e. whether it is evaluating a medical intervention, an epidemiological association of an environmental exposure) as Figure 4.2 illustrates.

Figure 4.2 Settings for the systematic review and meta-analysis articles



There is a tendency for meta-analyses of effects from environmental exposures to be based on evidence that does not originate from a systematic review. There are no articles reporting a systematic review *and* meta-analysis in an environmental setting, and all reports on an epidemiological association involve a systematic review. There has been a large increase in the number of systematic reviews published over time, particularly in the last five years (Figure 4.3); most of these are articles reviewing medical interventions (Figure 4.4). The use of systematic review and/or meta-analysis methods to review evidence for environmental exposures has decreased in the last five years (2001- 2005) compared to the five years previous to that (1996 – 2000).

Figure 4.3 Year of publication by type of article**Figure 4.4** Year of publication by setting of article

4.4.2 Features of the systematic reviews

Of the 86 articles reporting the use of systematic review methods, 52 articles (60% of the meta-analyses) report searching for both human and animal evidence to address the research question of interest and 34 (40%) consider only the animal evidence. Of the 52 systematic reviews including human and animal evidence, nine report a meta-analysis of some of the data. Five reviews report meta-analysing the human data only (Shepard and Fitcher, 1997; Zock and Katan, 1998; Collins *et al.*, 2001; de Nijs *et al.*, 2004; Singer and Thode, 2004), two meta-analyse the human data and then the animal data (Perna and Remuzzi, 1996; Biondi-Zoccai *et al.*, 2003), while two report meta-analysing the human evidence with the animal evidence (Jiao *et al.*, 2000; Woodruff *et al.*, 2004). The meta-analysis methods used to combine the animal data in Perna and Remuzzi (1996), Biondi-Zoccai *et al.* (2003), Jiao *et al.* (2000) and Woodruff *et al.* (2004) are described in the following section.

Within these 86 articles, details of search strategies are generally given comprehensively. 73 articles (85%) report the source(s) used to search the evidence along with details on at least two of the following: the search terms used, inclusion/exclusion criteria and any limitations placed in the search. However, only 26 of the 86 (32%) systematic reviews mention whether or not any language restrictions were placed on the systematic review (most limitations reported refer to the years searched in electronic databases). Seven of the systematic reviews report that there were no language restrictions, while 13 report that *only* English language articles were considered. The remaining eight articles list a number of languages in which articles were sought along with English, they are German, French, Italian, Spanish and Dutch. The reporting of these details is clearly deficient in this set of systematic reviews. Since there is mixed evidence as to whether language restrictions are a possible source of bias in systematic reviews (Egger *et al.*, 1997; Song *et al.*, 2000; Egger *et al.*, 2003), it is important that these details are reported so that the reader is aware of such limitations and possible sources of bias.

In terms of the sources of evidence searched, this set of systematic reviews comprehensively report methods. 73 of the 86 (85%) systematic reviews report searching more than one source of evidence with two articles reportedly searching

ten different sources of evidence (Borrelli and Ernst, 2002; Borrelli *et al.*, 2003). Electronic databases are used in all but one systematic review to search the evidence. The one study not using electronic databases was published in 1963, before such databases were available. The databases searched include Medline, Embase, Toxline, Cochrane library, Complementary Medicine and Current Contents. To supplement these electronic database searches, many articles also report searching various other sources, including the reference lists of relevant articles (n=54, 63%), contacting authors for further (un)published data (n=10, 12%), searching conference and meeting proceedings/abstracts (n=8, 9%), contacting companies for further articles (n=7, 8%) and searching the internet (n=2, 2%).

Unfortunately details on any quality assessment of the primary animal experiments are inadequately reported. Only 20 of the 86 (23%) systematic reviews reported an assessment of the quality of the primary experiments. Eight of these used the results as further inclusion and exclusion criteria and six studies use them as discussion points. The remaining six systematic reviews meta-analyse the animal data and use the quality score as a variable for subgroup analyses in an attempt to explain any between-study heterogeneity. Although there are many issues concerning the use of quality scores (Jüni *et al.*, 1999), MOOSE recommends their use as subgroup variables (Stroup *et al.*, 2000).

4.4.3 Features of the meta-analyses

46 articles report the use of meta-analysis methods to combine animal experiments. 29 (63%) of these report a systematic review of the literature prior to the meta-analysis of the relevant data (and have been included in the review of systematic review methods above). The data used in the other 17 articles (37%) include that from a set of replicate experiments (Tachibana, 1989; Tachibana *et al.*, 1996), an established database of results of experiments (Crump *et al.*, 1999) and non-systematic reviews (Corpet and Tache, 2002; Eichacker *et al.*, 2002; Brown and Strickland, 2003). For the remaining eleven studies not using a systematic review there are few details on the origin or identification of the primary data used in the meta-analyses. The reader is therefore prevented from making an informed decision on the search process and whether the findings are based on a biased set of evidence. As pointed out, the majority of these meta-analyses are for environmental

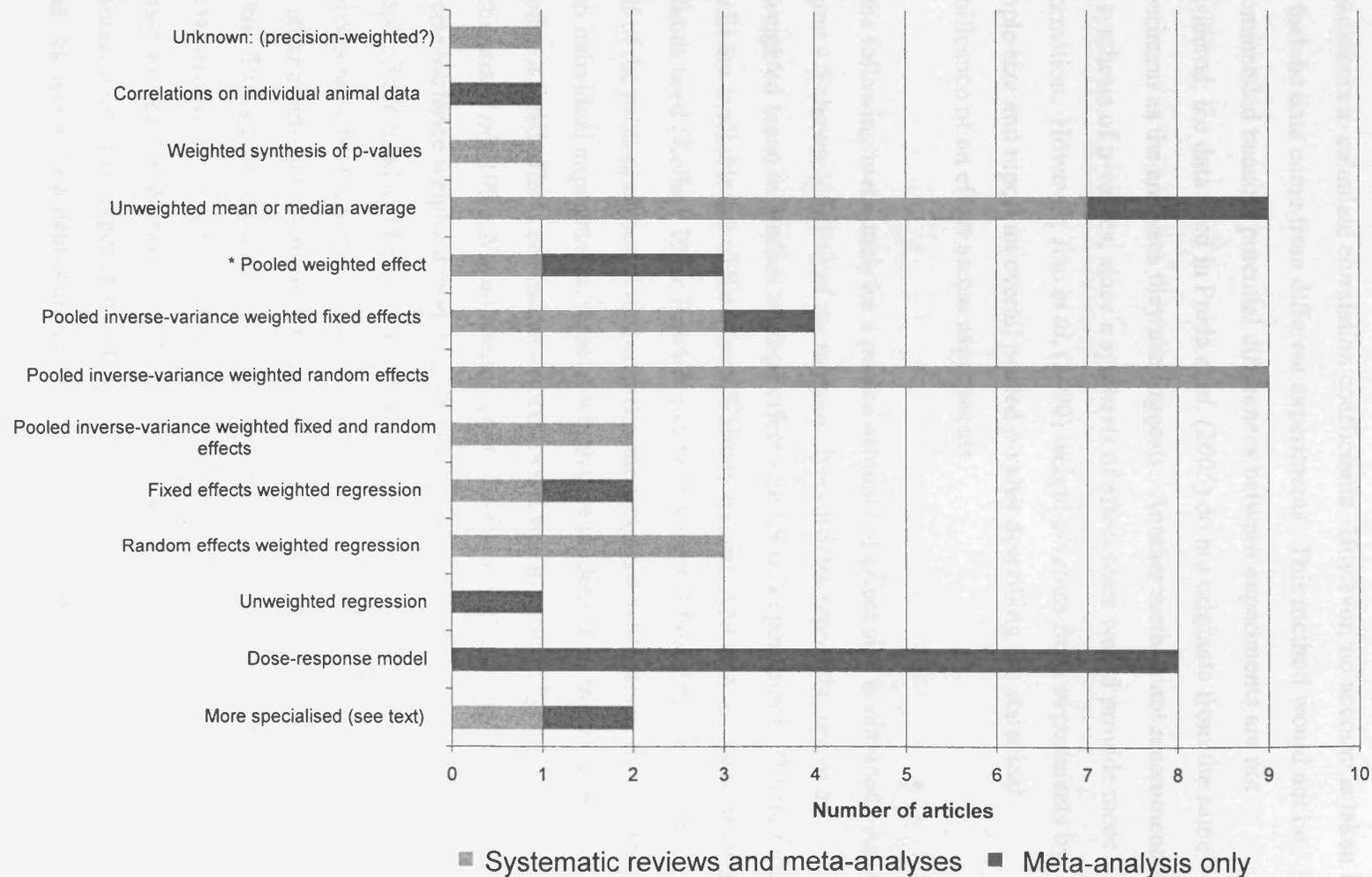
exposures; thus one cannot be fully confident of results since the basis for the meta-analysis is unclear.

The number of experiments combined in each article varies from meta-analyses of just three experiments (Horn *et al.*, 2001; Preda *et al.*, 2005) to one of 397 experiments (Crump *et al.*, 1999); the median number is 25, with 5 meta-analyses not reporting the number of experiments combined.

Only one study reports using individual animal data (Preda *et al.*, 2005), the rest synthesise aggregate data from each primary study. Obtaining individual animal data is likely to be more useful in a meta-analysis (Cook *et al.*, 1995), but it is often very hard to get hold of. None of the meta-analyses reviewed here reported seeking individual data from authors of the primary studies.

A variety of methodologies to synthesise data are used in these 46 articles. Individual summaries of the meta-analysis articles including the setting for the meta-analysis, the species/strain of animals included, the number of experiments included and some detail of the methods used (including effect estimates reported, whether and how heterogeneity was assessed, the synthesis methods used, any subgroup analyses and whether and how publication bias was investigated) are given in Appendix D.

Figure 4.5 illustrates the different methods used to synthesise evidence from individual animal experiments in the 46 meta-analyses. Although simple methods for obtaining a quantitative synthesis across experiments are popular, as are the usual inverse-variance weighted models described in Equations 3.1 and 3.2 in Chapter 3, so too are more complex models. Further description of the synthesis methods follows Figure 4.5.

Figure 4.5 Methods of synthesis used in meta-analyses with and without systematic reviews

* no details given on weights used and whether fixed or random effects model

At the top of Figure 4.5 is a meta-analysis where the method of synthesis used is not reported (Bertani *et al.*, 2002). The authors reference Cook *et al.* (1995) for the methods used and this suggests an inverse-variance weighted synthesis of the form described in Equations 3.1 and 3.2. One of the simplest approaches taken is by Preda *et al.* (2005) who combined the individual animal data from three experiments to calculate correlation coefficients. However, no account is taken for the fact the data came from different experiments. This method would not be recommended because potential differences between experiments are not considered; the data used in Preda *et al.* (2005) do not originate from the same experiment as the analyses they use suggests. Another method not recommend is the synthesis of p-values, since a synthesis of effect sizes would provide more information. However, Jiao *et al.* (2000) weight p-values from experiments by sample size and report an overall pooled p-value describing the statistical significance of an effect across experiments.

In the following meta-analyses a pooled estimate of effect size is obtained. As Figure 4.5 shows, 9 (20%) of the meta-analysis articles report the use of an unweighted mean or median average effect size across experiments. Where further details are available use of the Mann-Whitney test and ANOVA are cited among the methods used (Kelley, 1996; Rowlett and Woolverton, 1996; Pries *et al.*, 1998). Half of the meta-analyses (n=23) report some weighted synthesis of the evidence from individual experiments, three of which give no details on the weights used or whether a fixed effects or random effects model was assumed (Tachibana, 1989; Tachibana *et al.*, 1996; Nava-Ocampo *et al.*, 2000). Fixed and random effects inverse-variance weighted models of the form given in Equations 3.1 and 3.2 in Chapter 3 are used in 15 (33%) of the meta-analyses. All but one of these 15 meta-analyses have been carried out to investigate the efficacy of medical interventions: the other meta-analysis is investigating an epidemiological association (Glatt *et al.*, 2000). This suggests that use of an inverse-variance weighted meta-analysis model has translated from the methods used to synthesise human RCTs. In 5 (11%) of the meta-analyses, fixed or random effects weighted regression models are used to combine individual experiments, and in one meta-analysis unweighted regression is used. However, few details are given on these regression models in the articles.

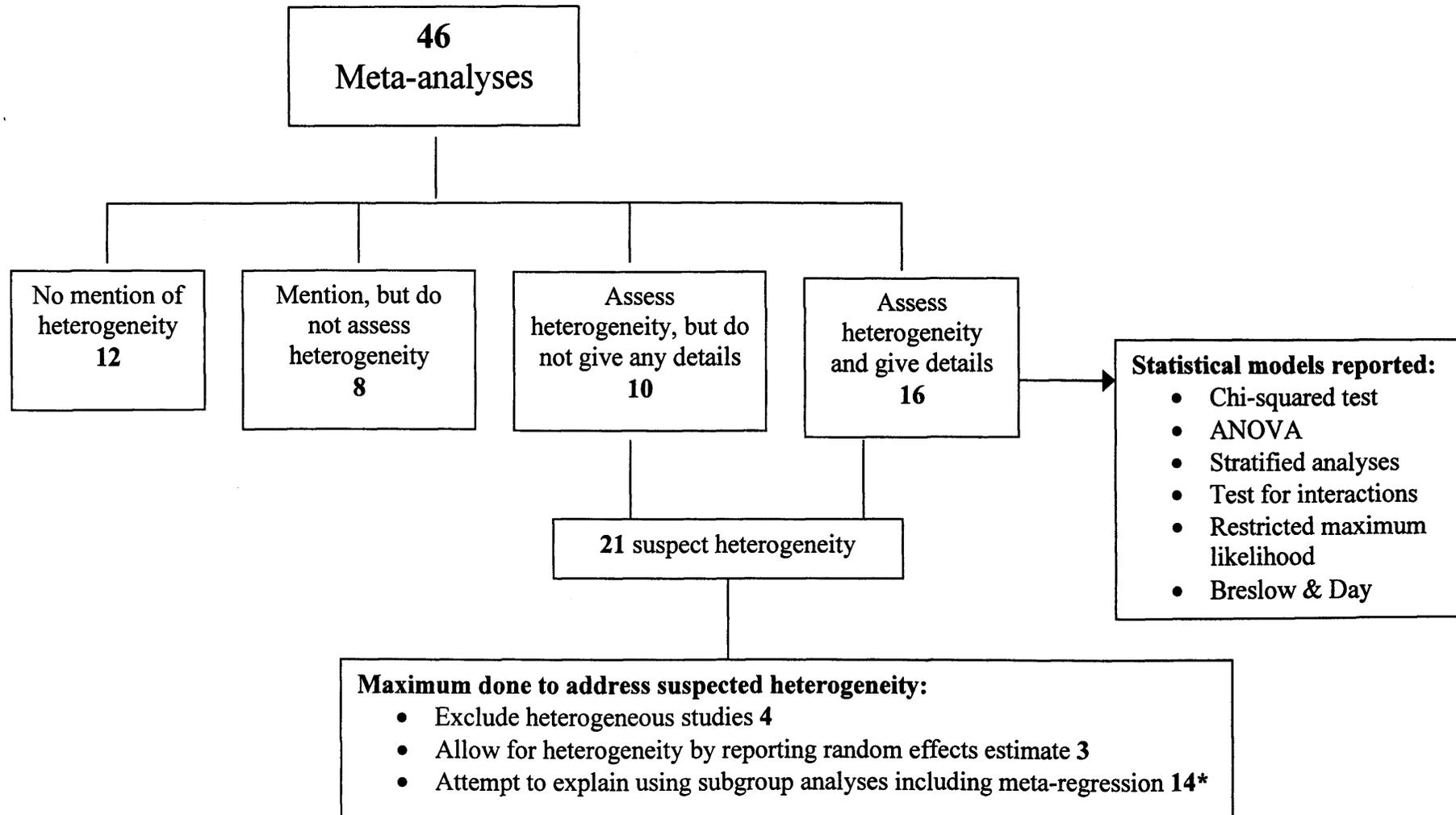
Of the 8 meta-analyses (17%) reporting use of dose-response models to synthesise evidence from individual experiments, six are from meta-analyses carried out to investigate health effects from exposure to environmental chemicals and two are estimating clinical effects of an intervention. It is interesting to note that none of these 8 meta-analyses were preceded by a systematic review. The different types of dose-response models used in these meta-analyses are summarised in Appendix D, with two authors citing the use and development of specialised software to apply the dose-response models.

Further specialised methods of synthesis reported in these meta-analyses are the meta-analysis of diagnostic data using a summary receiver-operator characteristic (SROC) curve (Craig *et al.*, 2000) and Crump *et al.* (1999) report modelling distributions of p-values from nearly 400 experiments.

Five of the 46 articles report the use of meta-analysis methods to combine human with animal data, and although the most appropriate method of analysis would be to account for the species differences, Kroll *et al.* (1993), Jiao *et al.* (2000) and Woodruff *et al.* (2004) combine the human and animal evidence without accounting for the fact data from different species are being synthesised. Carroll *et al.* (1994) and Guth *et al.* (1997), however, account for species differences using the regression model presented by Cox and Piegorsch (1994) described in Section 3.4.

As mentioned in Section 3.3.2, between-study heterogeneity is a widespread and important feature of meta-analyses (Engels *et al.*, 2000; Sutton *et al.*, 2000; Villar *et al.*, 2001). It is essential that the presence of between-study heterogeneity is not only considered in a meta-analysis, but also that possible sources are investigated, since any excess heterogeneity has implications for both the synthesis and inference of a meta-analysis (Sutton *et al.*, 2000). Figure 4.6 shows the number of meta-analyses reviewed in this chapter which do, and do not, report observing or assessing between-study heterogeneity. The methods used to assess between-study heterogeneity and how suspected heterogeneity is dealt with are also shown.

Figure 4.6 Flow diagram of how heterogeneity is considered in 46 meta-analyses



* this value includes studies that attempted to explain between-study heterogeneity and reported random effects estimate

Of the 21 meta-analyses in which between-study heterogeneity is suspected, four exclude the heterogeneous studies from the calculation of the pooled estimate of effect. The Cochrane Reviewer's Handbook states that if a heterogeneous study (or studies) is clearly an outlier then such an approach can be justified (Higgins and Green, 2005). However, they caution that there are often many reasons that can be found for why a particular study could be considered an outlier, hence such an approach could be dangerous and actually introduce bias into the meta-analysis. An attempt to explain why these studies were heterogeneous would be desirable, rather than just excluding them. However, this attempt was not made by any of these four meta-analyses. The meta-analyses reporting the random effects estimate rather than the fixed effects estimate when heterogeneity is suspected, allow for this between-study heterogeneity, but again, no attempt to explain possible sources of heterogeneity is made.

On the other hand, 14 out of 21 (67%) meta-analyses suspecting between-study heterogeneity report subgroup analyses to investigate possible sources of heterogeneity. This approach is advocated by many authors (Cook *et al.*, 1995; Moher *et al.*, 1999; Stroup *et al.*, 2000; Higgins and Green, 2005). However, there are important issues here as to the quality of these subgroup analyses. In some of the meta-analyses more than five variables are assessed in an attempt to explain heterogeneity and it is often unclear as to whether these analyses are *a priori* or *post hoc*. Interpretation of findings from these subgroup analyses should be made in light of whether they were pre-specified and how many were carried out, in addition to biological or methodological processes (Deeks *et al.*, 2001; Higgins and Green, 2005). Reports of meta-analyses of animal experiments must therefore clearly state the basis for any subgroup analyses and make sure they are pre-specified. This feature is inadequately reported by the meta-analyses reviewed here. Of particular importance to meta-analyses of animal experiments, are differences between animal species and strains. Seventeen of the 46 meta-analyses either failed to provide details on the species or strains used in the meta-analysis or gave no details on whether any differences were taken into account. 14 meta-analyses only included one species or strain of animal. Of the remaining 15 meta-analyses, five analysed different species separately and ten took species differences into account using regression methods.

Publication bias can significantly affect interpretation of a meta-analysis and is no less important in meta-analyses of animal experiments. Only 17 meta-analyses mention and consider, to some extent, publication bias. An assessment of publication bias is reported in six of these: funnel plots, Egger's regression test (Egger *et al.*, 1997) or the Failsafe number (Rosenthal, 1979) are generally used. The lack of any investigation into publication bias by authors of these meta-analyses is worrying although reasonable explanations are given by some authors: only a small number of studies are being reviewed (Craig *et al.*, 2000); "the current statistical procedures addressing this issue lack validity" (Kelley, 1996). This is a particularly important point from Kelley (1996) and in Chapters 7 and 8, performance of the usual tests for publication bias is investigated using simulation analyses.

In the systematic reviews and meta-analyses reviewed in this chapter there is a dearth of any graphical presentation of primary studies and overall estimates. Plots, such as forest and funnel plots, can be very informative and helpful to readers and are recommended in a number of guidance documents (Cook *et al.*, 1995; Moher *et al.*, 1999; Stroup *et al.*, 2000).

4.5 Quality reporting guidelines

4.5.1 Development of guidelines

As a consequence of the review of systematic reviews and meta-analyses in the previous section, guidelines for good quality reporting of systematic reviews and meta-analyses of animal experiments have been developed. These guidelines are largely based on QUOROM since animal experiments could be seen as analogous to human RCTs, but also include aspects of MOOSE. Further modifications make the guidelines specific to the reporting of meta-analyses of animal experiments and, from an evaluation of how well the 103 articles identified in this chapter reported their systematic reviews and meta-analyses, particular features have been incorporated that are poorly reported in the set of meta-analyses reviewed in Section 4.4, such as addressing species differences and whether literature searches are subject to any language restrictions. These guidelines are given in Figure 4.7.

Figure 4.7 Proposed guidelines for reporting systematic reviews and meta-analyses of animal experiments

Heading	Subheading	Descriptor
<i>Title</i>		Identify the report as a meta-analysis [or systematic review] of animal toxicology experiments
<i>Abstract</i>		Use a structured format
	<i>Objectives</i>	Describe explicitly the scientific question/ hypothesis
	<i>Data sources</i>	Describe the databases and other important information sources used
	<i>Review methods</i>	Describe the selection criteria (e.g. species, strain, intervention/exposure, outcome & study design); methods for validity assessment, data abstraction, and study characteristics, and quantitative data synthesis
	<i>Results</i>	Describe characteristics of the experiments included and excluded; qualitative and quantitative findings (e.g. point estimates and confidence intervals/ standard errors), stating clearly what is estimated: dose-response curves, LD50 etc; and subgroup analyses
	<i>Conclusion</i>	State the main results and their implications
<i>Introduction</i>		Describe the scientific problem explicitly, biological rationale for the intervention/exposure, and rationale for the review
<i>Methods</i>	<i>Searching</i>	Describe the information sources in detail (e.g. databases, registers, personal files, expert informants, agencies, hand-searching), including keywords, search strategy and any restrictions (years considered, publication status, language of publication)
		Describe special efforts to include all available data (e.g. contact with authors, searching the grey literature)
	<i>Selection</i>	Describe the inclusion and exclusion criteria (defining population, intervention/ exposure, principal outcomes, and study design) List excluded experiments and reasons for exclusion

	<i>Validity and quality assessment</i>	Describe the criteria and process used (e.g. blind assessments, quality assessment, and their findings)
	<i>Data abstraction</i>	Describe the process or processes used (e.g. completed independently, in duplicate), including details on reproducibility, inter-rate agreement. Whether aggregate data or individual animal data are abstracted
	<i>Study characteristics</i>	Describe the type of study designs, animals' characteristics (e.g. species, strain, age, sex), details of intervention/exposure (including route of administration, dose and duration), outcome definitions
	<i>Quantitative data synthesis</i>	Describe the principal measures of effect, method of combining results (e.g. fixed- and random-effects; meta-regression), handling of missing data; how statistical heterogeneity was assessed and investigated; how data from different species and strains were dealt with; adjustment for possible confounding variables; rationale for any a-priori sensitivity and subgroup analyses; and any assessment of publication bias – all in enough detail to replicate
<i>Results</i>	<i>Flow chart</i>	Provide a meta-analysis profile summarising experiment flow giving total number of experiments in the meta-analysis
	<i>Study characteristics</i>	Present descriptive data for each experiment (e.g. species, strain, age, sex, sample size, intervention/exposure, dose, duration)
	<i>Quantitative data synthesis</i>	Report agreement on the selection and validity assessment and relevance to the scientific question/ hypothesis; present simple summary results (e.g. forest plot); present data needed to calculate effect sizes and confidence intervals; identify sources of heterogeneity, impact of study quality and publication bias
<i>Discussion</i>		Summarise key findings; discuss scientific/clinical inferences and generalisability based on internal and external validity; interpret the results in light of the totality of available evidence, including data from human studies; discuss rationale for use of animal data to help inform human health outcomes; critically appraise potential biases in the review process (e.g. publication bias); suggest a future research agenda

To explore the relative quality of the meta-analyses of animal experiments, every one of the 46 meta-analyses (with or without a systematic review) has been assessed on the basis of the guidelines given in Figure 4.7. The findings are given in the following section.

4.5.2 Findings of the quality assessment

The quality assessment of the 46 meta-analyses suggest that reporting of methods, results and the discussion in the meta-analysis is particularly poor; so too are aspects of the abstract (first column of Table 4.1). However, every meta-analysis reviewed here adequately describes the scientific problem, biological rationale for the intervention/exposure, and rationale for the review. 27 of the 46 meta-analyses (60%) identify the article as a systematic review or meta-analysis in the title.

The quality of reporting is generally worse in meta-analyses concerning environmental exposures (n=10) than those concerning medical interventions (n=30) or epidemiological associations (n=6) as Figures 4.8a, 4.8b and 4.8c show. Part of this can be explained by the fact that none of the environmental exposure meta-analyses are preceded by a systematic review, so it is understandable that some details (e.g. information sources, validity and quality assessments) are less likely to be reported in these articles. Although environmental exposure meta-analyses are more likely to identify the use of animal experiments in the article's title, they appear less likely to fulfil elements of the abstract and methods sections compared to meta-analyses of animal experiments investigating a medical intervention or epidemiological association (Figures 4.8a, 4.8b and 4.8c).

Figure 4.8a Percentage of meta-analyses, by setting, fulfilling criteria from the proposed guidelines: title and abstract

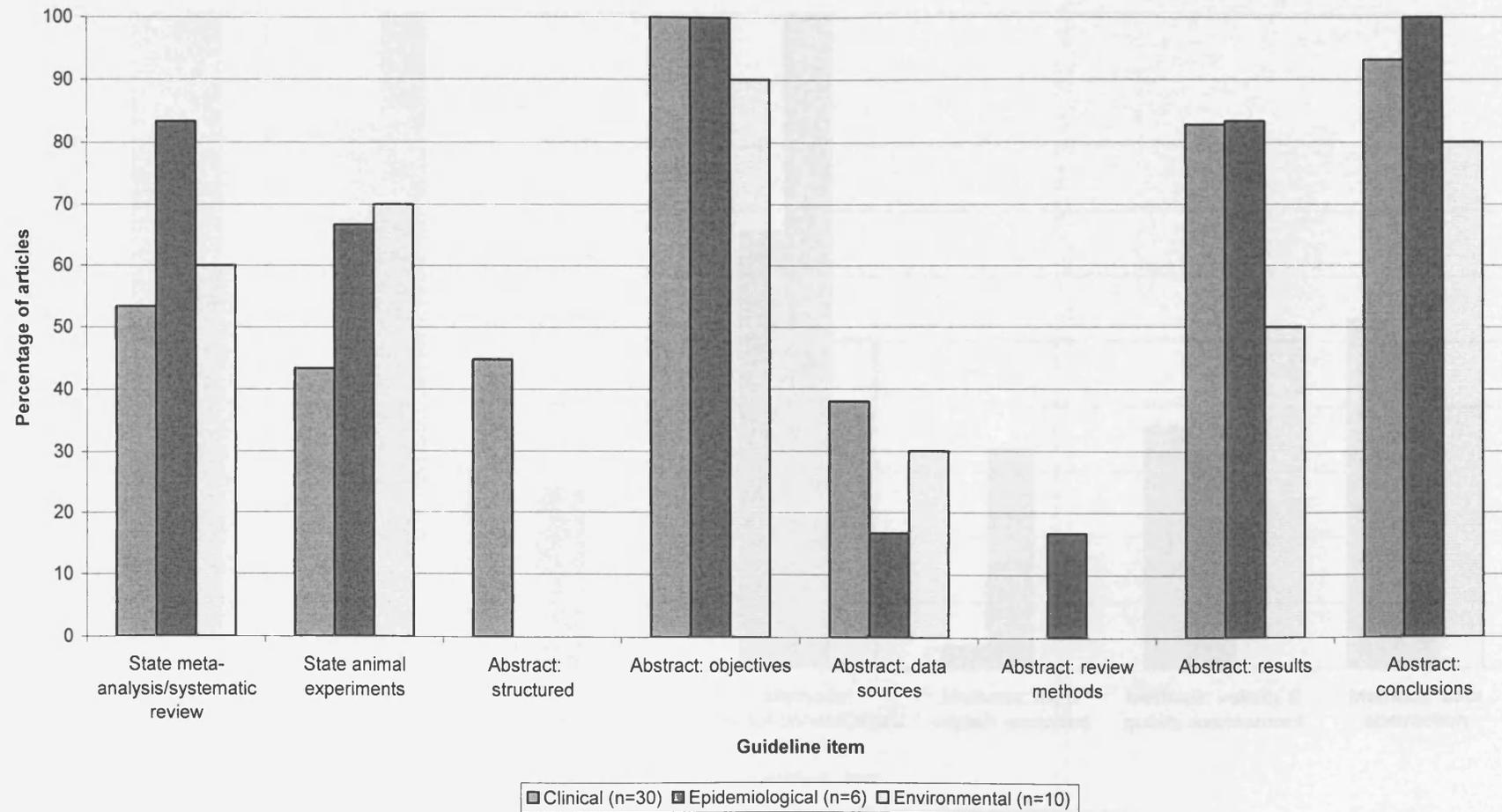


Figure 4.8b Percentage of meta-analyses, by setting, fulfilling criteria from the proposed guidelines: introduction and methods

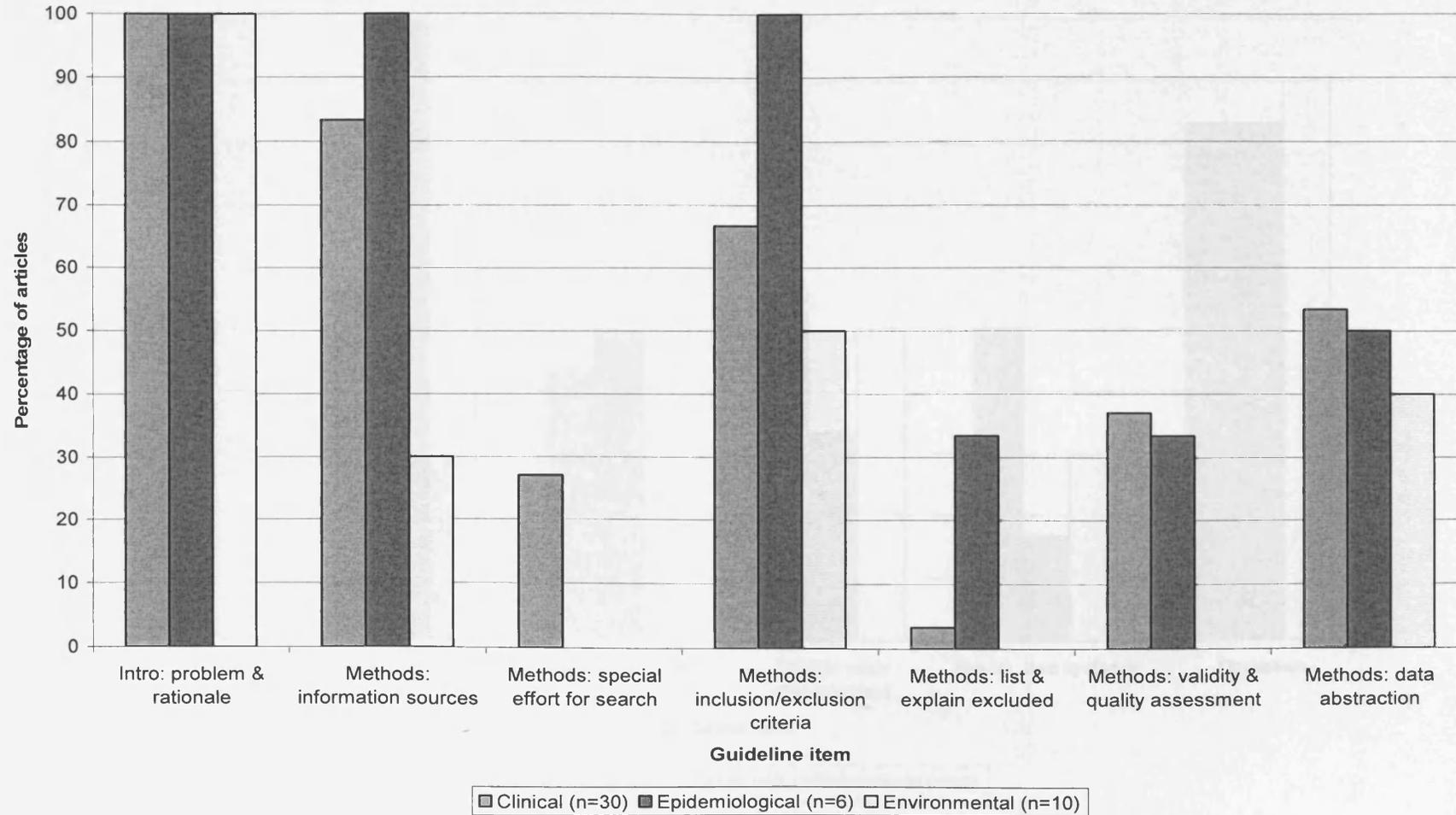
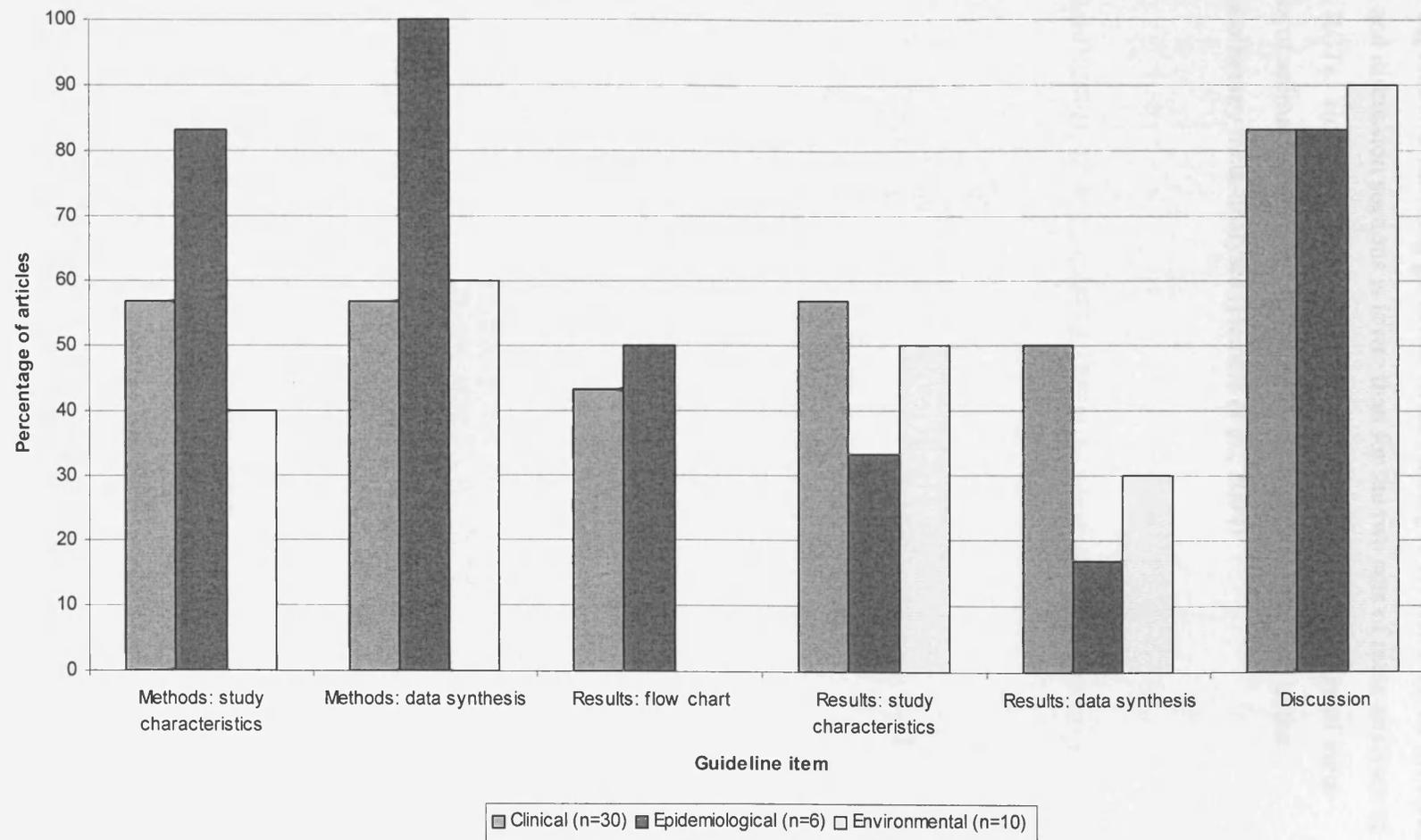


Figure 4.8c Percentage of meta-analyses, by setting, fulfilling criteria from the proposed guidelines: results and discussion



Considering the set of 46 meta-analyses of animal experiments as a whole, when compared with two sets of meta-analyses of human RCTs (Christensen *et al.*, 2001; Hemels *et al.*, 2004), Table 4.1 shows that the percentage of meta-analyses of animal experiments fulfilling guideline items relating to elements of the methods, results and discussion sections is lower than for the two sets of meta-analyses of human RCTs. However, for many of the other items, the percentage of meta-analyses of animal experiments fulfilling items is greater than that in the pharmacotherapy meta-analyses (Hemels *et al.*, 2004)

Table 4.1 Percentage of meta-analyses fulfilling each item by type of meta-analyses

Guideline item	Meta-analyses of animal experiments % (n=46)	Hepatology meta-analyses (Christensen, 2001) % (n=15)	Pharmacotherapy meta-analyses (Hemels <i>et al.</i>, 2004) % (n=32)
Title – define as meta-analysis	59	93	22
Title – define as animal data ^a	52	<i>Not an item in QUOROM, so no result for these meta-analyses</i>	
Abstract – structured	29	40	50
Abstract – state objectives	98	73	69
Abstract – data sources	33	80	16
Abstract – review methods	2	20	9
Abstract – results	76	73	0
Abstract – conclusions	91	80	94
Introduction	100	100	91
Methods – information sources	74	87	59
Methods – special effort in searching ^a	17	<i>Not an item in QUOROM, so no result for these meta-analyses</i>	
Methods – inclusion/exclusion criteria	67	73	56
Methods – list excluded studies ^a	7	<i>Not an item in QUOROM, so no result for these meta-analyses</i>	
Methods – validity	26	67	16
Methods – abstraction	50	87	22
Methods – characteristics	57	87	72

Methods – synthesis	63	100	69
Results – trial flow	35	47	6
Results – characteristics	52	87	81
Results – synthesis	41	57	75
Discussion	85	93	97

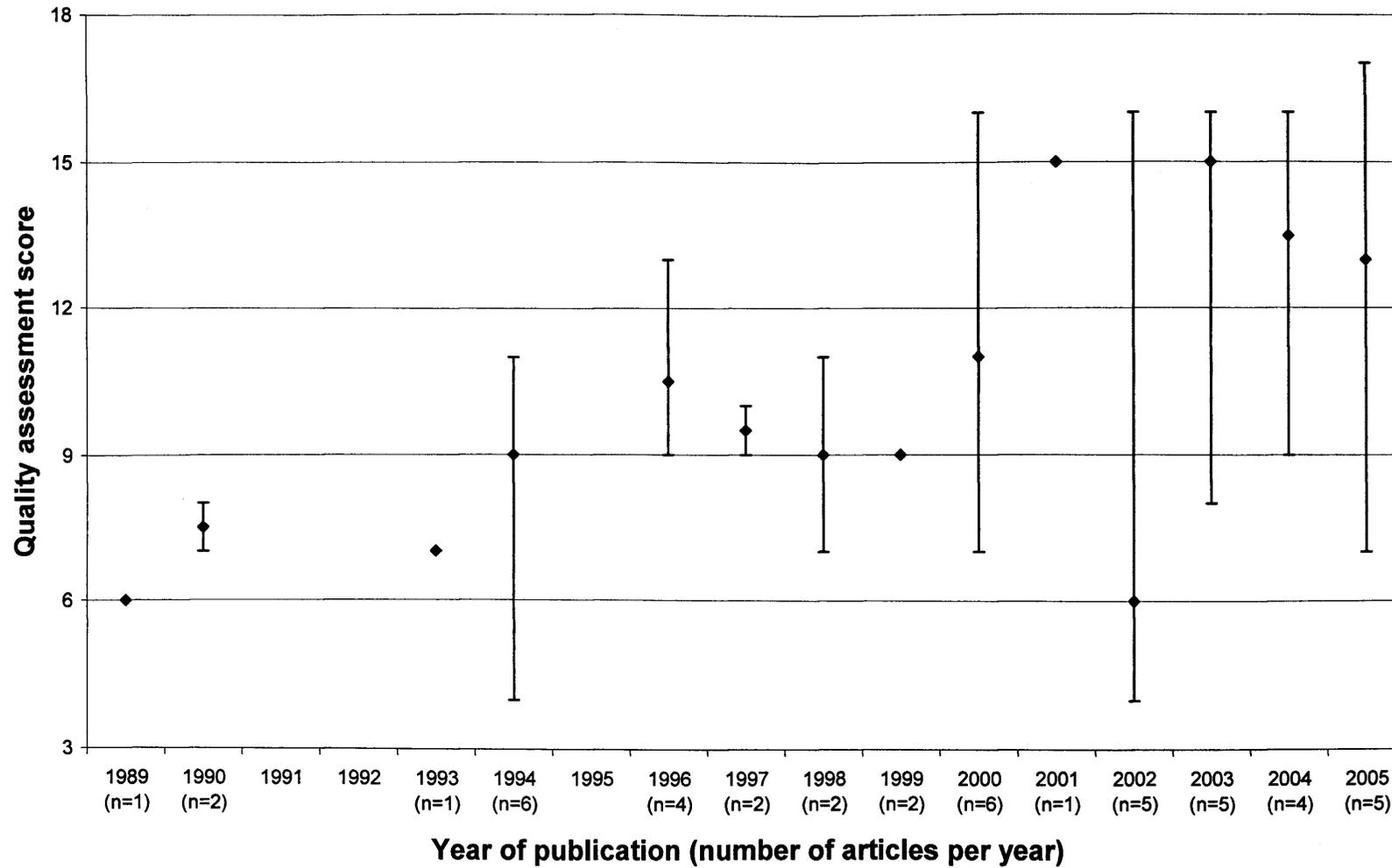
^a not an item in QUOROM

Finally, it is encouraging to note that there is some evidence to suggest the quality of reporting of meta-analyses of animal experiments is improving. Using the modified guidelines, a crude quality score was given to each meta-analysis, where a meta-analysis scored one point for every criteria they fulfilled in the guidelines (with a maximum possible score of 21). The median and range of quality scores for meta-analyses published in the same year are shown in Figure 4.9. This should help give some idea as to whether a change in the quality of these meta-analyses (according to the guidelines in Figure 4.7) has occurred over time. Figure 4.9 shows that the studies with the highest median quality scores were published in the last few years. However within the last few years a number of meta-analyses of low quality, according to the assessment carried out in this chapter, have also been published. This suggests that improvements in the quality of the reporting of meta-analyses of animal experiments are needed.

4.6 Summary

The systematic review described in this chapter has identified 103 articles reporting a systematic review and/or meta-analysis of animal experiments. This has allowed an investigation into the methods used, the appropriateness of these methods and has led to the development of guidelines to promote improved conduct and reporting of such research.

A number of deficiencies in the conduct and reporting of systematic reviews and meta-analyses of animal experiments have been identified in this review. Although the reporting of systematic reviews of animal experiments appears to be of reasonable quality, aspects of the methods and results sections of meta-analyses of animal experiments are particularly poorly reported. Moreover, results of this review suggest that features specific to meta-analyses of animal experiments, such as how to deal with different species and strains of animals, need to be addressed.

Figure 4.9 Median and range of quality assessment scores by year

n is the number of meta-analyses for that year

Several issues have emerged from this review, some of which are not specific to systematic reviews and meta-analyses of animal experiments. In relation to searches of the electronic database, the term ‘meta-analysis’ as a publication type (i.e. meta-analysis.pt) in MEDLINE identified a number of articles that were not quantitative syntheses of evidence but systematic reviews. This highlights an issue with terminology that has previously been noted (Chalmers *et al.*, 2002) where meta-analysis is thought of as a systematic review, rather than an extension to a systematic review in situations where it may be appropriate. For the systematic review described in this chapter, the issue regarding terminology was not particularly important since both systematic reviews and meta-analyses were being sought. However, it could be misleading for researchers and can greatly reduce the precision of carefully designed search strategies. Similarly, because the search term ‘animals’ was used, many ecological and veterinary studies were identified, decreasing the precision of the search process.

A large number of studies thought to be relevant for this review, were eventually excluded because they either referred to themselves as ‘systematic reviews’ but gave no details on any sources of evidence searched, the limitations or inclusion and exclusion criteria and the number of studies being reviewed (n=6) or gave one or two details of a systematic review which were not sufficient to meet the inclusion criteria given in Section 4.3.1 (n=5). For instance, Wegener (1979), Feihl *et al.* (2001), Kroeze *et al.* (2002) and Zhong *et al.* (2003) all refer to their articles as systematic reviews, but no details on the source(s) of evidence searched, how they were searched and the number of studies obtained are reported. Similarly, Loder (2003) describes her review as systematic in the abstract, yet only gives information on the sources searched, not on any inclusion/exclusion criteria or search strategy. Smorenburg and Van Noorden (2001) also refer to their review as systematic and state that “...all animal studies, published between 1960 and 1999... are reviewed”, but few further details are given.

A distinction should be made between articles that carry out a systematic review but do not publish sufficient detail on the process and articles that have not carried out a systematic review. Because of this it is not enough to just recommend the use of systematic reviews, the methods used must be clearly reported so that potential

sources of bias and details of the process can be examined and compared to the results and conclusions of the review.

This systematic review has shed light on the extent and style of systematic reviews and meta-analyses of animal experiments. It is interesting to note that none of the environmental meta-analyses were preceded by a systematic review, and that in many of these meta-analyses the origin of the data could not be identified.

However, it is encouraging to see that the number of meta-analyses without a systematic review is slightly smaller in the period 2001-2005, than in 1996-2000 (4 vs. 6 articles).

The quality assessment carried out in Section 4.5.2 has uncovered a number of areas of deficient reporting of the systematic reviews and meta-analyses of animal experiments. These must be addressed before systematic review and meta-analyses of animal experiments can make efficient use of the data to inform human health, while helping the aims of the UK 3Rs programme in reducing the number of animal experiments. In the following two chapters, systematic review and meta-analysis methods are used to identify, review and evaluate the available human and animal evidence relevant to risk assessments of two different environmental exposures in order to explore the potential benefits these methods may bring to the current risk assessment process. In Chapter 6, neurobehavioural risks associated with occupational exposure to Mn is assessed. First, in Chapter 5, a relationship between exposure to THMs in drinking water and an increase in the risk of a pregnant woman delivering a low birth weight baby is explored.

Trihalomethanes and low birth weight: example I

5.1 Chapter overview

The limitations of current risk assessment methods have been highlighted and the potential of systematic review and meta-analysis methods to overcome some of these limitations has been discussed (Chapter 2). However, the number of articles combining both human and animal data (the types of evidence expected in a risk assessment) appear to be small; the five articles identified in Section 4.4.3 (Kroll *et al.*, 1993; Carroll *et al.*, 1994; Guth *et al.*, 1997; Jiao *et al.*, 2000; Woodruff *et al.*, 2004) and the three articles described in Section 3.4.3 (DuMouchel and Harris, 1983; DuMouchel and Groër, 1989; Cox and Piegorsch, 1994). In this chapter, a cross-design synthesis approach (GAO, 1992) is taken to the evaluation of evidence for a risk assessment of exposure to a group of environmental chemicals. The possible association between exposure to trihalomethanes (THMs) in drinking water and an increased risk of giving birth to a low birth weight baby is investigated. Bayesian hierarchical meta-analysis models used by Prevost *et al.* (2000) and discussed in Chapter 3 are applied to evidence from human and animal studies to assess potential human health risks. In Section 5.2, the THM example is introduced and issues surrounding the risk assessment of this group of environmental chemicals are discussed. The search for relevant literature (Section 5.3), including that from both human and animal studies, the methods for obtaining study-specific estimates comparable across study design (Section 5.4) and the synthesis of these estimates within and across study type are all described (Section 5.5). Results of the evidence synthesis are presented in Section 5.6, with sensitivity analyses of assumptions given in Section 5.7. In Section 5.8 the results are interpreted and discussed in light of assumptions made and the findings from sensitivity analyses. Finally, in Section

5.9, implications for the use of systematic review and meta-analysis methods to assist in the human health risk assessment of trihalomethanes are discussed. The work presented in this chapter includes and extends the work described in Peters *et al.* (2005) (given in Appendix E).

5.2. Trihalomethane exposure and low birth weight

Trihalomethanes (THMs) are a group of chlorinated by-products (CBPs). They are formed when chlorine, added to drinking water supplies for disinfection purposes, reacts with organic and inorganic substances such as humic and fulvic acids already present in the water supply (Fawell *et al.*, 1997). The type and concentration of these CBPs depends upon many factors including the amount of chlorine added, time since the chlorine was added, water temperature and its pH level (Koivusaol and Vartianinen, 1997). THMs are the most common by-products of chlorination (Nieuwenhuijsen *et al.*, 2000). They include chloroform, bromodichloromethane (BDCM), dibromochloromethane (DBCM) and bromoform, but often *total* THMs are measured and reported in epidemiological studies. Correlations between the individual (chloroform, BDCM, DBCM and bromoform) and total THM concentrations suggest that total THMs are a good indicator for chloroform concentration in drinking water, but not for the other THMs (Whitaker *et al.*, 2003).

The evidence for possible health effects of exposure to THMs has been mixed. The health effects considered include cancers in general (Morris *et al.*, 1992; Hsu *et al.*, 2001; Lee *et al.*, 2004) and cancers specific to the pancreas (Do *et al.*, 2005), the bladder (McGeehin *et al.*, 1993; Villanueva *et al.*, 2003; Villanueva *et al.*, 2004) and the brain (Cantor *et al.*, 1999). Adverse reproductive and developmental effects have also been investigated (Reif *et al.*, 1996; Nieuwenhuijsen *et al.*, 2000). Because of concerns over adverse health effects from exposure to THMs, water companies are required to meet limits on the amount of THMs present in the water supply. In the US, the federal government has set a mandatory limit of 80 parts per billion (ppb) of total THMs in the water supply (EPA, 1998). Here in the UK, we are subject to EU directives of a limit of 100 ppb trihalomethanes in the water supply,

although it is stated that “where possible, without compromising disinfection, Member States should strive for a lower value” (EU, 1998).

Nieuwenhuijsen *et al.* (2000) published a narrative review of human and animal research investigating the potential association between THM exposure and adverse reproductive effects in humans. They suggested that there was a potential link between exposure to THMs and an increased risk of a pregnant woman delivering a low birth weight baby. The quality of the human studies, particularly for assessing exposure levels, was not high and the animal data were not comprehensive. In this chapter the narrative review by Nieuwenhuijsen *et al.* (2000) is built upon by systematically collating and quantitatively summarising the human and animal research investigating oral exposure to THMs and a possible risk of delivering a low birth weight baby. WHO has defined low birth weight as babies weighing less than 2500g at birth (WHO, 2004) and this is the definition used in the analyses presented here.

5.3 A systematic search of the literature

Since the focus of this chapter is on methods to combine different types of evidence, only limited details of the systematic review to identify literature on the potential association between oral exposure to THMs and low birth weight are reported here. Following guidelines set out by the CRD (Deeks *et al.*, 2001), evidence from both human studies and animal experiments was sought. Because of problems with translation, the search was limited to those reports published in the English language. The implications of this are discussed in Section 5.9. The electronic databases and search terms used are given in Appendix F. The literature search was supplemented by investigation of the references from published reports and reviews (e.g. Reif *et al.*, 1996; Fawell *et al.*, 1997) to identify further relevant research.

Data from eight articles were found to be relevant: five human epidemiological studies (Bove *et al.*, 1992; Kramer *et al.*, 1992; Savitz *et al.*, 1995; Gallagher *et al.*, 1998; Dodds *et al.*, 1999) and three animal studies (Thompson *et al.*, 1974; Ruddick *et al.*, 1983; Narotsky *et al.*, 1997). Multiple experiments were reported in each of

the animal studies, so that eight animal experiments were identified in total. The 13 relevant studies (five epidemiological studies and eight animal experiments) are summarised in Table 5.1.

There are a number of differences between the human epidemiology studies and the animal experiments in this example. Firstly, in the human studies, odds ratios (ORs) for a pregnant woman to deliver a low birth weight baby are presented at each exposure level category where the referent group is the lowest exposure group. In the animal experiments, the health outcome of interest, foetal weight, is subject to a litter effect since foetal weight is related to the number of pups in a litter (Healy, 1972), i.e. the more pups there are in a litter, the more likely those pups will be of low weight, as with multiple births in humans. An ideal analysis of foetal weight would take into account the litter effect (Williams, 1975), however, in the animal experiments used in this example, the individual animal data are not available and so the results, as reported in the animal experiments, are used. In two of the eight animal experiments the mean and standard deviation of the mean weights in each litter at each dose level are reported and used in subsequent analyses. The remaining six experiments report mean weights and standard deviations at each dose level, but do not report whether litter effect has been accounted for. Secondly, the ORs from the human studies are each adjusted for a number of different covariates (e.g. education, maternal age, parity), in the animal experiments different species and strains are used. Thirdly, exposure to THMs is measured and reported differently. Exposure in the human studies is measured in terms of ppb, while in the animal studies it is measured in mg per kg of body weight per day (mg/kg/day). In four of the human studies total THMs are measured while in all of the animal experiments and one human study (Kramer *et al.*, 1992), exposure to one or more of the individual THMs (e.g. chloroform, BDCM, DBCM, bromoform) is assessed.

The many differences in study design, measurement and reporting have to be addressed directly so that comparable study-specific estimates can be obtained for each of the 13 studies, and then synthesised for a meaningful summary of the relevant evidence to assist in the derivation of an exposure limit.

Table 5.1 Summary of the relevant studies identified from the literature search

Author	Study design	Population/ species/strain	Exposure	Exposure measure	No. dose levels	Outcome measure
<i>Five human epidemiological studies</i>						
Bove <i>et al.</i> , 1992	Cross-sectional	Population based	Total THMs	ppb	5	Term low birth weight (>37 weeks gestation and < 2500g)
Savitz <i>et al.</i> , 1995	Case-control	Population based	Total THMs	ppb	3	All low birth weight (< 2500g)
Dodds <i>et al.</i> , 1999	Retrospective cohort	Population based	Total THMs	ppb*	4	All low birth weight (< 2500g)
Gallagher <i>et al.</i> , 1998	Retrospective cohort	Population based	Total THMs	ppb	4	All low birth weight (< 2500g)
Kramer <i>et al.</i> , 1992	Case-control	Population based	Chloroform	ppb*	3	All low birth weight (< 2500g)
<i>Eight animal experiments</i>						
Thompson <i>et al.</i> , 1974	In vivo toxicology	Sprague-Dawley rats	Chloroform	mg/kg/day	4	Foetal weights at postpartum day 1
		Dutch-Belted rabbits	Chloroform	mg/kg/day	4	Foetal weights at postpartum day 1
Ruddick <i>et al.</i> , 1983	In vivo toxicology	Sprague-Dawley rats	Chloroform	mg/kg/day	3	Foetal weights at postpartum day 1
			BDCM	mg/kg/day	3	Foetal weights at postpartum day 1
			CDBM	mg/kg/day	3	Foetal weights at postpartum day 1
			Bromoform	mg/kg/day	3	Foetal weights at postpartum day 1
Narotsky <i>et al.</i> , 1997	In vivo toxicology	Fisher 344 rats	BDCM (in corn oil)	mg/kg/day	4	Foetal weights at postpartum day 1
			BDCM (in aqueous)	mg/kg/day	4	Foetal weights at postpartum day 1

* reported in µg/l, which is equivalent to ppb

5.4 Obtaining study-specific estimates

In order to obtain comparable estimates of effect, a dose-response slope estimate of the $\ln(\text{OR})$ for low birth weight was sought from each individual study. This meant that the measures of effect reported in the animal experiments (mean foetal weight at each dose) had to be converted to $\ln(\text{OR})$ s for low birth weight. To transform the mean foetal weights at each dose level in the animal experiments to ORs it is assumed that the foetal weights were normally distributed and that 7.6 % of the animals in the control (zero dose) group were of low weight. There appear to be no generally accepted cut off values for low foetal weight in different animal species and strains, and so the percentage of low birth weight babies, those weighing < 2.5 kg in England in 1999 (ONS, 2000), was used: 7.6%. Under these assumptions the number of animals in each dose group considered to be of normal and low foetal weight was obtained so that the OR for low birth weight could be calculated. The measured exposure scales (i.e. levels of THM reported in ppb or mg/kg/day) also had to be transformed so that they were comparable between the human studies and animal experiments. Since it is more difficult to obtain estimates of water intake and body weight for different animals, the exposure scale from the human studies was transformed to that in the animal experiments (mg/kg/day). Under the assumptions that average body weight is 60 kg and average water intake is 2 litres/day, as in standard risk assessment practice (DoE, 1993; WHO, 1996), exposures measured in the human epidemiology studies were calculated in terms of mg/kg/day.

Thus, 80ppb = 80 $\mu\text{g/l}$.

Applying the average water intake assumption

= 160 $\mu\text{g/day}$,

applying the average body weight assumption

$$\begin{aligned}
 &= \frac{160\mu\text{g/day}}{60\text{kg}} \\
 &= 2.6667\mu\text{g/kg/day} \\
 &= 0.0027\text{mg/kg/day}
 \end{aligned}$$

To calculate a dose-response slope in each study a weighted least squares linear regression of $\ln(\text{OR})$ on the natural log of dose, $\ln(\text{dose})$, was fitted. The weight was inversely proportional to the variance of the $\ln(\text{OR})$ estimate at each $\ln(\text{dose})$ level

$$\ln(\text{OR}_{ik}) = \alpha_i + \beta_i \ln(\text{dose}_{ik}) + \varepsilon_i \quad (5.1)$$

where k is the observation number in study i ($i:1, \dots, 13$), and ε_i is the error term.

The α_i 's represent the intercept which is forced through the minimum dose in each study, and the slope estimates β_i (and corresponding variances) are the study-specific estimates to be used in the subsequent synthesis.

Although there are differences in the exposures measured between the studies (i.e. individual and total THMs) and evidence suggests some are more highly correlated than others with total THMs (Whitaker *et al.*, 2003), for simplicity these exposures are assumed equivalent. Similarly, differences in the designs of the human epidemiology studies (case-control, cohort and cross-sectional) are not addressed here, i.e. the assumption is that each type of study is estimating the same effect and is subject to the same biases. This will be discussed further in Section 5.9.

5.5 Methods of synthesis

In Chapter 3, methods for the synthesis of human and animal evidence and methods for the generalised synthesis of human evidence were reviewed. A number of features common to the most appealing models were discussed and subsequently proposed for their use in the synthesis of human and animal data in this and the following chapter. Thus, because of flexibility in the specification of models, a

hierarchical Bayesian approach is taken here to combine the human and animal evidence for an assessment of exposure to THMs and an increased risk of delivering a low birth weight baby. A number of models are used to synthesise the evidence, ranging in sophistication from the separate synthesis of the human evidence and the animal evidence, using ‘vague’ and informative prior distributions, to more complex models allowing for differences in the species and strains used in each study (as in Prevost *et al.* (2000) and Sutton and Abrams (2001)). Pooled estimates from a classical random effects meta-analysis are calculated and compared with the estimates from all of the Bayesian models. The aim of applying these models to the data is to estimate the effect of exposure to THMs on the risk of a woman delivering a low birth weight baby, taking account of the relevant human and animal evidence.

The Bayesian analyses in this chapter and in Chapter 6 were carried out in WinBUGS (Spiegelhalter *et al.*, 2000b). This software package was designed specifically for Bayesian analyses. The statistical model is specified and Markov chain Monte Carlo (MCMC) simulation is used for estimation of the unknown parameters. MCMC simulations allow integration over high-dimensional probability distributions to provide parameter estimates for models. Samples are drawn from the probability distributions and approximated (Gilks *et al.*, 1999). The parameter estimates can be made more accurate by increasing the number of samples drawn. In WinBUGS, a Gibbs sampler is used to draw the samples from the full set of posterior conditional distributions which converge under Ergodic theory to the posterior marginal distributions (Spiegelhalter *et al.*, 2000a; Sutton and Abrams, 2001). A common issue in MCMC estimation is establishing convergence of the sampler. Assessing whether convergence has been achieved is not easy and is subjective. Within WinBUGS a number of diagnostic tools are available to help assess convergence. The Gelman-Rubin plot allows assessment of convergence of the sampler when more than one chain (all starting from different initial values) is used. The history plot displays successive sample values (i.e. plots of sample value against sample number) to illustrate how well the sampler is mixing (i.e. checking that it is moving around the sample space quickly), further indications of good or poor mixing can be obtained from the autocorrelation plot which shows the correlation between successive sample values and the density plot shows the posterior samples. Examples of these plots are given in Figures 5.4 – 5.7. In

addition to these plots, the changing of initial values, length of burn-in and the number of iterations can be used to help judge the convergence of models (see Section 5.7.3).

5.5.1 Synthesis of study-specific estimates within study type (human epidemiology studies and animal experiments)

A two-level hierarchical random effects Bayesian meta-analysis model (*Model 1*) was used to combine the study-specific dose-response slope estimates, β_i , within each of the two study types (human epidemiological studies and animal experiments). This random effects model is the same as that presented in Equation 3.5 and those used by Prevost *et al.* (2000), Sutton and Abrams (2001) and Simmonds *et al.* (2003). Within this framework, the data were analysed in several ways: *Model 1a* combined only the human epidemiological studies; *Model 1b* combined only the animal experiments. Both of these models had ‘vague’ prior distributions placed upon the unknown parameters, so that the data would dominate in the synthesis of the studies. In Section 5.7 sensitivity analyses for different choices of ‘vague’ prior distributions are described.

Two further models, *Model 1c* and *Model 1d*, are used to combine the epidemiological studies and animal experiments, respectively, but with informative prior distributions on the pooled slope parameter, μ , which is based on the posterior mean and variance from *Model 1a* and *Model 1b*. This follows the approach taken by Sutton and Abrams (2001) in the synthesis of evidence from RCTs and observational studies. Thus, the posterior mean and variance from the synthesis of the animal experiments (*Model 1b*), is used to inform the prior distribution for the synthesis of the human epidemiological studies (*Model 1c*). Similarly, the posterior mean and variance from synthesising the epidemiological studies (*Model 1a*), is used as a basis for the prior distribution in the synthesis of the animal experiments (*Model 1d*). The structure used by all forms of *Model 1* follows that given in Equation 3.5

$$\begin{aligned} \beta_i &\sim N(\theta_i, \sigma_i^2) & \mu &\sim N(a, b) \\ \theta_i &\sim N(\mu, \tau^2) & 1/\tau^2 &\sim \text{Gamma}(0.001, 0.001) \end{aligned} \quad (5.2)$$

where θ_i is the true dose-response slope in study i (for *Model 1a* and *Model 1c* $i: 1, \dots, 5$, for *Models 1b* and *Model 1d* $i: 1, \dots, 8$), μ is the pooled dose-response slope and τ^2 is the between study variance. In all four models (*Models 1a - d*), it is assumed that σ_i^2 are known, and the observed variances of the slope estimates in each study are used to represent them. In *Model 1a* and *Model 1b*, $a = 0$ and $b = 10^9$, allowing for a range of plausible values for the dose-response slope estimate of between -2×10^9 and 2×10^9 . In *Model 1c*, $a =$ posterior mean from *Model 1b* and $b =$ posterior variance from *Model 1b*. Similarly, in *Model 1d*, $a =$ posterior mean from *Model 1a* and $b =$ posterior variance from *Model 1a*. In all four models, a ‘vague’ prior, $\text{Gamma}(0.001, 0.001)$, is placed upon the between study precision parameter, $1/\tau^2$.

5.5.2 Synthesis of study-specific estimates across study type

A fifth form of *Model 1*, *Model 1e*, combines all 13 studies ($i: 1, \dots, 13$), taking no account of the fact studies are from two different sources, with ‘vague’ prior distributions on all unknown parameters. This is not an appropriate approach as the different sources of evidence should be accounted for in the synthesis because of their differing designs and relevance to the parameters of interest, i.e. human health effects (GAO, 1992). Three further models (*Models 2, 3a* and *3b*) do take into account that these data are from different study types.

A further hierarchical random effects Bayesian model (*Model 2*) estimates an overall pooled slope estimate, μ , but allows distinct estimates of the between study variances for the epidemiological studies and the animal experiments. Thus, a common average is assumed regardless of study design but different between-study variances are allowed for each study type. *Model 2* is given by

$$\begin{aligned}
 \beta_{ij} &\sim N(\psi_{ij}, \sigma_{ij}^2) & \mu &\sim N(0, 10^9) \\
 \psi_{ij} &\sim N(\mu, \tau_j^2) & 1/\tau_1^2 &\sim \text{Gamma}(0.001, 0.001) \\
 & & 1/\tau_2^2 &\sim \text{Gamma}(0.001, 0.001)
 \end{aligned} \tag{5.3}$$

where the β_{ij} are the observed slope estimates for the i th study of type j (where $j=1$ for epidemiological studies and $j=2$ for animal experiments), σ_{ij}^2 are the variances of the β_{ij} and ψ_{ij} is the true slope estimate for study i of type j . τ_j^2 is the variance between studies of type j and μ is the pooled slope estimate. ‘Vague’ prior distributions were placed on the unknown parameters μ , $1/\tau_1^2$ and $1/\tau_2^2$.

Model 3 can be thought of as an extension to *Model 2* where distinct study type estimates of effect are modelled. It is a three-level hierarchical random effects Bayesian model which includes a level to account for study type, j . Prevost *et al.* (2000) and Sutton and Abrams (2001) apply this model to combine data from RCTs with data from observational studies. This model was reviewed in Chapter 3 (Equation 3.10). In *Model 3a*, $j:1,2$ where $j=1$ represents the human epidemiological studies and $j=2$ the animal experiments). In *Model 3b*, the animal experiments are further divided in an attempt to account for the different species and strains of animal used in the studies, hence $j:1, \dots, 4$ where $j=1$ for human studies, $j=2$ for experiments using rabbits, $j=3$ for experiments using Sprague-Dawley rats and $j=4$ for experiments using F344 rats). In Equation 5.4, *Model 3* (Equation 3.10) is described alongside the prior distributions placed on the unknown parameters μ , $1/\tau_j^2$ and $1/\nu^2$.

$$\begin{array}{ll}
 \beta_{ij} \sim N(\psi_{ij}, \sigma_{ij}^2) & \mu \sim N(0, 10^9) \\
 \psi_{ij} \sim N(\theta_j, \tau_j^2) & 1/\tau_j^2 \sim \text{Gamma}(0.001, 0.001) \\
 \theta_j \sim N(\mu, \nu^2) & 1/\nu^2 \sim \text{Gamma}(0.001, 0.001)
 \end{array} \quad (5.4)$$

As well as estimating the variance between studies of type j , τ_j^2 , *Model 3* estimates the variance between study types, ν^2 , and the pooled dose-response slope for the j th type of study, θ_j . As with *Model 1e* and *Model 2*, the pooled slope estimate of the 13 studies is given by μ . ‘Vague’ prior distributions were placed upon the unknown parameters μ , $1/\tau_j^2$ and $1/\nu^2$. It may arguably be more appropriate to discuss the estimate of θ_1 , the overall human effect, from this model than μ , the overall species

effect. In Figure 5.2 both estimates are presented and this point is further discussed in Section 5.8.

WinBUGS (version 1.3) (Spiegelhalter *et al.*, 2000b) was used to estimate parameters for the Bayesian analyses using Markov Chain Monte Carlo simulation. For each model a ‘burn-in’ of 10,000 iterations was followed by a further 200,000 updates after which the median and the 2.5 and 97.5 percentiles of the posterior distribution were used to summarise the parameter estimates. All initial values were set to one and convergence and model performance were assessed visually from the trace and autocorrelation plots available within WinBUGS (see Figures 5.4-5.6).

As part of the sensitivity analyses reported in Section 5.7, the length of ‘burn-in’, number of updates and the initial values given above were changed, and different ‘vague’ prior distributions were placed upon the parameters in *Models 1e, 3a* and *3b* to assess the sensitivity of the results to these prior distribution specifications.

5.6. Results

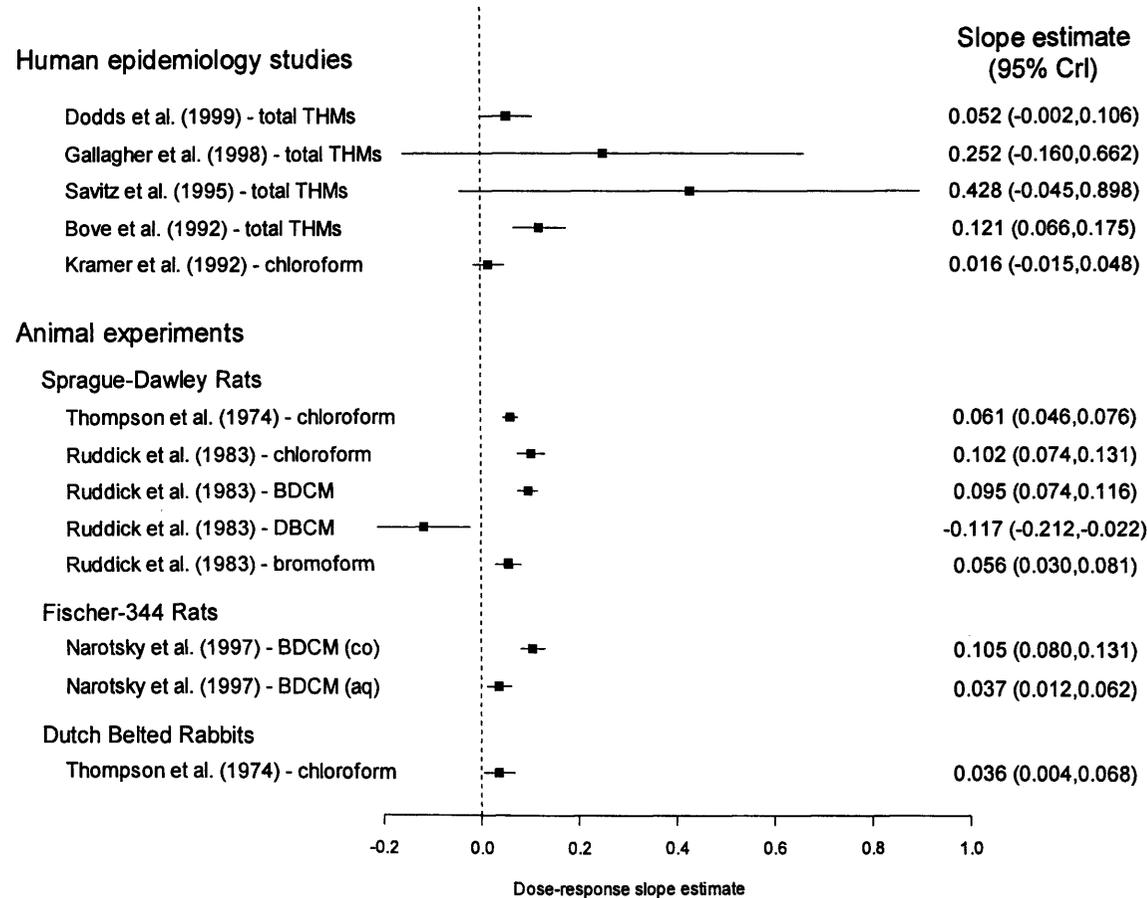
5.6.1 Study-specific dose-response slope estimates

The dose-response slope estimates (medians), β_i , and 95% credibility intervals (CrIs) calculated from each study are shown in Figure 5.1. Apart from the study of Ruddick *et al.* (1983) on DBCM exposure, the animal experiments all have very similar dose-response slope estimates and are quite precise, the human epidemiology study estimates appear more heterogeneous and generally less precise. The estimate of the between study variance (the posterior median), τ^2 , for the epidemiology studies is 0.0051, but for the animal experiments it is 0.0027.

5.6.2 Within study type pooled dose-response slope estimates

In Table 5.2 the pooled dose-response slope estimates, μ , and 95% CrIs obtained from *Models 1a-d* are compared with the results from a classical random effects synthesis fitted using the META command in Stata 8.2 (StataCorp, 2004).

Figure 5.1 Study specific dose-response slope estimates (and 95% CrIs) from $\ln(OR)$ vs. $\ln(dose)$ linear model



co, corn oil vehicle; aq aqueous vehicle

Table 5.2 Pooled slope estimates, μ , (and 95% CrIs/CIs) from the human epidemiological studies and animal experiments

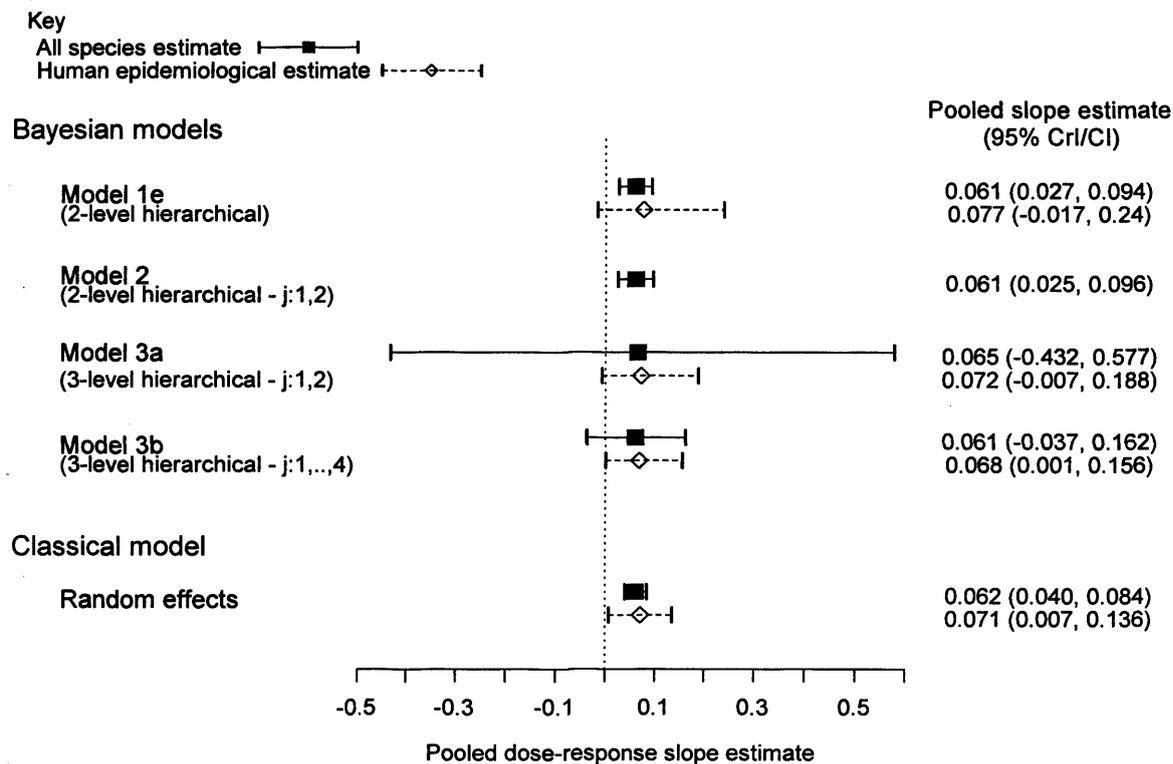
Model	Epidemiological studies (n=5)	Animal experiments (n=8)
Bayesian model with vague prior distribution (<i>Models 1a</i> and <i>1b</i>)	0.077 (-0.017, 0.240)	0.058 (0.006, 0.100)
Bayesian model with informative prior distribution (<i>Models 1c</i> and <i>1d</i>)	0.061 (0.023, 0.099)	0.060 (0.017, 0.100)
Classical random effects model	0.071 (0.007, 0.136)	0.062 (0.038, 0.086)

Results from the Bayesian models with ‘vague’ prior distributions (*Model 1a* and *Model 1b*) and the classical models show that the pooled slope estimates from the human data, μ , are slightly larger than the pooled slope estimates from the animal experiments and have much more variability associated with them. This reflects what was seen in Figure 5.1. As expected, the Bayesian models give pooled slope estimates with wider 95% CrIs than the estimates obtained from the classical random effects model. This is because more variability is taken into account in the Bayesian model. The results from the Bayesian models with informative prior distributions (*Model 1c* and *Model 1d*) are very similar to each other and are close to the pooled slope estimate of the animal experiments from *Model 1b* and the classical random effects model. Thus, it would appear that in *Model 1c* and *Model 1d* the data from the animal experiments are dominating, regardless of whether they form the prior or the likelihood in the Bayesian analysis. This reflects the fact that the slope estimates from the animal experiments are estimated more precisely than those from the human epidemiology studies.

5.6.3 Overall pooled dose-response slope estimates

The pooled slope estimates, μ , from *Models 1e*, *2*, *3a* and *3b*, and the pooled slope estimate for the human evidence, θ_i , in *Models 3a* and *3b* are shown in Figure 5.2. Again, pooled estimates from the classical random effects model are given for comparison.

Figure 5.2 Pooled dose-response slope estimates, μ and θ_i (95% CrIs) obtained from the five synthesis models used to combine all 13 studies



Model 1e, human epidemiology estimate is μ from Model 1a;
 Model 3, human epidemiology estimate is θ_i

It can be seen that the models give similar pooled dose-response slope estimates, μ , but the level of precision varies across models. *Model 1e* assumes no difference between the human studies and animal experiments. Results suggest that in comparison with the findings of *Models 1a-d* (Table 5.2), the animal data are dominating the estimate of the pooled slope response slope; this is not necessarily what is wanted. Although animal experiments are less likely to be prone to biases than the human epidemiology studies and potentially contribute useful information, in terms of relevance, evidence from human studies is more important. In *Model 2* although a common mean is again assumed, the between-study variance for human studies is allowed to differ to that from animal experiments. This reflects what is seen in Section 5.6.1 where dose-response slopes from the animal experiments are more similar than those from the human studies. Allowing for these different between-study variances in *Model 2* results in a slightly wider 95% CrI for μ than

that from *Model 1e*; the pooled estimate is the same. *Models 3a* and *3b* yield larger amounts of uncertainty in the pooled dose-response slope estimates, reflecting the variability allowed among species in the models. In particular, *Model 3a* (which distinguishes between the human studies and the animal experiments) gives a very wide 95% CrI compared to the pooled slope estimates from the other models in Figure 5.2. The reason for this, as Prevost *et al.* (2000) also noted, is that the variance among species is estimated with enormous uncertainty, partly since only two sources of evidence, in this example, the pooled slope estimate from the human epidemiology studies, θ_1 , and the pooled slope estimate from the animal experiments, θ_2 , are being synthesised to obtain the overall pooled slope estimate, μ . Hence, when the different species and strains are taken into account in *Model 3b*, so that there are four sources of evidence, the 95% CrI is much narrower. Before any interpretation of the parameter estimates from each of these models is discussed (Section 5.8), the sensitivity of some of the various assumptions made in this analysis are investigated.

5.7. Sensitivity analyses

Many assumptions were made to obtain comparable dose-response slope estimates from each of the 13 studies and to combine them using the synthesis models described. In this section some of these assumptions are changed to assess the sensitivity of the results. The impact of different dose-response models on the resulting slope estimates, β_i , is explored, in addition to changing the water intake and body weight values assumed in the initial analyses. The effect of different ‘vague’ prior distributions placed on the unknown parameters in the synthesis models defined in Section 5.5 is also investigated.

5.7.1 Dose-response models

For simplicity a linear regression slope was used to obtain study-specific dose-response slope estimates in Section 5.4. The sensitivity of the results based on this model is now explored using a number of alternative dose-response models. Where i is the study ($i:1, \dots, 13$) and j is the dose level in each study i , the alternative dose-response models are:

A. A logit model on $\ln(\text{dose})$ given by

$$\ln\left(\frac{p_{ik}}{1-p_{ik}}\right) = \alpha_i + \beta_i \ln(\text{dose}_{ik}) + \varepsilon_i \quad (5.5)$$

where p_{ik} is the probability of observing a low birth weight foetus at dose level k in study i .

B. A linear model as given in Equation 5.1, but with dose replacing $\ln(\text{dose})$ as the explanatory variable

$$\ln(OR_{ik}) = \alpha_i + \beta_i \text{dose}_{ik} + \varepsilon_i \quad (5.6)$$

C. A linear model of $\ln(\text{OR})$ on dose as in B (Equation 5.6) above, but taking into account the correlation structure between the $\ln(\text{OR})$ s within each study using the method from Greenland and Longnecker (1992). The covariance matrix C is obtained by letting A_0 and A_x be the fitted number of low birth weight foetuses at the reference exposure and exposure x , respectively, B_0 and B_x be the fitted number of foetuses that are not of low birth weight at the reference exposure and at exposure level x , respectively, such that $\frac{A_x B_0}{A_0 B_x} = \exp(L_x)$, where L_x is the adjusted $\ln(\text{OR})$ at exposure level x ($x \neq 0$). C then has diagonal elements v_x (the estimated variance of L_x) and off-diagonal elements c_{xz} , where

$$c_{xz} = r_{xz} (v_x v_z)^{1/2}, \quad r_{xz} = \left(\frac{1}{A_0} + \frac{1}{B_0} \right) s_x s_z \quad \text{and} \quad s_x = \frac{1}{A_x} + \frac{1}{B_x} + \frac{1}{A_0} + \frac{1}{B_0} \quad (5.7)$$

(Greenland and Longnecker, 1992).

D. A logit model as in A above (Equation 5.5), but with dose replacing $\ln(\text{dose})$ as the explanatory variable

$$\ln\left(\frac{p_{ik}}{1-p_{ik}}\right) = \alpha_i + \beta_i \text{dose}_{ik} + \varepsilon_i \quad (5.8)$$

The alternative dose-response models A, B and D above were all calculated in WinBUGs, dose-response model C was calculated classically in Splus (Splus, 1999).

The study-specific dose-response slope estimates are shown in Table 5.4. In the models where ‘dose’ is the independent variable (Models B, C and D), the study-specific slope estimates from Model B are generally larger than those from Models C and D. One difference between Model B and Models C and D that may explain the increase in the slope estimates is that Model B incorrectly assumes independence of estimates within a study. This is incorrect since all ln(ORs) refer to the lowest exposure category as the reference group, therefore they are not independent. Although estimates are not independent in Model C, the correlation structure of the ln(OR) is taken into account using the method of Greenland and Longnecker (1992), and in Model D, because ln(OR) are not being modelled, the estimates within a study are independent.

A clear difference between the slope estimates calculated from the different models is that for the epidemiology studies the estimates are much larger when ‘dose’ is used as the independent variable (Models B, C and D), rather than ‘ln(dose)’ (Equation 5.1 and 5.5 (Model A)). However such a difference is not seen in the slope estimates from the animal experiments. The slope estimates are defined as the increase in ln(OR) per unit increase in exposure (mg/kg/day), whether that be dose or ln(dose). Whereas the range of dose levels covered by the animal experiments reaches 400 mg/kg/day, the exposure reported in the epidemiology studies do not exceed 0.0042 mg/kg/day.

If a ‘dose’ dose-response model is used, the epidemiology studies will each have very little weight in the synthesis of all studies (as they have very low precision), and so the animal experiments will tend to dominate. However, if the ‘ln(dose)’ dose-response model is used, the animal experiments are less likely to dominate because the precision of the slope estimates is similar for the animal experiments

and epidemiology studies. The results of applying the synthesis *Models 1a, 1b and 1e* to the dose-response slope estimates obtained from the different linear dose-response models are given in Table 5.3.

The choice of dose-response model is critical for the synthesis of the studies. To investigate which dose-response model is most appropriate, the fit of each of the dose-response models was assessed using the Bayes Information Criterion (BIC) (Schwarz, 1978). BIC values for the dose-response model are given in Table 5.4. The shaded numbers in each row highlight the model with the lowest BIC value. The lowest BIC value indicates the dose-response model that best fits the data from that study (Schwarz, 1978). The linear models appear to give a better fit to the data than the logit models, since the BIC values are lower for the linear models. However, results from applying the BIC suggest that the use of ‘ln(dose)’ over ‘dose’ in the linear model improves the fit of the model.

5.7.2 Sensitivity of assumptions on body weight, water consumption and low birth weight cut off values

To convert the measurement of exposure reported in the epidemiology studies, standard default body weight and water consumption values were applied to the data. More recent surveys suggest that these values are inaccurate (ECETOC, 2001; EA and DEFRA, 2002). The dose-response slopes, β_i , from the epidemiology studies were re-analysed using estimates from these surveys in place of the former ‘default’ values. As described on page 88,

$$dose_{mg/kg/day} = \frac{dose_{ppb} * water}{weight * 1000} \quad (5.9)$$

where $dose_{mg/kg/day}$ is the dose level in mg/kg/day, $dose_{ppb}$ is the dose level in ppb, $water$ is the estimate of daily water intake per day and $weight$ is the average weight in kg. In the initial analysis, $water = 2$ and $weight = 60$ in Equation 5.9. By using a Bayesian model inclusion of uncertainty about the assumed values, their impact on the dose-response slope estimates obtained and the subsequent synthesis of the slopes can also be assessed. Thus, for the sensitivity analyses, a mean of 68.53 kg (standard deviation of 13.87 kg) was assumed as average body weight (EA and

DEFRA, 2002) such that $weight \sim N(68.53, 192.38)$ in Equation 5.9, and 0.991 litres/day (standard deviation of 0.0304) was assumed for average water intake (ECETOC, 2001), such that $water \sim N(0.991, 0.0009)$ in Equation 5.9. These sensitivity analyses were carried out for both the 'ln(dose)' linear dose-response model and the 'dose' linear dose-response model. The results of these sensitivity analyses are given in Table 5.5.

Table 5.3 Study-specific dose-response slope estimates (and 95% CrIs) obtained from the three different linear dose-response models (original dose-response model and Models A-D in Equations 5.5 –5.8)

	<i>Original Model</i>	<i>Model A</i>	<i>Model B</i>	<i>Model C</i>	<i>Model D</i>
Study	Linear model: ln(OR) and ln(dose)	Logit model: ln(dose)	Linear model: ln(OR) and dose	Linear model: ln(OR) and ln(dose)	Logit model: Dose
Epidemiological studies					
Dodds <i>et al.</i> (1999) (Total THMs)	0.052 (-0.002, 0.106)	0.045 (-0.028, 0.118)	27.37 (-3.38, 57.99)	16.92 (-20.75, 54.60)	17.78 (-20.83, 56.25)
Gallagher <i>et al.</i> (1998) (Total THMs)	0.252 (-0.160, 0.662)	0.130 (-0.335, 0.584)	253.2 (-117, 622)	210.24 (-27.53, 448.01)	183.3 (-261.20, 597.10)
Savitz <i>et al.</i> (1995) (Total THMs)	0.428 (-0.045, 0.898)	0.210 (-0.283, 0.705)	150 (-30.17, 329)	62.26 (-122.77, 247.30)	55.73 (-121.30, 231.10)
Bove <i>et al.</i> (1992) (Total THMs)	0.121 (0.066, 0.175)	0.124 (0.056, 0.193)	98.83 (53.47, 143.8)	95.40 (39.78, 151.01)	95.99 (41.14, 150.80)
Kramer <i>et al.</i> (1992) (Chloroform)	0.016 (-0.015, 0.048)	0.016 (-0.019, 0.052)	638.2 (-487.8, 1763)	650.71 (-606.00, 1907.4)	665.5 (-610.60, 1917.00)
Animal experiments					
<i>Sprague-Dawley Rats</i>					
Thompson <i>et al.</i> (1974) (Chloroform)	0.061 (0.046, 0.076)	0.070 (0.047, 0.096)	0.017 (0.013, 0.021)	0.013 (0.010, 0.016)	0.132 (0.010, 0.016)

Ruddick <i>et al.</i> (1983) (Chloroform)	0.102 (0.074, 0.131)	0.160 (0.132, 0.191)	0.022 (0.017, 0.028)	0.023 (0.017, 0.029)	0.031 (0.026, 0.037)
Ruddick <i>et al.</i> (1983) (BDCM)	0.095 (0.074, 0.116)	0.097 (0.065, 0.135)	0.015 (0.011, 0.019)	0.004 (0.001, 0.007)	0.006 (0.003, -0.008)
Ruddick <i>et al.</i> (1983) (DBCM)	-0.117 (-0.212, -0.022)	-0.147 (-0.273, -0.063)	-0.038 (-0.070, -0.007)	0.038 (0.007, 0.069)	-0.054 (-0.110, -0.020)
Ruddick <i>et al.</i> (1983) (Bromoform)	0.056 (0.030, 0.081)	0.030 (0.003, 0.061)	0.008 (0.005, 0.011)	0.009 (0.006, 0.012)	0.016 (0.013, 0.021)
<i>Fischer-344 Rats</i>					
Narotsky <i>et al.</i> (1997) (BDCM – corn oil vehicle)	0.105 (0.080, 0.131)	0.108 (0.077, 0.144)	0.042 (0.029, 0.056)	-0.002 (-0.017, 0.012)	0.010 (-0.001, 0.021)
Narotsky <i>et al.</i> (1997) (BDCM – aqueous vehicle)	0.037 (0.012, 0.062)	0.041 (0.009, 0.079)	0.008 (-0.003, 0.019)	0.001 (-0.018, 0.020)	-0.007 (-0.017, 0.003)
<i>Dutch Belted Rabbits</i>					
Thompson <i>et al.</i> (1974) (Chloroform)	0.036 (0.004, 0.068)	0.044 (-0.002, 0.103)	0.021 (0.002, 0.040)	0.014 (-0.008, 0.065)	0.017 (-0.004, 0.040)

Table 5.4 Bayesian Information Criterion (BIC) values for the different dose-response models

	<i>Original Model</i>	<i>Model A</i>	<i>Model B</i>	<i>Model D</i>
Study	Linear model: ln(OR) & ln(dose)	Logit model: ln(OR) & ln(dose)	Linear model: ln(OR) & dose	Logit model: ln(OR) & dose
Epidemiological studies				
Dodds <i>et al.</i> (1999) (Total THMs)	2.04[†]	4.26	2.58	4.89
Gallagher <i>et al.</i> (1998) (Total THMs)	3.34	6.46	3.14	6.05
Savitz <i>et al.</i> (1995) (Total THMs)	3.80	4.84	2.79	5.15
Bove <i>et al.</i> (1992) (Total THMs)	5.10	9.90	5.32	11.03
Kramer <i>et al.</i> (1992) (Chloroform)	0.90	2.52	0.70	2.20
Animal experiments				
<i>Sprague-Dawley Rats</i>				
Thompson <i>et al.</i> (1974) (Chloroform)	9.12	35.54	4.50	7.22
Ruddick <i>et al.</i> (1983) (Chloroform)	27.77	208.14	9.19	23.93
Ruddick <i>et al.</i> (1983) (BDCM)	10.77	59.86	24.86	91.99
Ruddick <i>et al.</i> (1983) (DBCM)	1.21	4.54	1.32	2.96
Ruddick <i>et al.</i> (1983) (Bromoform)	21.06	135.63	14.96	56.54
<i>Fischer-344 Rats</i>				
Narotsky <i>et al.</i> (1997) (BDCM–corn oil vehicle)	3.83	67.99	39.73	125.16
Narotsky <i>et al.</i> (1997) (BDCM–aqueous vehicle)	11.62	33.92	17.79	38.51
<i>Dutch Belted Rabbits</i>				
Thompson <i>et al.</i> (1974) (Chloroform)	1.21	2.96	1.54	3.66

Table 5.5 Dose-response slope estimates and 95% CrIs from the epidemiological studies obtained from applying different body weight and water intake assumptions

Study	<i>Original Model</i> Linear: ln(OR) vs. ln(dose)			<i>Model B</i> Linear: ln(OR) vs. dose	
	Original assumptions	Including variability of assumptions	Excluding variability of assumptions	Including variability of assumptions	Excluding variability of assumptions
Dodds <i>et al.</i> (1999)	0.052 (-0.002, 0.106)	0.052 (-0.002, 0.107)	0.052 (-0.002, 0.107)	72.66 (-9.97, 168.60)	62.99 (-8.12, 133.8)
Gallagher <i>et al.</i> (1998)	0.252 (-0.160, 0.662)	0.253 (-0.155, 0.662)	0.253 (-0.155, 0.662)	674.70 (-312.50, 1798)	585.40 (-260.3, 1433)
Savitz <i>et al.</i> (1995)	0.428 (-0.045, 0.898)	0.428 (-0.043, 0.897)	0.428 (-0.043, 0.897)	400.40 (-81.91, 958.50)	345.9 (-68.39, 757.9)
Bove <i>et al.</i> (1992)	0.121 (0.066, 0.175)	0.121 (0.066, 0.175)	0.121 (0.066, 0.175)	267.30 (134.50, 438.10)	229.2 (125.2, 333.3)
Kramer <i>et al.</i> (1992)	0.016 (-0.015, 0.048)	0.018 (-0.017, 0.052)	0.018 (-0.017, 0.052)	1696 (-1332, 5032)	1472 (-1124, 4055)
Pooled estimate (Model 1a)	0.077 (-0.017, 0.24)	0.078 (-0.016, 0.239)	0.078 (-0.016, 0.239)	154.20 (19.48, 513.80)	148.60 (20.28, 466.20)

For the 'ln(dose)' model, if the variability of the assumptions from the survey data is taken into account the estimates and 95% CrIs do not change (compare results in column two and three in Table 5.5). However, taking into account the variability of the body weight and water intake assumptions in the 'dose' model, larger dose-response slope estimates are obtained and the variability associated with these estimates is much greater than when the variability is not taken into account. The model that includes information on the variability of the water intake and body weight assumptions is more appealing as it shows from the outset the amount of variability associated with the results. Clearly, as more accurate data are collected, the models can be updated.

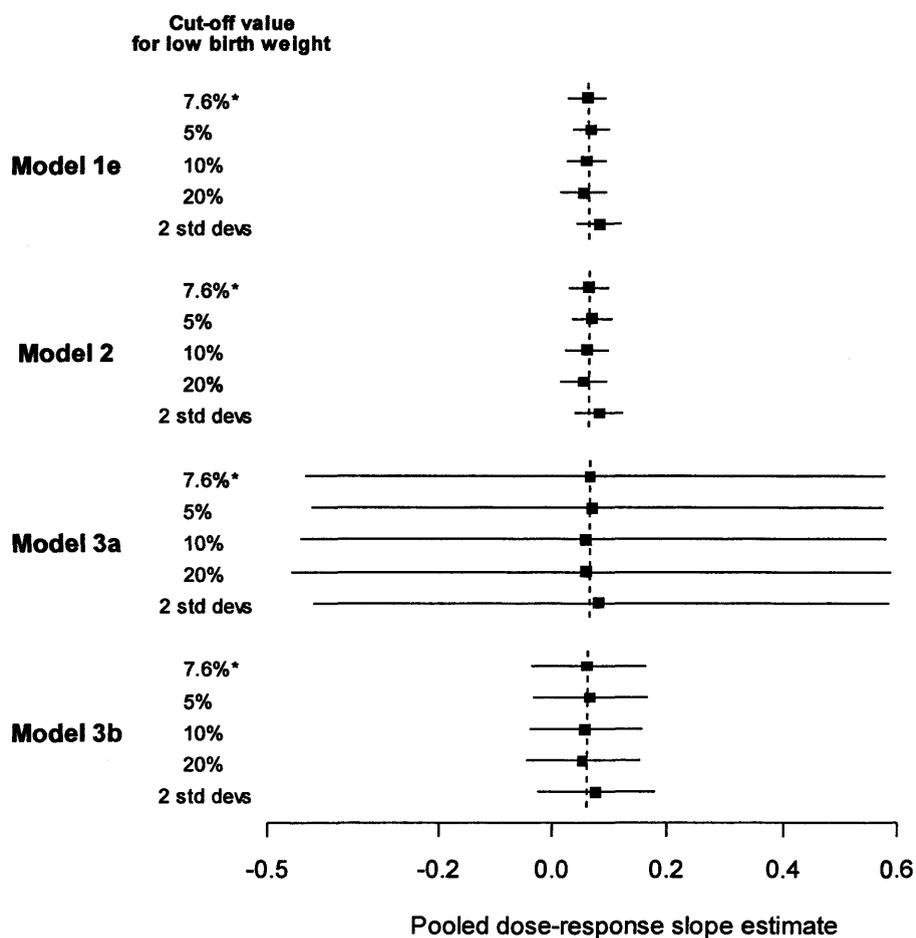
Further assumptions were made in defining the percentage of control group animals that were of low birth weight in the animal experiments so that ORs could be calculated from the reported means (and standard deviations) of the foetal weights. Four further cut-off levels of low birth weight were applied: 5 %, 10 %, 20 % and taking two standard deviations below the control group mean (Foster and Auton, 1995). The results are given in Figure 5.3.

When the cut-off values for low birth weight are changed there is little impact on the pooled slope estimates. Use of Foster and Auton's definition of two standard deviations below the mean as low birth weight gives the largest departure from the estimate calculated in the initial analyses, however, this is a relatively small difference.

5.7.3 Checking of model fitting assumptions and sensitivity of prior distributions

Use of different 'burn-in' lengths, number of iterations and initial values in the Bayesian analyses in WinBUGS suggested that the models had converged satisfactorily. Convergence was assessed using a number of diagnostic tools available in WinBUGS: Gelman-Rubin plots for the convergence of multiple chains, the autocorrelation plot, the density plot and the history plot. Examples of these are given in Figures 5.4-5.7 for Model 3b (Equation 5.4).

Figure 5.3 Pooled dose-response slope estimates obtained for different low birth weight cut-off values



* Cut-off value used in the initial analyses. The dashed line indicates the pooled slope estimate from initial analysis for that Model.

Figure 5.4 Gelman-Rubin plots for μ , θ_i and τ^2 in Model 3b (see Equation 5.4)

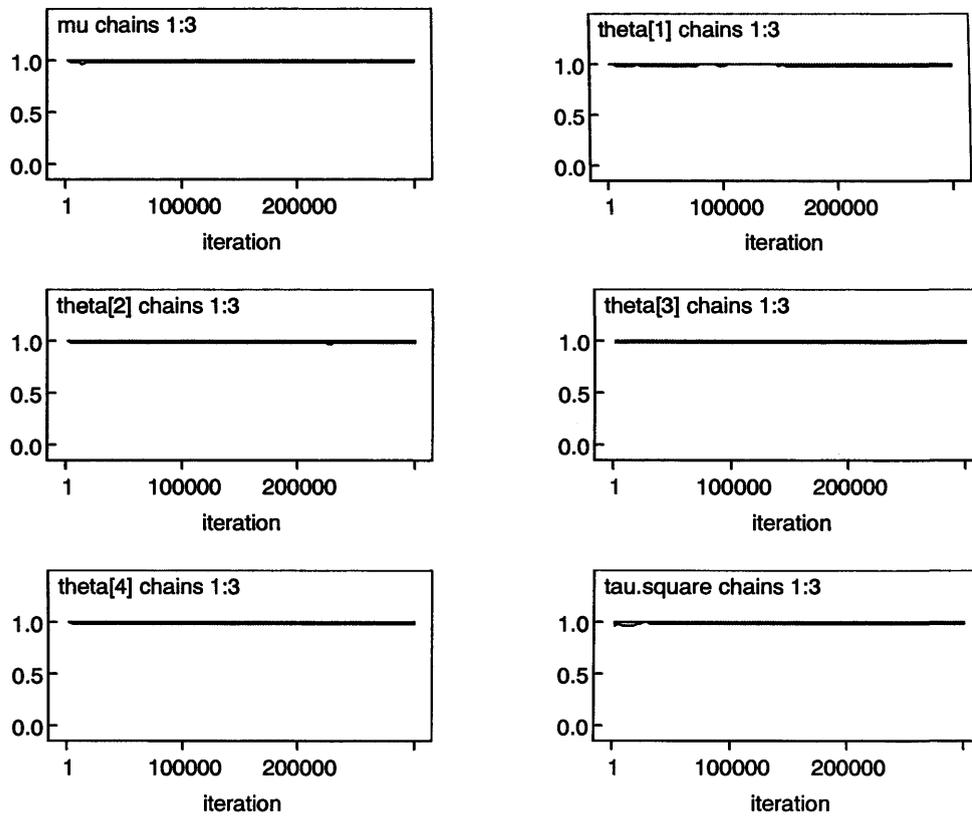


Figure 5.5 Autocorrelation plots for μ , θ_i and τ^2 in Model 3b (see Equation 5.4)

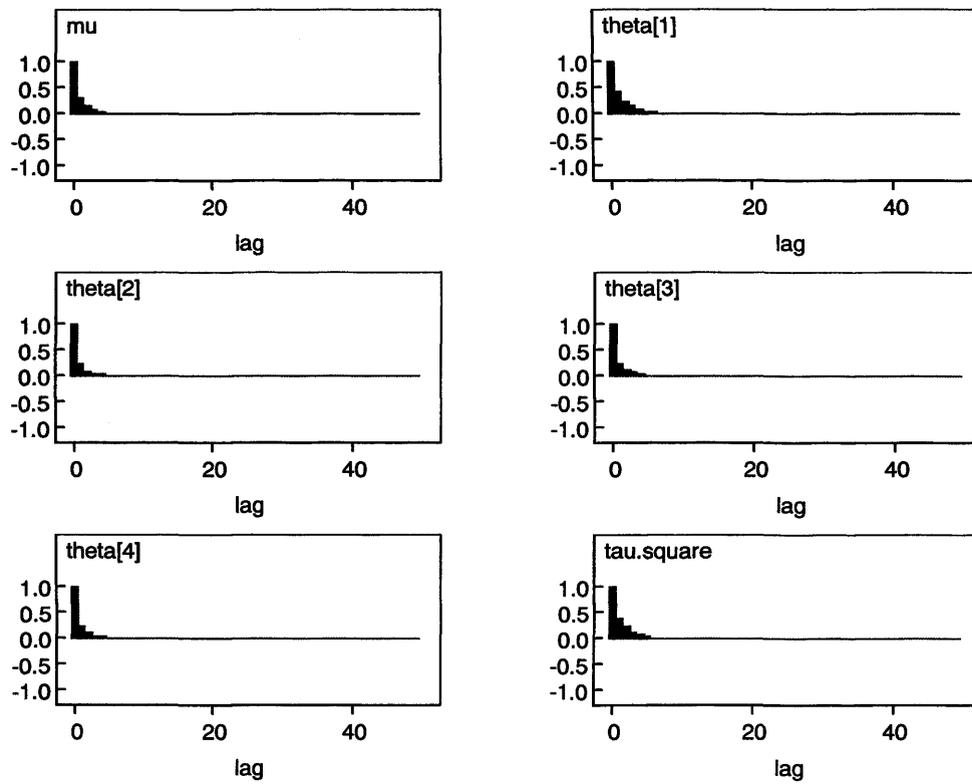


Figure 5.6 History plots for μ , θ_i and τ^2 in Model 3b (see Equation 5.4)

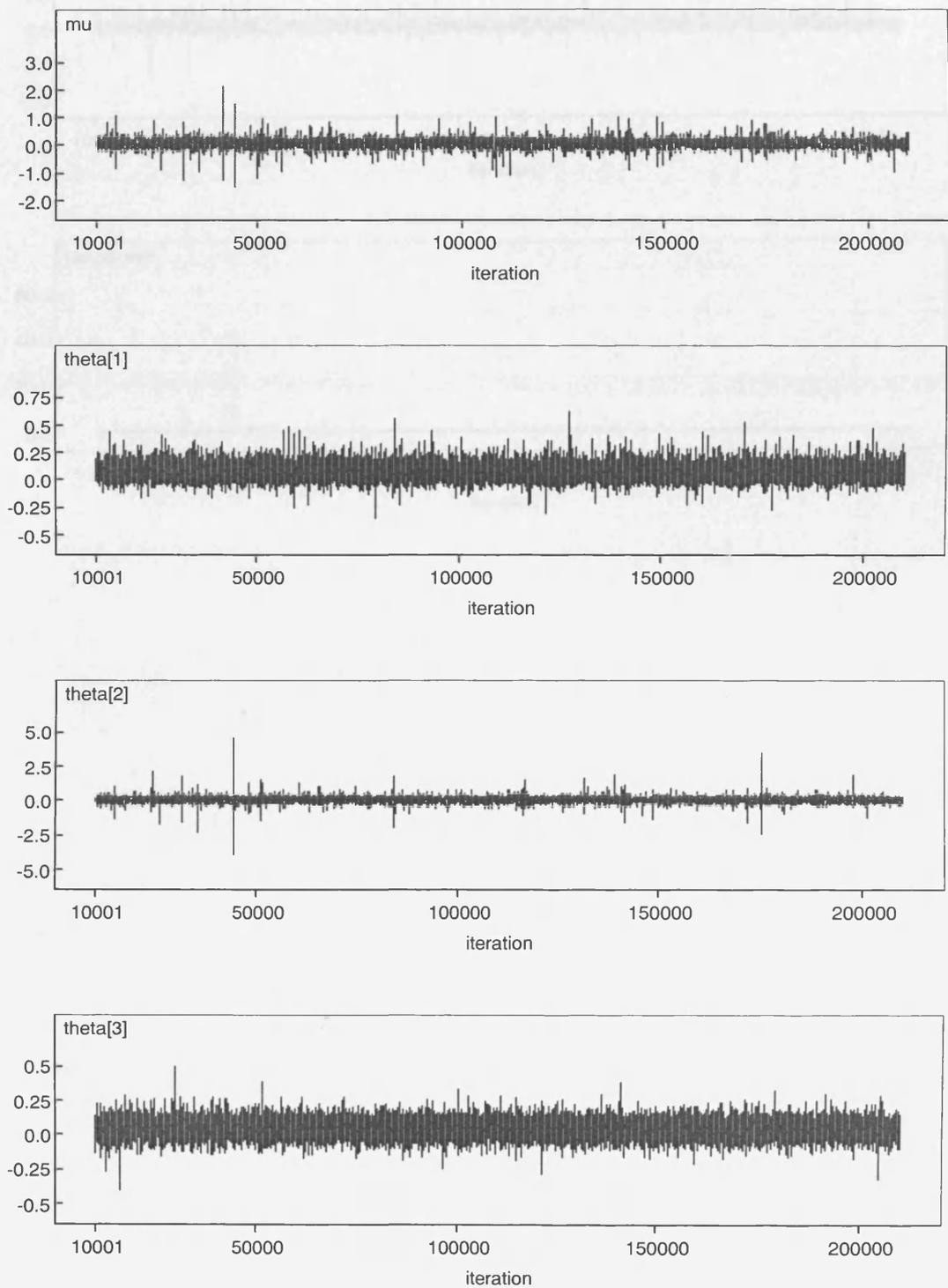
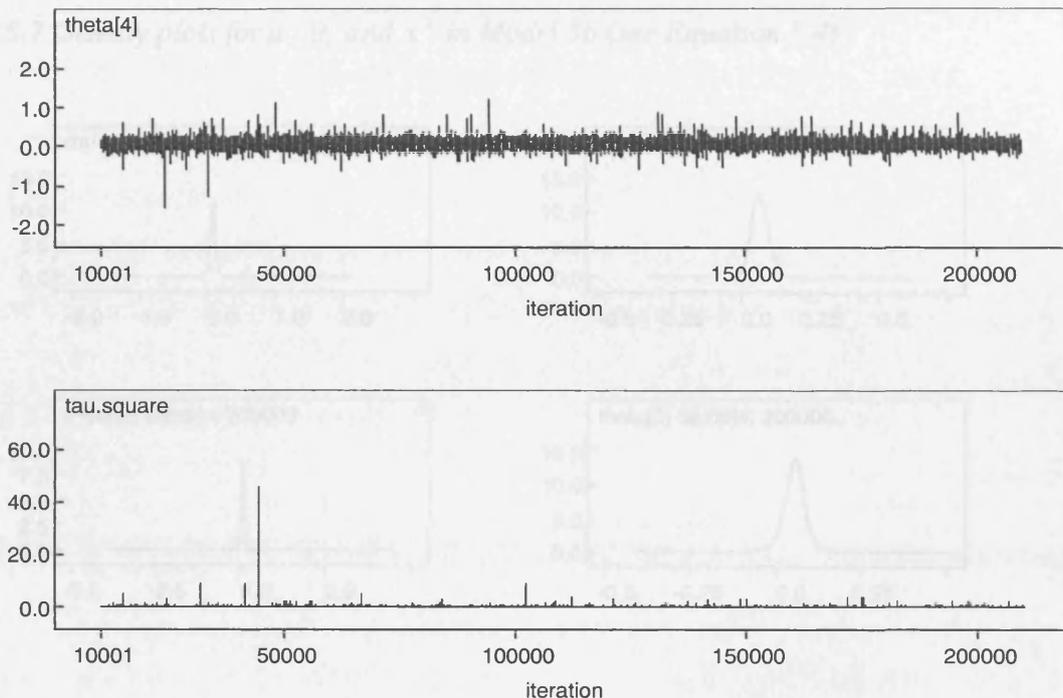


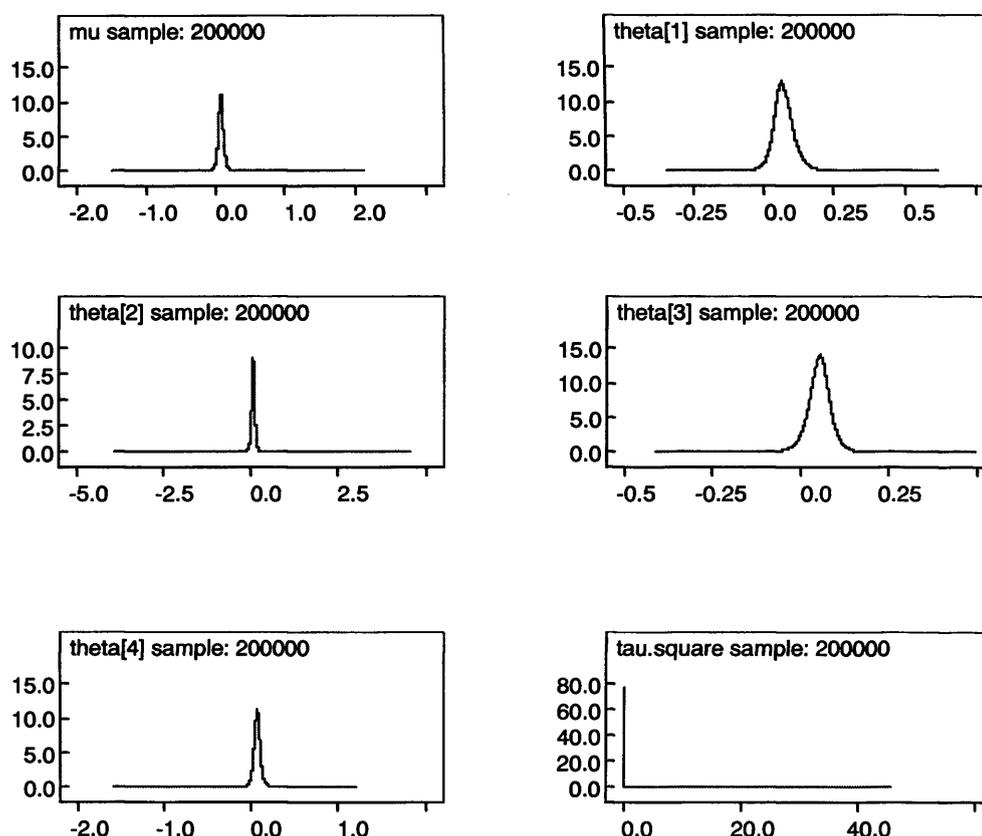
Figure 5.7: Trace plots for μ , σ , and α^2 in Model 5b using Equation 5.4



Assessment of the Gelman-Rubin plots (Figure 5.8) shows that the MCMC chains within each chain and the average of the chains have converged. This suggests that the MCMC algorithm has converged to the target distribution. The Gelman-Rubin diagnostic is a measure of convergence that is achieved if the ratio of the variance of the chains to the variance of the average is approximately 1.0. In this case, the ratio is approximately 1.0, indicating convergence. The trace plots (Figure 5.6) show that the chains have converged to the target distribution. The trace plots show that the chains have converged to the target distribution. The trace plots show that the chains have converged to the target distribution.

As shown in the trace plots, the chains have converged to the target distribution. The trace plots show that the chains have converged to the target distribution. The trace plots show that the chains have converged to the target distribution. The trace plots show that the chains have converged to the target distribution. The trace plots show that the chains have converged to the target distribution.

Figure 5.7 Density plots for μ , θ_i and τ^2 in Model 3b (see Equation 5.4)



Assessment of the Gelman-Rubin plots (Figure 5.4) indicates that the 80% interval within each chain and the average 80% interval across the chains have a ratio of one, thus suggesting that convergence between the three different chains has been achieved (Brooks and Gelman, 1998). Figures 5.5 and 5.6 indicate that the Gibbs sample is mixing well, since there is little autocorrelation for all parameters (Figure 5.5) and no evidence of the sample getting ‘stuck’ in a particular sample space (Figure 5.6).

As recommended by many authors including DuMouchel and Harris (1983), DuMouchel and Groer (1989), Prevost *et al.* (2000), Sutton *et al.* (2000), Sutton and Abrams (2001), Spiegelhalter *et al.* (2004) and Lambert *et al.* (2005) the sensitivity of vague prior distributions placed on the unknown parameters were assessed. In *Model 1e*, a normal prior distribution, $N(0, 10^{11})$, which is slightly more diffuse than

that used in the initial analyses, is placed on the pooled dose-response slope estimate, μ . Two further prior distributions were placed upon the unknown between study precision component. They were: $\sigma \sim N(0,10^4)$, for $\sigma > 0$, which is uniform on the variance, and $1/\sigma^2 \sim Uniform(0,10^4)$ which allows values very close to zero for the variance component. The results obtained from using different prior distributions in *Model 1e* to synthesise all 13 studies are shown in Table 5.6.

Table 5.6 Results (slope estimates and 95% CrIs) from using different ‘vague’ priors in Model 1e

Prior for between-study precision, $1/\sigma^2$	Prior for pooled slope estimate, μ	
	$N(0,10^9)$	$N(0,10^{11})$
Gamma (0.001, 0.001)	0.061 (0.027, 0.094)	0.061 (0.027, 0.094)
$N(0, 10^4)^\dagger$, $\sigma > 0$	0.061 (0.026, 0.095)	0.061 (0.026, 0.095)
Uniform (0,10 ⁴)	0.061 (0.030, 0.091)	0.061 (0.030, 0.091)

[†] Prior distribution placed on between-study standard deviation

The findings suggest that changing the prior distributions in *Model 1e* has very little effect on the estimate for the pooled dose-response slopes. Hence we can be confident that the *Model 1e* is robust to the ‘vague’ priors used in the initial synthesis.

The prior distributions placed upon the unknown parameters μ , $1/\sigma_j^2$ and $1/\nu^2$ in *Models 3a* and *3b* are shown in Table 5.7.

Table 5.7 Prior distributions used in the sensitivity analyses for Models 3a and 3b

Priors for between study-type precision, $1/\nu^2$	Priors for between study precision, $1/\sigma_j^2$	Priors for pooled slope estimate, μ
Gamma (0.001, 0.001)	Gamma (0.001, 0.001)	N(0, 10^9)
Gamma (0.1, 0.1)	Gamma (0.1, 0.1)	N(0, 10^{11})
Gamma (1, 1)	Gamma (1, 1)	
N(0, 10^6) [†] for $1/\nu^2 > 0$	N(0, 10^6) ^{††} for $\sigma_j > 0$	
N(0, 10^4) [†] for $1/\nu^2 > 0$	N(0, 10^4) ^{††} for $\sigma_j > 0$	
N(0, 10^2) [†] for $1/\nu^2 > 0$	N(0, 10^2) ^{††} for $\sigma_j > 0$	

[†] Prior distribution placed upon between study-type standard deviation, ν

^{††} Prior distribution placed upon between-study standard deviation, σ_j

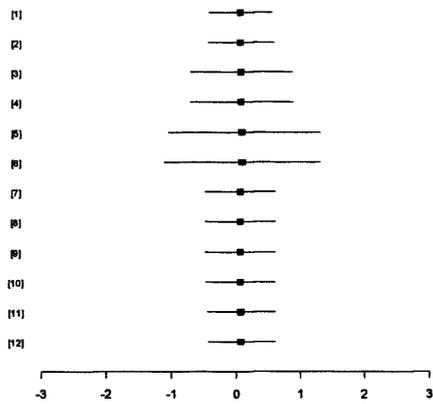
The Gamma (0.1, 0.1) and Gamma (1, 1) prior distributions placed on the precision parameters will allow for larger values for the variance component than those from using a Gamma (0.001, 0.001) prior distribution as in the original analysis. The half-normal prior distributions placed upon the standard deviation are uniform on the variance, and allow for differing ranges of plausible values for the variance component, with N(0, 10^6) being the most diffuse of the half-normal variance priors.

The results of these sensitivity analyses are shown in Figures 5.8 and 5.9 for *Models 3a* and *3b*, respectively. Each plot A-F gives the pooled estimate and 95% CrI from the different vague prior distributions placed on the pooled slope μ and the between-study precision $1/\sigma_j^2$. For instance, in plot A the 12 different combinations for vague prior distributions on μ and $1/\sigma_j^2$ are compared when the between study-type precision $1/\nu^2 \sim \text{Gamma}(0.001, 0.001)$; in plot B when $1/\nu^2 \sim \text{Gamma}(0.1, 0.1)$. Since there is a great deal of variability in the variance estimates for the different between study-type vague prior distributions in *Model 3a*, plots D-F in Figure 5.8 are on different scales. There is less variability in the estimates in Figure 5.9, so each plot has the same scale to help comparison between the different specifications of the between study-type precision in *Model 3b*. The numbers 1-12 in each plot correspond to the following vague prior specifications,

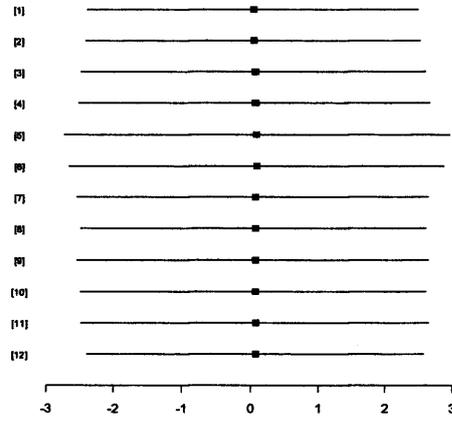
[1]	$1/\sigma_j^2 \sim \text{Gamma}(0.001, 0.001);$	$\mu \sim \text{N}(0, 10^9)$
[2]	$1/\sigma_j^2 \sim \text{Gamma}(0.001, 0.001);$	$\mu \sim \text{N}(0, 10^{11})$
[3]	$1/\sigma_j^2 \sim \text{Gamma}(0.1, 0.1);$	$\mu \sim \text{N}(0, 10^9)$
[4]	$1/\sigma_j^2 \sim \text{Gamma}(0.1, 0.1);$	$\mu \sim \text{N}(0, 10^{11})$
[5]	$1/\sigma_j^2 \sim \text{Gamma}(1, 1);$	$\mu \sim \text{N}(0, 10^9)$
[6]	$1/\sigma_j^2 \sim \text{Gamma}(1, 1);$	$\mu \sim \text{N}(0, 10^{11})$
[7]	$\sigma_j \sim \text{N}(0, 10^6), \sigma > 0;$	$\mu \sim \text{N}(0, 10^9)$
[8]	$\sigma_j \sim \text{N}(0, 10^6), \sigma > 0;$	$\mu \sim \text{N}(0, 10^{11})$
[9]	$\sigma_j \sim \text{N}(0, 10^4), \sigma > 0;$	$\mu \sim \text{N}(0, 10^9)$
[10]	$\sigma_j \sim \text{N}(0, 10^4), \sigma > 0;$	$\mu \sim \text{N}(0, 10^{11})$
[11]	$\sigma_j \sim \text{N}(0, 10^2), \sigma > 0;$	$\mu \sim \text{N}(0, 10^9)$
[12]	$\sigma_j \sim \text{N}(0, 10^2), \sigma > 0;$	$\mu \sim \text{N}(0, 10^{11})$

Figure 5.8 Pooled dose-response slope estimates and 95% CrIs from sensitivity analyses of the prior distributions in Model 3a

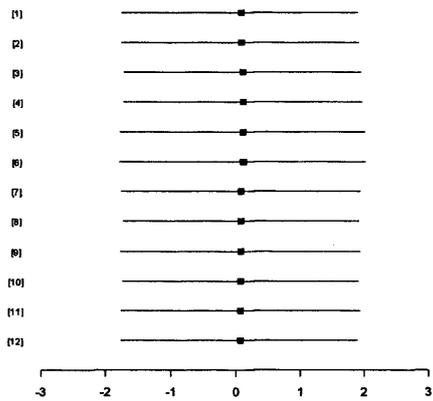
A. $1/\nu^2 \sim \text{Gamma}(0.001, 0.001)$



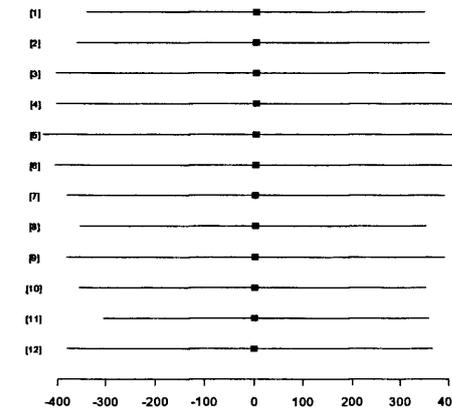
B. $1/\nu^2 \sim \text{Gamma}(0.1, 0.1)$



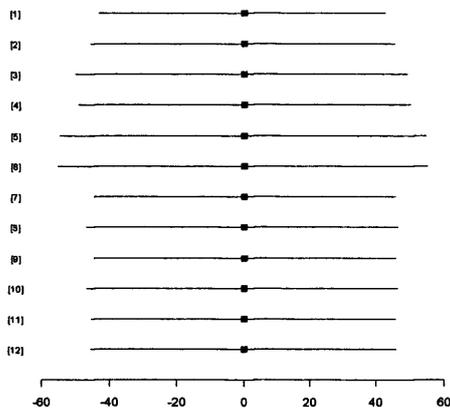
C. $1/\nu^2 \sim \text{Gamma}(1, 1)$



D. $\nu^2 \sim N(0, 10^6), \nu > 0$



E. $\nu^2 \sim N(0, 10^4), \nu > 0$



F. $\nu^2 \sim N(0, 10^2), \nu > 0$

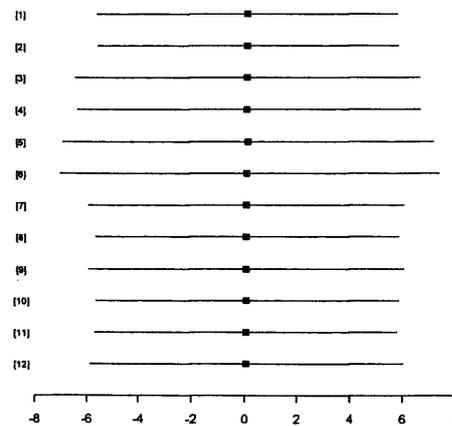
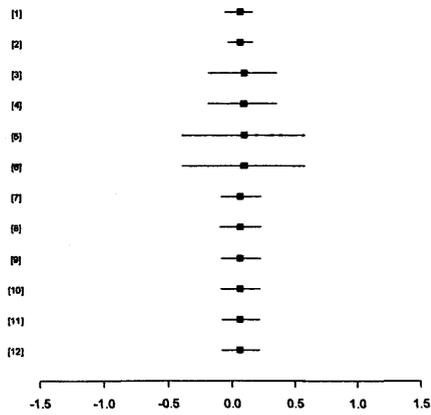
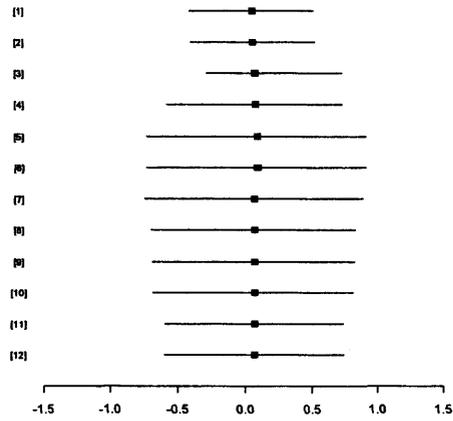


Figure 5.9 Pooled dose-response slope estimates and 95% CrIs from sensitivity analyses of the prior distributions in Model 3b

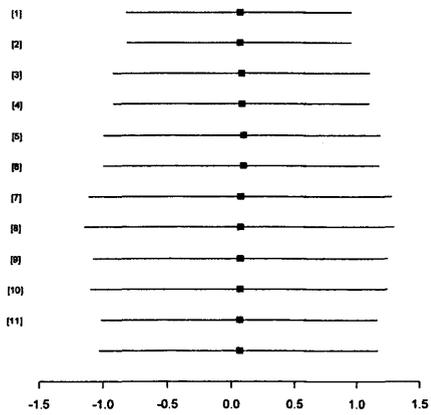
A. $1/\nu^2 \sim \text{Gamma}(0.001, 0.001)$



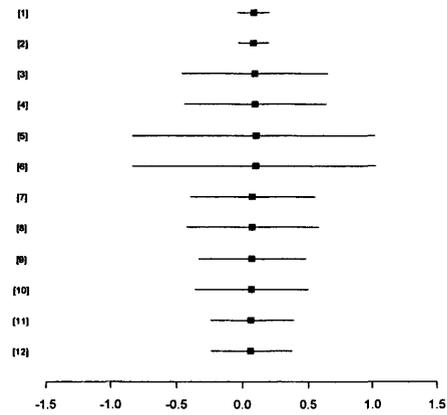
B. $1/\nu^2 \sim \text{Gamma}(0.1, 0.1)$



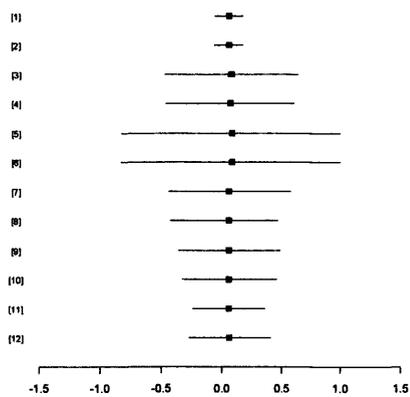
C. $1/\nu^2 \sim \text{Gamma}(1, 1)$



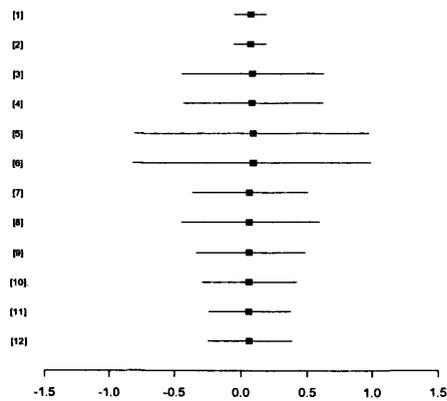
D. $\nu^2 \sim N(0, 10^6), \nu > 0$



E. $\nu^2 \sim N(0, 10^4), \nu > 0$



F. $\nu^2 \sim N(0, 10^2), \nu > 0$



Priors for the pooled dose-response slope parameter, μ

Changing the prior distribution placed on μ has very little effect on the pooled estimates and 95% CrIs in both *Models 3a* and *3b* (Figures 5.8 and 5.9). This suggests that the prior distribution used in the initial analyses, $N(0, 10^9)$, is sufficiently diffuse to let the data dominate in the synthesis.

Priors for the between study precision parameter, $1/\sigma_j^2$

Changing the between study precision prior in *Models 3a* and *3b* does have an impact on the estimated pooled dose-response slopes and 95% CrIs. In *Model 3a*, the impact is generally small (see Figure 5.8). The half-normal prior distributions on $1/\sigma_j^2$ ([7] – [12]) give very similar estimates and 95% CrIs to each other, regardless of the prior placed on the between study-type precision parameter, $1/\nu^2$. But the gamma prior distributions ([1] – [6]) give slightly different estimates and 95% CrIs. Larger slope estimates and 95% CrIs are obtained when the Gamma (1, 1) prior distribution is applied than when the Gamma (0.1, 0.1) or Gamma (0.001, 0.001) prior distributions are applied. Use of the Gamma (1, 1) prior, compared to the Gamma (0.1, 0.1) and Gamma (0.001, 0.001) priors, means that larger positive values for the variance components can be sampled and so the estimates are likely to be larger when Gamma (1, 1) is used than when Gamma (0.1, 0.1) and Gamma (0.001, 0.001) are used.

In *Model 3b* a pattern, similar to that in *Model 3a*, can be seen when gamma prior distributions are placed upon $1/\sigma_j^2$, regardless of the prior placed on the between study-type precision, $1/\nu^2$. When the half-normal prior distributions are placed on $1/\sigma_j^2$, they generally give similar results when a gamma prior is placed on the between study-type precision parameter, $1/\nu^2$ (see Figures 5.9 A-C), but give slightly different estimates and 95% CrIs when a half-normal prior is placed on $1/\nu^2$ (see Figures 5.9 D-F). In this instance, the half-normal prior which allows for the most variation in the estimation, $N(0, 10^6)$, provides results with wider 95% CrIs than those from the $N(0, 10^4)$ and $N(0, 10^2)$ priors, as might be expected.

Priors for the between study-type precision, $1/\nu^2$

Comparing analyses by the between study-type precision prior means comparing results across the plots A-F in Figures 5.8 and 5.9. In *Model 3a* (Figure 5.8) the dose-response slope estimates from the different prior distributions are very similar, although there are differences in the precision of these estimates. Results from placing gamma prior distributions on $1/\nu^2$ are not that different in terms of point estimate and precision, but there are considerable differences in the precision of each estimate when a half-normal prior distribution is used (see Figure 5.8).

In *Model 3b* (Figure 5.9) the different half-normal prior distributions placed upon the between study-type precision, $1/\nu^2$, gave similar results to each other. When the Gamma (1, 1) prior distribution is placed upon $1/\nu^2$, the results obtained have very little precision.

These sensitivity analyses demonstrate that the choice of prior distributions placed upon the unknown precision parameters in *Model 3* may have a very important influence on the results obtained. As Lambert *et al.* (2005) discuss in their simulation study of the impact of prior distributions in a meta-analysis scenario, it is difficult to specify truly vague priors, particularly for variance parameters. This is more problematic when there are, in effect, only two pieces of evidence from which to estimate the variance, as is the case with *Model 3a*. However, none of the pooled slope estimates obtained from the analyses of *Model 3* suggested a significant increase or decrease in the risk of low birth weight with exposure to THMs, and so the pooled estimates are robust to different choices of 'vague' prior, but the variability if these estimates are sensitive to the 'vague' prior placed on the parameters.

5.7.4 Changing relevance of the animal data

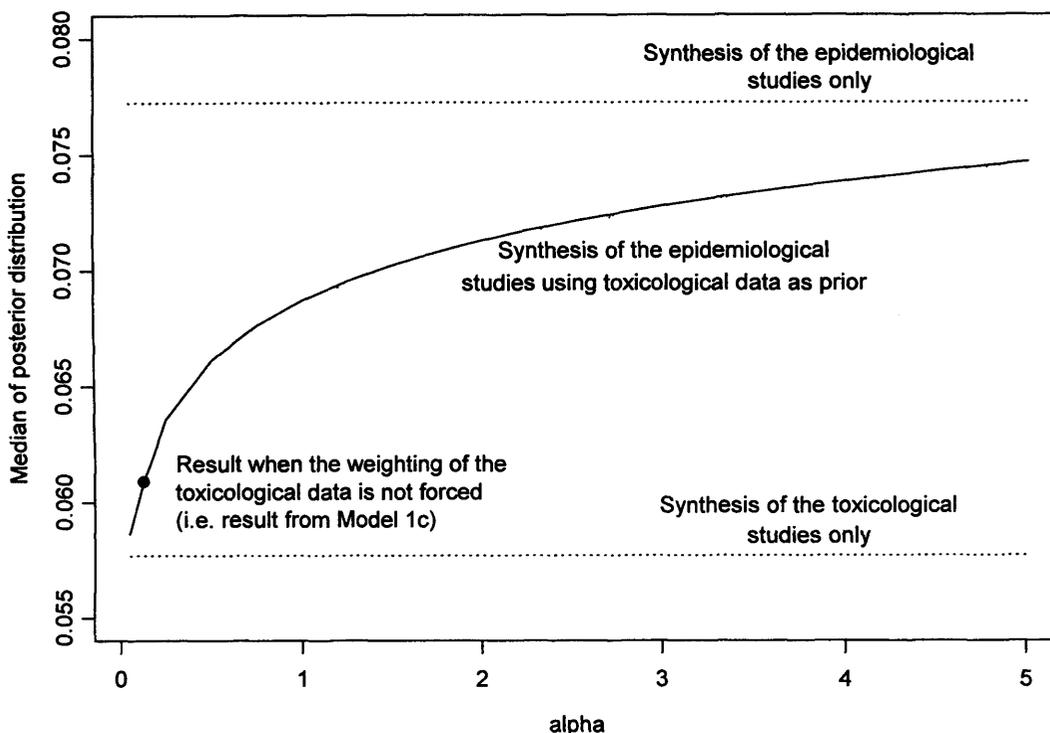
Using a Bayesian model, it is possible to include judgements on the relevance of the data from the animal experiments to the setting of standards of safe exposure to humans. In Section 5.5.1, how prior distributions can be placed on the human data that relates to the information available from the animal data was demonstrated (*Model 1c*). As reviewed in Chapter 3, Prevost *et al.* (2000) and Sutton and Abrams

(2001) have looked at how the weight of one source of evidence may be changed in accordance with prior beliefs concerning the relevance of that evidence in the synthesis. In this chapter, the animal data dominate in *Model 1c*, but each source of evidence can be forced to have the same weight in the synthesis by making the variance of the animal data (which is used to form the prior distribution) the same as the variance of the human data. The effects on the overall pooled estimate of the human data when the variance of the animal data (used to form the prior distribution) is changed can be investigated. Rather than using the three prior ‘beliefs’ that Sutton and Abrams (2000) used (see Section 3.5.2) to weight the animal data, a range of weights are employed here. Figure 5.10 shows the pooled estimates corresponding to the weight given to the animal experiments. α is the ratio of the variance of the human data and the animal data, so that, for example, when $\alpha = 2$, the variance of the animal data is twice that of the human data, so the human data has twice as much weighting in the synthesis. In the analyses reported in Section 5.5.1, where the animal data were used to form the prior for the synthesis of the human data (*Model 1c*), α is equivalent to 0.13 (as indicated on Figure 5.8). Thus highlighting the fact that the animal data have much more weighting in the synthesis than the human data.

Using different weights for the animal data to inform the synthesis of the human data has quite an impact on the results of the synthesis. Expert judgment can be incorporated to assess the relevance of the animal data to the human data in setting environmental exposure limits. Other methods addressing the issue of weighting in the synthesis of studies are available. Ibrahim and Chen (2000) explore the use of a power transform prior, whereby the likelihood function of historical evidence is raised to the power of ρ , where $0 \leq \rho \leq 1$ (such an approach could be applied to our example and possibly extended to include more sources of evidence). Prevost *et al.* (2000) extend their Bayesian three-level hierarchical model by incorporating constraints on certain parameters. For instance, they force the evidence from RCTs to be less variable than the evidence from observational studies. These additional methods, use of power transform priors (Ibrahim and Chen, 2000) and placing constraints on the different sources of evidence (Prevost *et al.*, 2000), are

demonstrated in Chapter 6 using evidence for an assessment of the neurobehavioural health risks from exposure to Mn.

Figure 5.10 Results of changing the relevance of the animal data on the pooled dose-response slope estimate from the human data



Alpha is the ratio of the variance of the animal data & human data, when $\alpha=1$ the animal data and the human data have equal weighting in the synthesis.

5.8. Interpretation of dose-response estimates

The pooled dose-response slopes estimated from *Models 1, 2* and *3* must be interpreted with care. It is not the intention of this assessment to obtain an average dose-response effect between exposure to THMs and low birth weight *across* species, but rather to estimate the effect in humans while taking account of all available and relevant evidence. This has meant the use and inclusion of evidence from animal experiments. It may therefore be argued that in *Models 3a* and *3b* the

main parameter of interest is θ , the pooled dose-response slope for the human evidence, rather than μ , the overall species effect.

The pooled dose-response slope estimates obtained in this chapter could be thought of as a first step in the BMD approach described in Section 2.3.1, i.e. calculation of a dose-response slope estimate. However, since regulatory agencies are now tending towards the use of slope estimates as the basis of setting exposure limits (Edler *et al.*, 2005) rather than use of a single figure (since slope estimates are more informative (e.g. an estimate of risk for any level of exposure can be obtained) and are also relatively easy to interpret (assuming parsimony has been used to model the dose-response relationship)), no choice of critical effect needs to be made. Furthermore, since human and animal evidence are considered and, as shown, their relevance investigated using prior distributions, UFs are not required.

The aim of this chapter was to explore the usefulness of systematic review and meta-analysis methods in risk assessments, not to identify an exposure limit for THMs and the risk of delivering a low birth weight baby. However, in Table 5.8 details of the estimates of risk for low birth weight based on results from *Models 1c* and *3b* are presented. This provides some illustration of how the dose-response slopes could be used in practice to inform human health risk assessments.

Table 5.8 Risk of delivering a low birth weight baby derived from synthesis Models 1c and 3b

Synthesis model	Parameter	Dose-response slope estimate	OR (95% CrI) for delivering a low birth weight baby when exposed to 100 ppb THMs
Model 1c	Human effect, μ	0.061 (0.023, 0.099)	1.0002 (1.0001, 1.0003)
Model 3b	Human effect, θ	0.068 (0.001, 0.156)	1.0002 (0.9999, 1.0005)

* defined as the increase in $\ln(\text{OR})$ per unit increase in $\ln(\text{dose})$

Estimates suggest a very small increased risk of delivering a low birth weight baby when exposed to 100 ppb of THMs compared to no exposure to THMs. Differences in the dose-response slope estimates from *Models 1c* and *3b*, have little effect on the resulting OR for low birth weight. There is, however, more uncertainty in the estimate from *Model 3b*, providing less evidence than that from *Model 1c* on the increased risk of delivering a low birth weight baby from exposure to 100ppb of THMs. The increased risk indicated from the ORs reported here, is very small, but considering the large number of pregnant women potentially exposed to this level of THMs, such estimates could represent a large number of babies in practice.

It could be argued, however, that it is not the posterior distribution that should be used here (Ades and Higgins, 2005). If these findings are to be applied to a slightly different human population, it would be more appropriate to use the predictive distribution. For example, the human epidemiology studies included in this example were all carried out in North American, and so any uncertainty in the likely effect seen in the UK could be accounted for in the predictive distribution. This is, however, beyond the scope of this thesis, but discussion of this can be found in Prevost et al (2000) and Spiegelhalter et al (2004) among others.

5.9 Summary

The potential for systematic review and formal synthesis models to help in a risk assessment of an increased risk of delivering a low birth weight baby for exposure to THMs has been shown in this chapter. Many assumptions have had to be made, but use of these methods has forced explicit acknowledgement and description of these assumptions. The sensitivity of many of these assumptions has been investigated and depending on the dose-response model used, the assumptions made can have a critical effect on the estimates obtained.

The sensitivity analyses have also demonstrated the necessity of checking the appropriateness of the chosen dose-response model. For simplicity, a linear relationship between the $\ln(\text{OR})$ for delivering a low birth weight baby and $\ln(\text{dose})$ was initially assumed. However, the logit model is among the most commonly

applied dose-response models in toxicology, together with the probit and Weibull models (Crump, 1984; Covello and Merkhofer, 1994; Lovell and Thomas, 1997). Here, the application of the BIC has suggested that the linear model provides a better fit to the data than the logit model. The use of 'ln(dose)' would seem more sensible than 'dose', as it reflects a multiplicative effect with increasing exposure. In fact, in toxicology and pharmacology, 'ln(dose)' is generally used (Rang *et al.*, 1999). In this example 'ln(dose)' is more appealing as the epidemiological slope estimates from the 'ln(dose)' dose-response model have comparable magnitude and precision to those from the animal experiments. All of the dose-response models have assumed that, for the outcome of low birth weight, THMs are non-threshold substances (any exposure, however small, could increase the risk of delivering a low birth weight baby, see Sections 2.2 and 2.3). Although most exposures are believed to be threshold substances, given the lack of evidence on a possible threshold, this assumption appears reasonable. However, further work could consider whether a threshold effect exists for exposure to THMs.

A particular issue related to the construction of the dose-response slopes is that the animals are generally subjected to much higher doses than humans (i.e. 400 mg/kg/day vs. 0.0042 mg/kg/day). Thus, evidence from the animal experiments concerns exposures way beyond those likely to be experienced by humans and so one could question the relevance here of the animal experiments. Subjecting animals to lower doses may help to increase relevance to the human situation, however, in order to have the power to detect any effects at these lower doses, many more animals may be required in the experiments, an approach that is not in line with the 3Rs (see Section 2.6).

A number of synthesis models have been used to combine the relevant THMs and low birth weight evidence, and in assessing which to recommend, account must be taken of the fact that, in this example, two different sources of evidence are being combined. Hence use of *Model 1e* to synthesise all 13 studies is not advocated since it does not distinguish between the human and animal evidence. For this example, it is not clear which synthesis models are the most advantageous. In terms of taking into account the two different sources of evidence, Models 1c and 1d, which used informative prior distributions, are of interest. Since the human

epidemiological data is of primary importance, *Model 1c* which synthesises the human evidence with the animal evidence forming the prior distribution is the more appealing of these two models. By assessing the relevance of the animal evidence and allowing it to have less influence in the synthesis, the impact of this particular source of evidence in *Model 1c* can, and has been, explored quantitatively.

However, in *Model 3b* the human and species/strain specific evidence are modelled explicitly, with the hierarchical framework allowing the pooled human estimate to borrow strength from the pooled animal estimates. Moreover some idea of the between species-strain variability is obtained through the ν^2 parameter in Equation 5.4. Although estimated with a great deal of uncertainty (since it is based on very little information and the prior distribution placed on it is vague) ν^2 could be used to investigate the relevance of the evidence from different species and strains to humans as part of future work. Since choice of synthesis model may not be clear, methods of model comparison could be extended (e.g. from use of BIC to inform choice of dose-response model in Table 5.4) to the use of Bayes Factors and averaging over the synthesis models (Kass and Raftery, 1995), as carried out by Sutton and Abrams (2001) in WinBUGS.

The sensitivity of limiting the systematic review in Section 5.3 to English language articles only has not been assessed. Such limitations could possibly lead to a biased set of evidence in the systematic review as pointed out in Section 4.4.2. Although problems of language translation arise when a search is not restricted to language, some foreign language articles tend to provide an English language abstract. This could provide sufficient information for inclusion into a systematic review when translation of the full text is not possible. However it is still unclear as to the impact language restrictions have (Egger *et al.*, 2003).

The impact of two assumptions made for the synthesis of the different evidence in this chapter has not been discussed. Dose-response slopes from two cohort, two case-control and one cross-sectional epidemiological studies were synthesised without account taken of the difference in designs. These differences could be important if bias in the study estimate is related to study design. Thus, one is assuming that estimation of the dose-response relationship is the same regardless of

the study design and measures of effect used. Methods for the synthesis of different epidemiological study designs are available (Müller *et al.*, 1999; Martin and Austin, 2000; Spiegelhalter and Best, 2003; Wolpert and Mengersen, 2004) and could be used and incorporated into the hierarchical models used in Sections 5.6. The effect of these different techniques to account for the fact different epidemiological study designs are synthesised would form further work into the sensitivity of the analyses carried out.

In this chapter, it was assumed that levels of total THMs, chloroform, BDCM, CDBM and bromoform are all equivalent (i.e. Total THMs = chloroform = BDCM = CDBM = bromoform). In reality the following relationship is true: Total THMs = chloroform + BDCM + CDBM + bromoform. Research has suggested that measures of chloroform and total THMs are highly correlated, but that correlations between total THMs and other individual THMs are not convincing (Whitaker *et al.*, 2003). One approach would be to consider that taken by DuMouchel and Harris (1983) where data from slightly different exposures are used to inform the exposure of interest and this is being investigated (Peters *et al.*, 2003).

Since this work was carried out three articles have been found that are relevant to this example: two human epidemiological papers each reporting a study of exposure to total THMs and the risk of delivering a low birth weight baby (Toledano *et al.*, 2005; Wright *et al.*, 2005) and one paper reporting two animal experiments investigating exposure to BDCM in rats and rabbits (Christian *et al.*, 2002). Toledano *et al.* (2005) report increasing adjusted ORs for delivering a low birth weight baby with increasing levels of exposure to total THMs, although the 95% CIs suggest that the evidence of an effect is not strong (low exposure, OR = 1; medium exposure, OR = 1.05 (95% CI 0.96, 1.15); high exposure, OR = 1.09 (95% CI 0.93, 1.27)). Wright *et al.* (2005) do not observe strong evidence of an increase in the adjusted OR for low birth weight risk with exposure to total THMs (low exposure, OR = 1; medium exposure, OR = 0.97 (95% CI 0.81, 1.26); high exposure, OR = 1.05 (95% CI 0.85, 1.29)). Christian *et al.* (2002) observe an increase in foetal weight with exposure to DCBM of up to 150 parts per million (ppm) in both the rat and rabbit experiments compared to no exposure to DCBM, but a decrease in foetal weight for exposures of 450ppm and 900ppm. As part of

further work these articles, and any subsequently identified, will be included in a re-analysis to provide an up-dated synthesis of the evidence, hopefully offering more information.

In this example a reasonable amount of similar evidence was available to apply systematic review and meta-analysis methods in an attempt to illustrate their potential in human health risk assessments. However, in reality this example was quite artificial in terms of human health risk assessment as a single health outcome was chosen (rather than assessing a range of possible health outcomes that may or may not be seen as more severe or occurring at lower levels of exposure as would be done in a traditional risk assessment). For instance, other related outcomes should be considered e.g. premature birth, stillbirths. When dealing with a number of potentially adverse effects a more formal decision modelling approach may be desirable so that both the evidence on, and the impact of, of these effects may be simultaneously assessed whilst allowing for appropriate correlation and uncertainty (Cooper *et al.*, 2004). In the next section a second risk assessment example is used to illustrate the potential usefulness of systematic review and meta-analysis methods in human health risk assessments.

Manganese and neurobehavioural effects: example II

6.1 Chapter overview

The potential usefulness of systematic review and meta-analysis methods in human health risk assessment for environmental chemicals is further explored in this chapter using the example of occupational exposure to manganese (Mn) and adverse neurobehavioural effects. This example is quite different to that used in Chapter 5, which allows assessment of the general, and more specific, issues involved in applying systematic review and meta-analysis methods to human health risk assessment. The example of Mn and neurobehavioural effects is introduced in Section 6.2 and compared to the THMs and low birth weight example of Chapter 5. In Section 6.3, a systematic review of the evidence relevant to a risk assessment of exposure to Mn and neurobehavioural health effects is described. Advantages of this approach to identifying and reviewing the evidence are discussed. The diverse human and animal data found to be relevant for this risk assessment are described in Section 6.4. Data on a particular outcome, activity level are synthesised within each different species in Section 6.5. Advantages of synthesising pieces of similar evidence and graphical presentation of the data are discussed. In Section 6.6, a subset of the activity level evidence in which data exist from human occupational epidemiology studies, and from rat, mouse and bird experiments are used to illustrate methods to synthesise evidence from diverse sources. Finally, in Section 6.7, further meta-analysis techniques are discussed in terms of their ability to help overcome some of the limitations of current risk assessment methods.

6.2 Introduction

To evaluate the potential of systematic reviews and meta-analyses in human health risk assessments it is advantageous to have illustrative examples where many differences exist between them so that general issues of the application of systematic reviews and meta-analyses can be investigated in addition to those specific to the individual examples. Previous reports suggest that neurobehavioural effects are the most common health effect of increased exposure to Mn generally observed at lower levels of exposure (Clewell *et al.* 2003). As a consequence, neurobehavioural health effects are investigated in this chapter.

The example of occupational exposure to Mn and neurobehavioural health effects is quite different to THMs and low birth weight (Chapter 5) for a number of reasons. The exposure profiles (the *levels* of exposure, the likely *populations* exposed and the *routes* of exposure) represent many of these differences. Mn is an essential element; so not only is there a level at which exposure to Mn is harmless, but exposure at a certain level is required to maintain good health (WHO, 1981). In Chapter 5, THMs were assumed to be non-threshold substances, and so any dose-response modelling for Mn will differ to that for THMs, since the threshold effect needs to be accounted for.

Everyone is exposed to Mn, but although it is an environmental substance found in rock, soil, water and food, the major source of exposure of Mn at toxic levels is occupational (WHO, 1981). Therefore unlike THMs where whole populations are potentially exposed but at quite modest levels, fewer people are at risk from high levels of exposure to Mn (i.e. only those working in particular industries), but the levels of exposure they experience are quite substantial. Another difference in the exposure profile of Mn compared to THMs, is the route of exposure: the main route of exposure to toxic levels of Mn is via inhalation (e.g. dust), whereas the main route of THMs exposure is via drinking water. This difference has implications when comparison of the human and animal studies is considered. In the THMs example all of the animal experiments reported oral exposure to THMs to coincide with the main route of human exposure, however, as will be seen in Section 6.3, only a small percentage of the relevant animal experiments expose animals via

inhalation (as the main route of human exposure). Thus, comparison across human and animal experiments on Mn is made more complex.

The health outcomes considered in the two examples are also quite different. Whereas low birth weight is easily measured and defined, and can be applied globally (WHO, 2004), neurobehavioural health effects represent a number of different outcomes involving adverse effects relating to emotion, learning and behaviour, which are often measured using a number of different tests and techniques. As a consequence, adverse neurobehavioural effects are much more difficult to quantify and define as will be seen in this chapter. This means that consistency across outcomes is a particularly important feature of human health risk assessments where neurobehavioural effects are of interest, and as will be shown, systematic review and meta-analysis methods can facilitate such assessments.

To build upon the work of Chapter 5, a broader range of possible health effects are assessed in this chapter, neurobehavioural effects, as opposed to a specific outcome (as in the THMs example, low birth weight). Initial species-specific meta-analyses use the models given in Equations 3.1 and 3.2 in a classical statistical framework (Section 6.5). In Section 6.6, Bayesian methods of synthesis, following on from those applied in Chapter 5, are used.

6.3 The systematic review

As with Chapter 5, the focus here is on the potential of evidence synthesis methods and not the systematic review methods used to identify the evidence and so only limited details are given. Appropriate guidelines were followed (Deeks *et al.*, 2001) to conduct a systematic search of all potentially relevant human studies and animal experiments for a risk assessment of the neurobehavioural effects from exposure to Mn. The search was carried out in Medline and the search strategy used is given in Appendix G. No limitations were placed on the language of the article, although articles not in English were not translated. As discussed in Sections 4.4.3 and 5.9 current research suggests the implications of using English language articles only are unlikely to induce bias into the review. The Medline search was supplemented

by a review of references used in a draft version of the IEH/IOM joint report on occupational exposure limits for manganese (IEH and IOM, 2004) and the ATSDR Toxicological Profile for Manganese (ATSDR, 2000). Relevant articles had to report details of a human study or animal experiment on possible neurobehavioral effects from exposure to Mn. Neurobehavioural effects were defined to be any adverse effects associated with learning, emotion and behaviour. The bibliographies of all relevant articles were searched to identify further references that may be of interest. From a list of 579 potentially relevant articles, 92 were found to be relevant, 55 of which were human epidemiology studies and 37 were *in vivo* animal experiments. A list of these 92 references is given in Appendix G. There are many differences between these studies in terms of the species and strains of animals and the routes of exposure used in the experiments. These differences are illustrated in Table 6.1.

There is a great amount of diversity in the evidence relevant for this risk assessment. In the animal experiments, not only are different strains of species used (e.g. Sprague-Dawley and Wistar rats), but experiments have been carried out in quite different species, from mice and birds to dogs and monkeys. Moreover, the route of exposure to Mn used in a study appears to be related to the species used. The majority of experiments in rats assess oral exposure to Mn, monkey experiments generally involve injecting Mn and human studies mainly concern the inhalation of Mn. In the following section the diversity of the relevant human and animal evidence is described and discussed in further detail.

Table 6.1 Diversity of evidence available for the risk assessment of manganese and neurobehavioural effects

Species	Strain	Route of exposure		
		Inhalation	Oral	Injection
Rat	Sprague-Dawley	1	4*	3*
	Wistar		2	
	ITRC albino		2	
	CD		1	
	ITRC		1	
	Albino		3	
	<i>Total</i>		<i>1</i>	<i>13</i>
Mice	CD-1		1	
	ddY		2	
	Swiss	1		
	ICR Swiss	1		
	<i>Total</i>	<i>2</i>	<i>3</i>	<i>0</i>
Birds	Quail		1	
Dogs	Beagle			1
Rabbits	unknown	1		2
Monkeys	Rhesus	2	1	2
	Macaca fasticularis			3
	Indian red-haired			1
	Squirrel			1
	Cebus			1
	<i>Total</i>	<i>2</i>	<i>1</i>	<i>8</i>
Humans				
	- Occupational studies	50		
	- Environmental studies		5	
	<i>Total</i>	<i>50</i>	<i>5</i>	<i>0</i>
Overall total		56	23*	14*

* Two experiments (oral and injection) are reported in one article

6.4 The evidence

6.4.1 Human epidemiological evidence

All but five of the human studies are occupational epidemiology studies of a cross-sectional design where inhalation is the route of exposure to Mn (Table 6.1). The earliest human epidemiology studies relevant to this review were published in 1941 (Flinn *et al.*, 1941; Kawamura *et al.*, 1941), and have been carried out in many countries across the world including those in Europe, Asia, South and North American and Australasia. The different occupations considered include mining, steel works, dry cell battery factories, welding, enamels production factories and alloy plants. Although the focus of this chapter is occupational exposure to Mn, non-occupational studies may provide useful information on health effects. The five environmental studies concern oral exposure to increased levels of Mn in the water supply due to environmental levels (Kondakis *et al.*, 1989), intoxication from nearby industry (Beuter *et al.*, 1999; Vieregge *et al.*, 1995), and accidental intoxication (Kawamura *et al.*, 1941; Kilburn, 1987).

Exposure to manganese

Many different Mn compounds are assessed in the occupational epidemiology studies, e.g. pyrolusite, Mn dioxide, ferromanganese. For an exposure assessment typically an 'amount' of exposure has to be determined from a measured level of some form of Mn and take into consideration length of exposure. Often in an occupational study a level of exposure cannot be measured or estimated, and so to classify the amount of Mn exposure, workers are defined in terms of the type of job they do or in which areas of a factory, say, they spend most of their working time. Often exposure assessment can be quite uncertain. Comparisons across studies between the exposures individuals experience is difficult. The different approaches used to monitor and measure exposure, the different levels of exposure experienced (e.g. very high for miners, much lower for welders) and the different compounds individuals may be exposed to all increase the complexity of any cross-study comparison.

In the occupational epidemiology studies where Mn has been measured, levels in air, blood, urine and hair are reported. In some studies only length of exposure

(estimated by length of employment) is reported (Abd El Naby and Hussanein, 1965; Kilburn, 1987; Sjögren *et al.*, 1990; Huang *et al.*, 1993; Kilburn, 1998; Kilburn, 1999). Other studies do not attempt to determine an 'amount' of exposure, but just report whether subjects were exposed or unexposed (Rodier, 1954; Mena *et al.*, 1967; Stýblová *et al.*, 1979). Generally, these are the oldest studies, when methods to measure Mn exposure were not available.

In cases where Mn has been measured or estimated, it is typically done so as a level of exposure in air and reported as a volume, mg/m³. However, there are yet more differences in how Mn levels in air are estimated and reported in the occupational epidemiology studies varying in cost, practicality and accuracy. These levels may be measured from a stationery device (placed in a particular area of the factory) or a personal device (attached to an individual in an attempt to measure their personal exposure levels). Additionally, some studies measure and report *total* Mn dust levels in air (the amount of dust taken in by an individual) while other studies measure and report *respirable* levels (the amount of dust likely to enter the respiratory tract – a proportion of the total dust levels).

Attempts to compare exposure levels across studies are further limited by how the levels are reported; for instance, the range of exposure levels in the exposed and control groups (Brown *et al.*, 1991), the average estimated or measured exposure level (Gibbs *et al.*, 1999), upper limits for exposure (Flinn *et al.*, 1941; Hua and Huang, 1991). Thus comparison of exposure levels is incredibly difficult with such diverse measuring and reporting of Mn levels in occupational settings, reflecting issues of cost, practicality and accuracy. This not only has implications for comparison across a number of human studies, but also in comparison with findings from relevant animal studies. If individual subject (patient and animal) data were available, many of these problems could be overcome to ease comparison. However, these data are not available and so comparison is made at an aggregate level. This not only brings challenges for the comparison of exposure levels across studies, but introduces further problems such as the ecological fallacy (this is discussed further in Section 6.5.2.)

Neurobehavioural health effects

The different health outcomes and how they are tested add even further complexity to the comparability of the human epidemiology studies. The types of neurobehavioural health effects can be summarised using categories set out by Iregren (1999). Although Iregren (1999) describes these categories as arbitrary, since different tests often contain aspects of more than one particular category, they provide a useful basis for assessment of the outcomes and have been used by others (Fern-Pollak *et al.*, 2004; DuMont *et al.*, 2005). They are:

- Motor function
- Memory
- Reaction time
- Other cognitive function
- Mood and subjective symptoms

In Table 6.2 the different neurobehavioural health effects assessed in the human epidemiological studies are presented. Each row represents a type of study which has been defined by the health effects measured. The first row indicates there are four studies where all five types of neurobehavioural outcomes were investigated; the second row shows that two studies report an assessment of motor function, memory, reaction time and other cognitive function, but not mood or subjective symptoms.

27 of the 55 epidemiology studies report using tests to assess neurobehavioural effects from at least two of Iregren's outcome categories; seven studies only test aspects of motor function (penultimate row of Table 6.2). In these 34 epidemiology studies test results are reported in terms of mean scores (and usually standard deviations) for the exposed group and the control group. Many different tests are reportedly used to measure the same neurobehavioural effect. For example, 13 different types of motor function tests are identified in the 55 epidemiology studies, with some of these having up to 5 different versions (e.g. for the pegboard test which measures dexterity, Purdue, Santa Ana, Grooved, SPES and undefined versions are reportedly used).

Table 6.2 *The type and frequency of neurobehavioural effect outcomes investigated in the 55 human epidemiological studies*

No. studies assessing combinations of outcomes	Motor function	Memory	Reaction Time	Other Cognitive function	Mood/ subjective symptoms
4	X	X	X	X	X
2	X	X	X	X	
2	X	X	X		X
1	X	X	X		
1	X	X		X	X
3	X	X		X	
2	X		X	X	X
1	X		X	X	
1	X		X		X
1	X		X		
1	X			X	X
1	X			X	
3	X				X
1		X	X	X	X
2		X		X	X
1		X		X	
7	X				
21					X

In the 21 studies that only reported outcomes for mood or subjective symptoms, no tests are carried out (last row of Table 6.2); instead the number of exposed or control subjects judged to be suffering a particular symptom is reported (although some of these studies do not report findings for a comparative control group). Thus, some studies report mean scores for many different tests and neurobehavioural outcomes while others report the proportion of individuals indicating changes in mood or subjective symptoms. These differences in how the outcomes are assessed and presented impacts upon the comparability of these studies, and have severe implications for the synthesis of such evidence.

6.4.2 Animal experiment evidence

Of the 37 relevant animal experiments, 16 are on rats (including one study presenting results for both oral and injection exposure), 11 on monkeys, 5 are experiments on mice, 3 on rabbits, one on dogs and one on birds. Within these species a number of different strains are used (see Table 6.1).

Exposure to manganese

As with the human evidence, there is a great deal of variation in how exposures are measured and reported in the animal experiments. The routes of exposure used in the experiments differ between the animals used. Half of the experiments use oral exposures to Mn, typically given in drinking-water. In general it is rats that are orally exposed to Mn, while monkeys are much more likely to be exposed through injection. Six of the experiments use inhalation as the route of exposure reflecting the usual route of human occupational exposure to Mn.

Different Mn compounds are assessed including manganese chloride, manganese oxide and manganese carbonate. The frequency and length of exposure also varies between experiments. Although dose levels are generally well reported in these experiments when drinking water is the route of exposure to Mn, animals are generally caged together, and so how much a particular animal drinks is unknown. Hence, the amount of Mn an animal is exposed to in practice may not be as accurately reflected in the reporting of the experiment, as one may have been led to believe by the fact that animal experiments are more controlled than human epidemiology studies.

Neurobehavioural health effects

Animals are tested for neurobehavioural effects in 18 of the 37 experiments. The mean scores and standard deviations for each test at each dose level are usually reported in the experiments. In the other 19 experiments certain behaviours and symptoms related to neurobehavioural health effects are observed in the animals by investigators (such as lethargy, unsteady gait, tremor, muscular weakness). The number of animals per dose group exhibiting these various behaviours and symptoms are reported in the experiments.

6.4.3 Summary

In this section, the diversity in design, measurement and reporting of studies relevant to an assessment of the risks to neurobehavioural health effects from exposure to Mn has been highlighted. The many differences in study population, Mn exposure (in terms of compound, level, length and frequency), neurobehavioural health effects assessed and the tests used for an assessment are

unlikely to be specific to this example. The diversity described in this section highlights some of the challenges faced by risk assessors in an attempt to review and evaluate all of the relevant evidence for a comprehensive risk assessment. Having already shown that a systematic review offers a way of identifying and reviewing evidence that allows transparency and reproducibility, in the next section I demonstrate how meta-analysis methods can help overcome some of the limitations of current risk assessment approaches. Because of the diversity of the evidence a synthesis of all the relevant data may be limited. Instead systematic review and meta-analysis methods can be used to summarise evidence on similar exposures and effects, strengthening the risk assessment by offering a more structured and transparent framework. Meta-analysis methods are applied to a particular neurobehavioural health effect, activity level, in an attempt to demonstrate how these methods can be used in the risk assessment process, overcoming some of the current limitations of the risk assessment process.

6.5 Synthesis of the evidence: activity level

This particular outcome, activity level, was chosen as research suggests that changes in motor function abilities are the first symptoms of adverse exposure to Mn (ATSDR, 2000; WHO, 1981) and activity level is one of many markers for motor function ability. However, this health outcome is quite subjective and, perhaps, vague in that there is no threshold corresponding to normal activity levels by which to make comparisons. Moreover, many factors could contribute to reduced or increased activity levels, but because such vague and subjective health outcomes are often the first indications of an adverse effect in humans they are important to monitor. Sixteen human epidemiology studies report level of activity as a health outcome in the Mn literature; a further 17 animal experiments are also relevant to an assessment of this outcome. Activity level scores are available in four of the human epidemiology studies and ten of the animal experiments. For each of these, the mean activity scores in the exposed and control groups are reported. In the other human epidemiology studies and animal experiments, activity level is reported as the number of subjects reporting, or observed to have, a negative symptom relative to activity level (e.g. fatigue, exhaustion, restlessness). Some studies report both

types of responses. Application of meta-analysis methods to the human evidence are described in Section 6.5.1, in addition to the difficulties and limitations of any synthesis. Similarly, meta-analysis methods to evaluate the animal activity level evidence are described in Section 6.5.2. In Section 6.6, evaluation of all the evidence (human and animal) using meta-analysis methods is demonstrated.

6.5.1 Assessing the human evidence

Before any synthesis of the human epidemiological evidence on activity level and Mn exposure can be done, four issues need consideration. Firstly, relevant data from two studies (Wennberg *et al.*, 1991; Camerino *et al.*, 1993) is not clearly reported and so these studies are excluded from further analyses. The second issue is that a number of different symptoms have been investigated and reported. Fatigue is the most commonly investigated symptom and so in the synthesis of activity level evidence the focus will be on fatigue and fatigue-like symptoms of drowsiness, asthenia, tiredness and lethargy. Scores for vigour from the four epidemiology studies using questionnaires will also be considered and synthesised. Thirdly, a number of studies report more than one outcome as a measure of activity level for the same set of subjects. Although these can be compared to assess the consistency of reporting, these symptoms are likely to be correlated (e.g. vigour and fatigue), and so this correlation should be taken into account. However there is a lack of individual data on how the various symptoms correlate with each other and so to avoid this correlation, only one activity level outcome from each study will be included in the analyses. The fourth issue is that measures of fatigue or vigour from four studies are reported as mean scores and their variances, while the remaining studies report the number of subjects in the control and exposed groups exhibiting a particular fatigue-like symptom. Clearly these different measures of fatigue cannot be compared directly. For the four studies reporting mean fatigue scores from a questionnaire, the standardised mean difference between scores in the control group and those in the exposed group have been calculated by

$$d = \frac{x_0 - x_1}{\sigma_{01}} \quad (6.1)$$

where x_0 is the mean score for the control group, x_1 is the mean score in the exposed group and σ_{01} is the sample standard deviation defined below, where σ_0 and σ_1 are the standard deviations in the control and exposed group, and n_0 and n_1 are the number of subjects in the control group and the exposed group, respectively.

$$\sigma_{01} = \sqrt{\frac{(n_0 - 1)\sigma_0^2 + (n_1 - 1)\sigma_1^2}{n_0 + n_1 - 2}} \quad (6.2)$$

The variance of d , the standardised mean difference is given by Equation 6.3 (Sutton *et al.*, 2000).

$$\text{var}(d) = \frac{n_0 + n_1}{n_0 n_1} + \frac{d^2}{2(n_0 + n_1)} \quad (6.3)$$

For studies in which the number of subjects with a fatigue-like symptom are reported, the $\ln(\text{OR})$ for reporting a fatigue-like symptom in the exposed group compared to the control group is calculated. Let n_{1e} be the number of subjects in the exposed group exhibiting a symptom, n_{2e} be the number of subjects in the exposed group *not* exhibiting a symptom, with n_{1c} the number of subjects in the control group exhibiting a symptom and n_{2c} the number of subjects in the control group not exhibiting a symptom by

$$\begin{aligned} \ln(\text{odds})_e &= \ln\left(\frac{n_{1e}}{n_{2e}}\right) \\ \ln(\text{odds})_c &= \ln\left(\frac{n_{1c}}{n_{2c}}\right) \end{aligned} \quad (6.4)$$

$$\ln(\text{OR}) = \frac{\ln(\text{odds})_e}{\ln(\text{odds})_c}$$

where a comparative control group is not reported only the $\ln(\text{odds})_e$ is calculated. The studies, the symptoms reported and the outcome measures calculated are displayed in Table 6.3.

Table 6.3 Evidence from the human epidemiology studies used in subsequent analyses

Study	Year	Symptom	No. exposed	No. controls	Standardised mean difference in symptom score		Ln(OR)		Ln(odds) in exposed group	
					Estimate	Variance	Estimate	Variance	Estimate	Variance
Huang <i>et al.</i>	1990	Vigour	61	61	0.181	0.033				
Huang <i>et al.</i>	1990	Fatigue	61	61	-0.545	0.034				
Mergler <i>et al.</i>	1994	Vigour	74	74	0.293	0.027				
Mergler <i>et al.</i>	1994	Fatigue	74	74	-0.630	0.028				
EC	1997	Vigour	151	98	-0.226	0.017				
EC	1997	Fatigue	151	98	-0.238	0.017				
Kilburn	1998	Vigour	41	66	1.066	0.045				
Kilburn	1998	Fatigue	41	66	-1.190	0.046				
Flinn <i>et al.</i>	1941	Drowsiness	23	16			0.788	2.772	-2.708	0.711
Mena <i>et al.</i>	1967	Asthenia	13	8			6.129	4.192	3.296	2.074
Jonderko <i>et al.</i>	1971	Asthenia	46	45			0.143	0.338	-1.661	0.159
Saric <i>et al.</i>	1977	Fatigue	369	204			0.164	0.032	-0.355	0.011
Wang <i>et al.</i>	1989	Fatigue	8	32			1.059	0.603	0	0.444
EC	1997	Tiredness	151	98			0.675	2.683	-4.608	0.673
Gibbs <i>et al.</i>	1999	Fatigue	75	75			0	0.10953	-0.400	0.055
Deschamps <i>et al.</i>	2001	Asthenia	138	137			0.334	0.063	-0.408	0.030
Rodier	1955	Fatigue	115						-2.027	0.084
Schuler <i>et al.</i>	1957	Fatigue	15						1.686	0.474
Abd El Naby & Hussanein	1965	Lethargy	45						-1.997	0.206

Using classical fixed and random effects meta-analysis methods (see Equations 3.1 and 3.2 in Chapter 3), pooled estimates for the standardised mean difference in fatigue and vigour scores, $\ln(\text{ORs})$ for reporting a fatigue-like symptom for an exposed individual compared to a control individual and the $\ln(\text{odds})$ of reporting a symptom if exposed are given in Table 6.4.

As described in Section 3.3.2, I^2 gives an estimate of the amount of heterogeneity between studies in a meta-analysis (Higgins and Thompson, 2002). For each meta-analysis carried out here, I^2 was calculated. Since the estimates of I^2 suggest between-study heterogeneity in all of the meta-analyses Table 6.4, only the random effects pooled estimates are discussed. The pooled standardised mean difference in scores from the four studies using a questionnaire to assess activity level suggest that exposed workers experience stronger feelings of fatigue than workers in the control group. This is also reflected in the pooled $\ln(\text{OR})$ for reporting a fatigue-like symptom, where exposed subjects are more likely to exhibit symptoms of fatigue than the control subjects, though not significantly so ($\text{OR}=1.35$; 95% CI 0.91, 2.00). Interpretation of the pooled log odds of exposed workers reporting a symptom is difficult without a control. However since inclusion of the data from three studies that do not have a comparative control group does not appreciably change the pooled $\ln(\text{odds})_e$ calculated for the eight studies with comparative control groups one may argue that the conclusions would not have changed substantially if such control groups did exist for these three studies. This gives confidence to the findings that exposed workers experience more fatigue than workers in the control group. Although some comparison across these types of pooled estimates can be made, i.e. assessing whether they are all indicating a similar pattern, any further comparison or powerful estimate of an effect is not possible with the data as they currently are.

Table 6.4 Combined effects for measures of fatigue and vigour from the relevant human epidemiology studies

Symptom	No. studies in analysis	Standardised mean difference in score			Ln(OR) for reporting a symptom			Ln(odds) for reporting a symptom (exposed group)		
		Fixed	Random	I ²	Fixed	Random	I ²	Fixed	Random	I ²
Vigour	4	0.18 (0.02, 0.34)	0.31 (-0.19, 0.81)	89 (75, 95)
Fatigue	4	-0.54 (-0.70, -0.37)	-0.63 (-1.00, -0.26)	80 (46, 92)
Fatigue/ drowsiness/ asthenia/ tiredness	8	.	.	.	0.24 (-0.01, 0.48)	0.30 (-0.09, 0.69)	33 (0, 70)	-0.47 (-0.63, -0.31)	-0.93 (-1.52, -0.33)	90 (84, 94)
Fatigue/ drowsiness/ asthenia/ tiredness/ lethargy	11	-0.59 (-0.74, -0.45)	-0.98 (-1.59, -0.38)	86 (75, 92)

Methods do exist for the synthesis of outcomes on a continuous scale with those on a binary scale (Whitehead *et al.*, 1999; Chinn, 2000). However, these methods are not applied here for a number of reasons. Firstly, the method proposed by Whitehead *et al.*, (1999) cannot be used as there is no threshold for 'normal' scores of fatigue from the questionnaires. Secondly, to use Chinn's method the assumption that fatigue is measured on the same underlying scale in the different types of studies has to be made. This is a strong assumption to make here as fatigue is actually measured in some studies, but observed in others. If this assumption is made then use of Chinn's transformation could be used to compare levels of fatigue between control and exposed subjects in the 11 studies. Methods also exist to combined evidence from one-armed and multi-armed trials (Begg and Pilote, 1991). This approach could be used here, if there were evidence from studies that only looked at the symptoms of fatigue for control subjects.

So far the health effects have been described in terms of those exposed to Mn and those not exposed, no levels of exposure have been used in the analyses. To help derive an exposure limit for Mn some consideration of the level of Mn workers are exposed to must be made. However as pointed out in Section 6.4.1 exposure data can be poorly reported in studies.

Of the 14 epidemiology studies in Table 6.3, three present no information on exposure levels for the subjects in their study (Rodier, 1955; Mena *et al.*, 1967; Kilburn, 1998). One study uses length of exposure to categorise individuals (Abd El Naby and Hussanein, 1965), a further three studies provide estimates of likely exposure levels in air (Flinn *et al.*, 1941; Wang *et al.*, 1989; Huang *et al.*, 1990), although one of these also measures levels of Mn in blood (Wang *et al.*, 1989). In the remaining seven epidemiology studies, measured exposure levels are reported. Six report measured levels of Mn in air in terms of ranges (Mergler *et al.*, 1994; EC, 1997; Saric *et al.*, 1977; Deschamps *et al.*, 2001), upper limits (Schuler *et al.*, 1957) or calculate a cumulative exposure index (Gibbs *et al.*, 1999). A number of these studies report ranges of Mn levels in blood and/or urine, as well as in air. The final study (Jonderko *et al.*, 1971) reports a lower limit for Mn levels in blood.

This diversity in the reported exposure data adds a further level of complexity to the setting of an occupational exposure limit for Mn and restricts a more quantitative assessment of exposure as subjective judgement is required in making assumptions so that the levels of measured and estimated Mn exposure in different mediums can be compared. However, if such diversity in the reported exposure levels did not exist, the results of an assessment of the exposure levels should be interpreted cautiously, since aggregate exposure data are used. An assessment of exposure would be much more informative if individual subject data were available.

6.5.2 Assessing the animal evidence

As with the evidence from the human epidemiology studies, many differences exist between the 17 animal experiments, and these must be addressed before the data can be synthesised. First, evidence from four of the 17 experiments cannot be included in the analyses as there are insufficient data on either the number of animals in each dose group (Calabrese *et al.*, 1999) or the number of animals exhibiting a negative activity symptom (Chandra, 1972; Subhash and Padmashree, 1991; Dorman *et al.*, 2000). Secondly, in a number of experiments, the animals' activity levels are measured at numerous time-points. To overcome differences in time-points and frequency of the measurements between different experiments, the final (or only) time-point measure reported in each experiment is used as the main outcome.

Thirdly, two experiments (Bonilla, 1984; Pappas *et al.*, 1997) report results for more than two groups, while the remaining experiments have only one exposed group which is compared to a control group. Although reducing the amount of information that can be used, for ease of comparison, the lowest and highest dose groups in the two experiments are used in the subsequent analysis. Fourthly, most experiments test activity levels and report mean scores (and standard deviations) for the exposed and control groups. Since the measurement scales are not the same for all experiments, the standardised mean difference between control and exposed animal activity scores have been calculated using Equation 6.2. A positive standardised mean difference suggests control animals are more active than exposed animals. In the other three experiments animals are observed and the number in each group showing adverse activity level effects is given. From these proportions, the ln(ORs) for an exposed animal to exhibit a negative activity symptom compared

to a controlled animal is calculated using Equation 6.3. Finally data on a number of different species and strains of animal are available in these experiments. The data are analysed by the different species used, thus giving some idea of the observed effect from exposure to Mn across different species. The data available from the 17 animal experiments and the standardised mean differences or lnORs for an effect on activity level are shown in Table 6.5.

The pooled standardised mean differences in activity counts between exposed and control animals (where a positive value indicates less activity in the exposed group, see Equation 6.1) and ln(ORs) for observing a negative activity symptom (see Equation 6.4) using classical fixed and random effects meta-analyses (see Equations 3.1 and 3.2 in Chapter 3) are shown in Table 6.6 by animal species. The pooled standardised mean differences suggest that rats exposed to Mn are much more active than control rats, but this pattern is not seen with exposed mice and birds compared to unexposed mice and birds. In fact the evidence for mice and birds tends to suggest animals exposed to Mn are less active than control animals. This suggests that Mn exposure has an adverse effect on all species, but for rats its effect is to increase activity levels, while for humans (Table 6.4), mice and birds (Table 6.6) activity levels tend to decrease with exposure to Mn.

The pooled ln(ORs) from experiments on monkeys suggests that exposed monkeys are over 7 times more likely (95%CI: 1.05, 51.42) to be observed exhibiting an adverse activity symptom than unexposed monkeys. The uncertainty associated with this estimate is clear from the very wide CI, yet it does not include the null. This finding must be interpreted with care as the adverse symptom refers to extremes of behaviour; it does not refer to increased or decreased activity levels unlike the rest of the animal and human evidence. Thus, general patterns in activity levels from exposure to Mn can be compared between species but, because of differences in the measuring and reporting of activity levels, such comparisons are limited.

Table 6.5 Available animal evidence

Study	Year	Species	Symptom	No. exposed	No. controls	Standardised mean difference in activity score		Ln(OR)	
						Estimate	Variance	Estimate	Variance
Chandra <i>et al.</i>	1981	Rat	Activity counts per 15 min - day 7	6	6	-2.21	0.54		
Murthy <i>et al.</i>	1981	Rat	Activity - 21% casein diet	6	6	-13.03	7.41		
Murthy <i>et al.</i>	1981	Rat	10% casein diet	6	6	-11.50	5.85		
Bonilla	1984	Rat	No. movements	8	8	4.67	0.93		
Nachtman <i>et al.</i>	1986	Rat	last time measurement	12	12	-1.42	0.21		
Pappas <i>et al.</i>	1997	Rat	Activity levels	10	10	-5.24	0.89		
Gray & Laskey	1980	Mouse	last time measurement	6	6	2.15	0.53		
Lown <i>et al.</i>	1984	Mouse		45	38	0.07	0.05		
Komura & Sakamoto	1991	Mouse	% rate of activity – day 105	8	8	0.97	0.28		
Laskey & Edens	1985	Bird		11	15	3.82	0.44		
Suzuki <i>et al.</i>	1975	Monkey	Hyperexcitability	6	2			2.20	3.02
Eriksson <i>et al.</i>	1987	Monkey	Hyperactive, then hypoactive	4	2			3.81	4.62
Shinotoh <i>et al.</i>	1995	Monkey	Hypoactive	3	3			1.02	2.13

Table 6.6 Pooled fixed and random effects estimates for activity levels from the relevant animal experiments

Activity outcome	Species	No. studies in analysis	Standardised mean difference between activity score*			Ln(OR) for reporting a symptom		
			Fixed	Random	I ²	Fixed	Random	I ²
Counts	Rat	6	-1.69 (-2.34, -1.04)	-4.02 (-7.30, -0.74)	95 (91, 97)	.	.	.
	Mouse	3	0.35 (-0.04, 0.74)	0.91 (-0.24, 2.05)	78 (29, 93)	.	.	.
	Bird	1	3.82 (2.52, 5.12)
Symptom	Monkey	3	.	.	.	2.00 (0.05, 3.94)	2.00 (0.05, 3.94)	0 (0, 90)

* a positive value indicates less activity in the exposed group

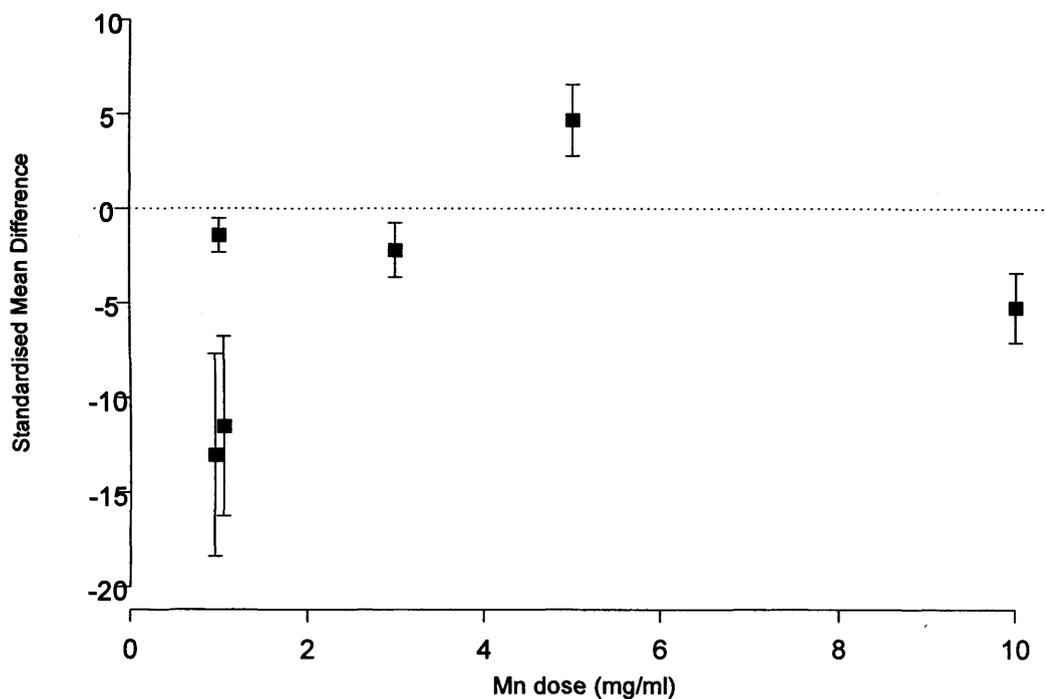
This is also an issue when exposure data are considered, since it is measured and reported in diverse ways. For example a number of different routes of exposure are used, as are different periods of exposure (ranging from four days to two years depending on the species and route of exposure) and frequency of dosing. Meta-regression techniques may be useful in exploring whether an effect exists at different dose levels and Figure 6.1 shows how this might be achieved by plotting the standardised mean differences from the six rat studies against the reported dose level of exposed rats (daily dose of Mn, in mg/ml of drinking water). It may be tempting to suggest from this plot that low and high levels of manganese in drinking water are associated with increased levels of activity in rats. However, there are a number of dangers in the assessment and interpretation of Figure 6.1. Firstly, there are only six experiments and the measures of exposure are so diverse that one should be cautious of over-interpretation of this plot. Perhaps if data from a larger number of experiments were available and the exposure data were less diverse, this plot may be more useful in identifying an exposure limit for Mn. Secondly, and more importantly, the use of subject-related averages in a meta-regression (such as average Mn exposure) introduces the problem raised earlier of ecological fallacy. This is where relationships between subject-related averages and outcomes are not the same as the relationship between exposure and outcome in the individual (Thompson and Higgins, 2002). The only way to overcome this problem is by using the individual subject data and modelling the individual exposures and outcomes.

6.5.3 Summary of species-specific evidence

Although the differences between the sources of evidence which must be taken into account in a risk assessment are great, use of systematic reviews and meta-analysis methods in this example, so far, have assisted in making it easier to evaluate and understand the available human and animal data by combining similar types of evidence. There are now eight pieces of evidence to consider regarding activity levels rather than data from 27 studies. These data are presented in Table 6.7. The pooled standardised mean differences in activity levels are displayed graphically in Figure 6.2 to help assess consistency of effects across species (as advocated by Roberts *et al.* (2002a)). Thus, systematic review and meta-analysis methods have a

useful role in summarising the evidence, regardless of whether a cross design synthesis is undertaken.

Figure 6.1 A plot of the standardised mean differences in activity score from six rat experiments against Mn dose



NB negative standardised mean difference suggests increased activity levels in rats exposed to Mn

In fact, in this example, there are a number of possible reasons why one may decide not to carry out a cross-species synthesis. For instance, whereas the evidence from the human studies is based on self-assessment regarding feelings of fatigue and vigour, the animal evidence is from experiments measuring the amount of activity animals display. These two types of assessment therefore may not be considered comparative and synthesis across species may produce misleading and uninformative findings. Furthermore, the evidence in this example suggests that humans, mice and birds experience a decrease in activity level with Mn exposure while rats increase their activity (see Table 6.7 and Figure 6.2). Both of these effects can be termed adverse, but a synthesis of all these results, as they are, is likely to result in a dilution of this effect (i.e. provide a standardised mean

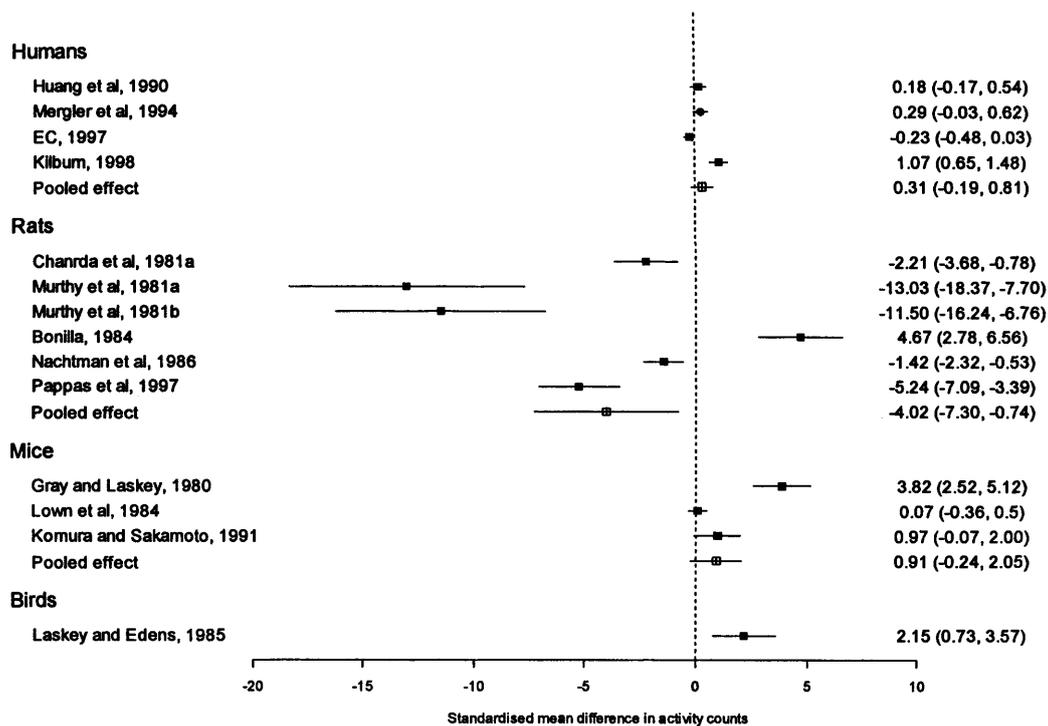
difference close to zero). Methods to account for this could be carried out, such as taking the modulus of each study estimate; however the overall estimate may be difficult to interpret.

Table 6.7 Random effects pooled evidence on activity levels and exposure to Mn

Activity level outcome	Species	Number of studies	Effect measure	Summary of effect
Fatigue	Humans	4	Standardised mean difference	-0.63 (-1.00, -0.26)
	Humans	8	Ln(OR)	0.30 (-0.09, 0.69)
	Humans	11	Ln(odds)	-0.93 (-1.52, -0.33)
Vigour	Humans	4	Standardised mean difference	0.31 (-0.19, 0.81)
Activity counts	Rats	6	Standardised mean difference	-4.02 (-7.30, -0.74)
	Mice	3	Standardised mean difference	0.91 (-0.24, 2.05)
	Birds	1	Standardised mean difference	3.82 (2.52, 5.12)
Negative activity symptom	Monkeys	3	Ln(OR)	2.00 (0.05, 3.94)

Nevertheless a synthesis of the available evidence taking account of the different species and study types can be achieved using the meta-analysis models defined in Chapter 5 (Equations 5.2 – 5.4). To illustrate the use of these models, and some possible extensions, they are applied to the human and rat evidence presented in Figure 6.2. It should be noted that because of the inconsistency between the results from the human and rat data one would not usually combine the evidence, however for illustrative purposes only, it is carried out here. The models and results are described in the following section.

Figure 6.2 Standardised mean differences in activity scores from four human epidemiology studies and ten animal experiments



6.6 Further application of meta-analysis methods

6.6.1 Synthesis across species

A naïve synthesis of the data from the four human studies and six rat experiments may assume that all 10 studies are estimating the same effect, so that there are no differences between species, routes of exposure or the slightly different health outcomes. This assumption clearly does not hold here, but such a synthesis is provided for a comparison with subsequent analyses. There is some evidence of heterogeneity between the outcomes from the 4 human studies and 6 rat experiments (Figure 6.2) and so a random effects model is used to synthesise the data. A random effects meta-analysis model, defined in Equation 3.2, is used here under a Bayesian framework (see Equation 6.5)

$$\begin{aligned}
 d_i &\sim N(\theta_i, \sigma_i^2) & \mu &\sim N(0, 10^9) \\
 \theta_i &\sim N(\mu, \tau^2) & 1/\tau^2 &\sim \text{Gamma}(0.001, 0.001)
 \end{aligned}
 \tag{6.5}$$

where d_i is the standardised mean difference in scores from study i ($i: 1, \dots, 10$), σ_i^2 is the variance of d_i , calculated from Equation 6.3[†]. θ_i is the true standardised mean difference in study i , μ is the pooled estimate and τ^2 is the estimate of between study variance. Results are compared with the pooled estimate from a classical random effects meta-analysis model.

The following models allow importance of the animal evidence in assessing human health risks from Mn exposure to be modelled.

In the first of these models, a prior, based on the rat data, is placed on the synthesis of the human evidence. The Bayesian model applied here is that defined in Section 5.5.1 (Model 1c), where the rat data are taken at ‘face value’ and the combined estimate is used as a prior for the synthesis of the human data

$$\begin{aligned}
 d_i &\sim N(\theta_i, \sigma_i^2) & \mu &\sim N(\mu_{rat}, \nu_{rat}^2) \\
 \theta_i &\sim N(\mu, \tau^2) & 1/\tau^2 &\sim \text{Gamma}(0.001, 0.001)
 \end{aligned}
 \tag{6.6}$$

where d_i is the standardised mean difference in scores from study i ($i: 1, \dots, 4$), σ_i^2 is the variance of d_i . θ_i is the true standardised mean difference in study i , μ is the pooled estimate, τ^2 is the estimate of between study variance, μ_{rat} is the combined estimate from the six rat experiments and ν_{rat}^2 is the variance of this estimate.

[†] Abrams *et al.* (2005) report that a correlation may be induced between d_i and its variance when calculated by Equation 6.3 since the variance is a function of d_i . Thus a simplification of Equation 6.3 (as given on page 31 of Sutton *et al.* (2000)) may be more appropriate to avoid this correlation.

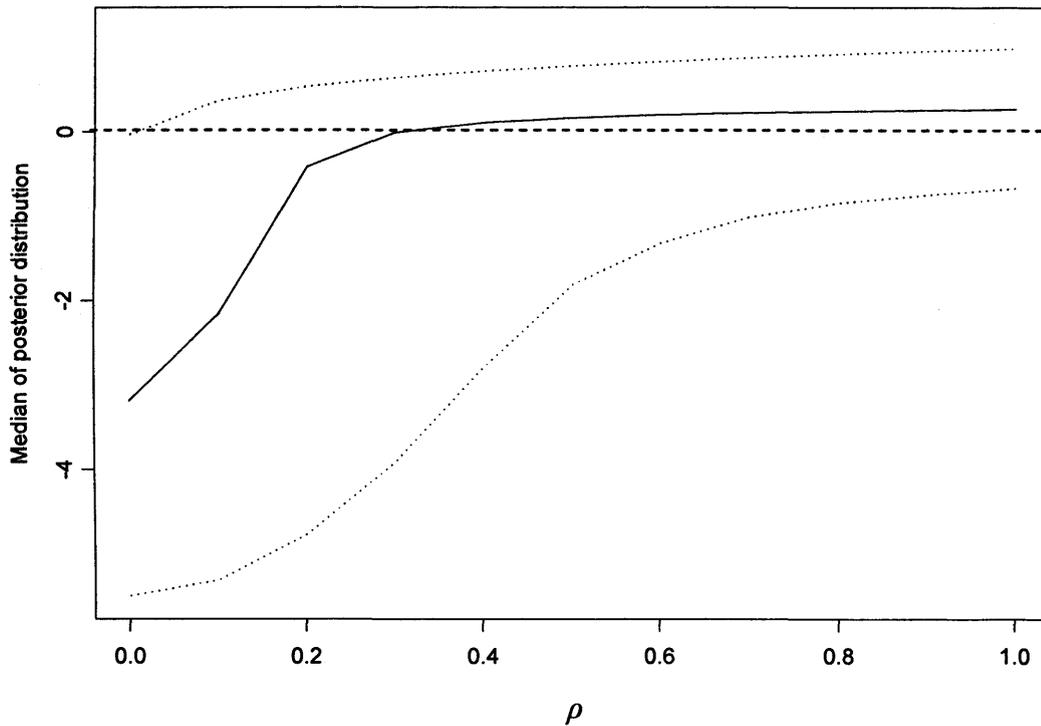
In the following models different levels of relevance of the rat data to the human data are incorporated. This is achieved through the use of power transform priors (Ibrahim and Chen, 2000). The prior distribution for the human evidence, here the rat data, is raised to the power ρ , where $0 \leq \rho \leq 1$ ($\rho = 1$ is equivalent to the informative prior analysis of Equation 6.6 above)

$$\begin{aligned}
 d_{ik} &\sim N(\theta_{ik}, \sigma_{ik}^2) & \mu_k &\sim N(\mu_{rat}, \nu_k^2) \\
 \theta_{ik} &\sim N(\mu_k, \tau_k^2) & 1/\tau_k^2 &\sim \text{Gamma}(0.001, 0.001) \\
 & & \nu_k^2 &= \nu_{rat}^{2\rho}
 \end{aligned} \tag{6.7}$$

where d_{ik} is the standardised mean difference in study i ($i:1, \dots, 4$ and $k:1, \dots, 11$ for differing values of ρ). σ_{ik}^2 is the estimated variance of d_{ik} , θ_{ik} is the true standardised mean difference in study i , μ_k is the combined estimate and τ_k^2 is the between study estimate of variance. μ_{rat} is the combined estimate from the six rat studies and ν_{rat}^{2a} is the variance of this estimate, raised to the power ρ . The median estimates and 95% CrI for the pooled human effect, μ_k for the 11 different values of ρ , are given in Figure 6.3.

From Figure 6.3 one can see that as the rat data are given more weight, i.e. as $\rho \rightarrow 1$, the pooled estimate from the human evidence increases in precision. In Figure 6.3 the dilution of adverse effects, seen in humans and rats, resulting from the cross-species synthesis can be seen. Nevertheless, this is a useful model to assess the sensitivity of the pooled estimate to differing levels of perceived relevance of the rat data, although it is difficult to determine how to measure relevance.

Figure 6.3 Plot of pooled human effect μ_k (95% CrI) with changing values of ρ (relevance weighting for the rat data) [$\rho=0$ implies rat data are totally discounted; $\rho=1$ implies rat data are included at face value]



In the following, perhaps more realistic scenario, relevance is not measured, instead judgements are made on some evidence being more relevant than other evidence.

The three-level hierarchical model described in Equation 5.4 is used here

$$\begin{aligned}
 d_{ij} &\sim N(\psi_{ij}, \sigma_{ij}^2) & \mu &\sim N(0, 10^9) \\
 \psi_{ij} &\sim N(\theta_j, \tau_j^2) & 1/\tau_j^2 &\sim \text{Gamma}(0.001, 0.001) \\
 \theta_j &\sim N(\mu, \nu^2) & 1/\nu^2 &\sim \text{Gamma}(0.001, 0.001)
 \end{aligned} \tag{6.8}$$

where d_{ij} is the standardised mean difference in study i ($i:1, \dots, 10$), species j ($j=1, \text{humans}, j=2, \text{rats}$), σ_{ij}^2 is the (estimated) variance of d_{ij} and ψ_{ij} is the true standardised mean difference in study i , species j . θ_j is the combined estimate of studies using species j , τ_j^2 is the estimate of variance between studies on species j ,

μ is the overall combined estimate and ν^2 is the estimate of variance between studies of different species.

Two versions of this model are applied to the human and rat Mn and activity level evidence. The first is as described above: the *unconstrained analysis*. In the second version of this model the following constraint, used by Prevost *et al.* (2000), is made to model the order of relevance of one species over another: the *constrained analysis*. In Equation 6.9, θ_1 is the pooled human estimate, and θ_2 is the pooled rat evidence, so that the human evidence are forced to be less biased than the animal evidence, which one is likely to believe given it is the human effect that is ultimately of interest.

$$|\mu - \theta_1| < |\mu - \theta_2| \quad (6.9)$$

The pooled estimates from the different models described here, except the power prior model, are given in Figure 6.4, so that their results can be compared.

Figure 6.4 Pooled standardised mean difference effects from the human studies and rat experiments

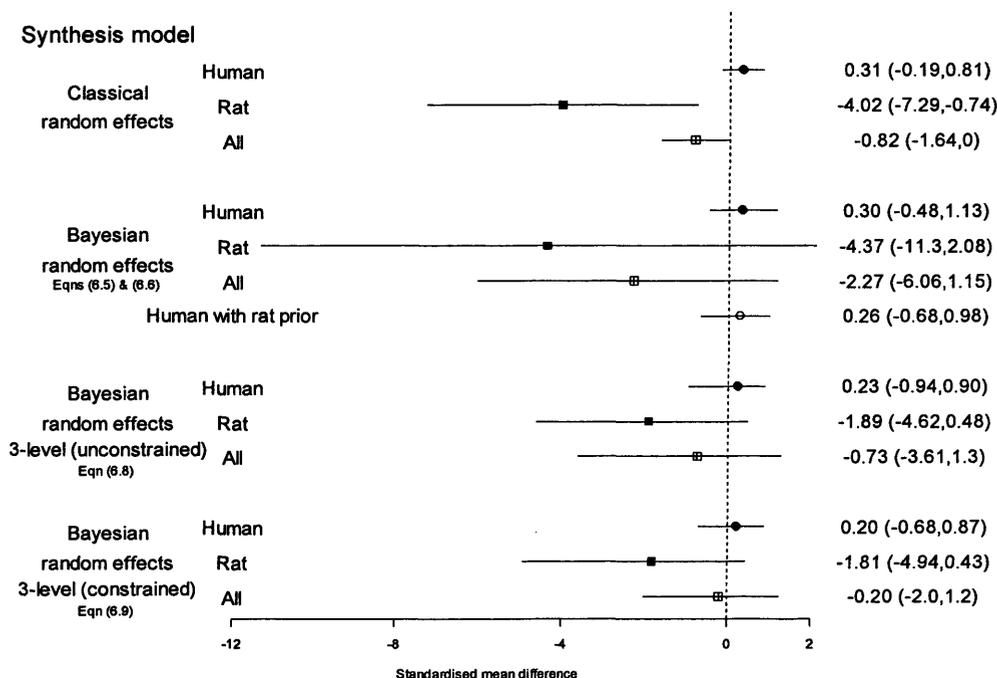


Figure 6.4 highlights findings from Table 6.7 and Figure 6.2 that the human evidence is much more precise than the rat evidence and that Mn exposure affects humans and rats differently according to the data used here. A naïve synthesis of the evidence in a classical random effects model results in a reasonably precise pooled estimate suggesting a significant increase in activity levels for those exposed to Mn. The naïve analysis using a Bayesian model is less precise, as one would expect from a Bayesian model but still suggests an increase in activity level with Mn exposure. The result of using rat data as a prior for the synthesis of the human data is very similar to if the prior data were not used at all. This is because the evidence from the rat experiments has a large variance and so the prior distribution is diffuse and has little impact on the synthesis of the human evidence. Use of the 3-level model gives pooled estimates of effect across rats and humans that lies between the naïve synthesis pooled estimate and that from use of an informative prior distribution to synthesise the human evidence. Comparing the estimates from the two 3-level models (unconstrained and constrained) shows that θ_1 , the estimate from the human evidence, and θ_2 , the estimate from the rat evidence, change very little. However, the overall pooled estimate of μ from the constrained model has been pulled towards the estimate of the human evidence, θ_1 , as would be expected.

As briefly discussed in Chapter 5, it may be more appropriate for estimates of θ_1 , the human effects, to be reported and interpreted, rather than estimates of μ , the overall species effect. Since human health effects are of interest here, it appears reasonable to use θ_1 . However, choice of θ_1 over μ is not as clear when different types of human evidence are being synthesised, for example, as in Prevost *et al.* (2000) and Sutton and Abrams (2000). In both of these papers evidence from human RCTs are combined with evidence from observational studies on an intervention, and it will no doubt be due to the main question of interest: the *effectiveness* or the *efficacy* of the intervention.

The application of these methods to the human and rat Mn and activity level data is illustrative only. In the following section, illustration of the advantages of a meta-analysis framework in a risk assessment context is extended to demonstrate possible sensitivity analyses.

6.6.2 Sensitivity analyses

Having a quantitative framework for the evaluation of evidence for a risk assessment allows investigation of the sensitivity of some of the decisions and assumptions made in the risk assessment process. In this section, the sensitivity of conclusions are investigated by assessing the impact of a number of studies on the meta-analyses carried out in Section 6.5.1 and more explorative analyses common in meta-analyses are described (between-study heterogeneity and publication bias).

Study impact

A common sensitivity analysis in meta-analysis is to investigate whether inclusion or exclusion of certain studies has a significant impact on the conclusions. For example, if, in earlier stages of the systematic review process, there has been difficulty assessing the quality, or the inclusion/exclusion criteria for one or more studies, an investigation as to the impact of including or excluding a study can be carried out. To demonstrate this Table 6.8 shows the pooled rat estimate of the standardised mean difference in activity when data from Chandra *et al.* (1981) are included (as in the initial analyses), and excluded, from a classical random effects meta-analysis.

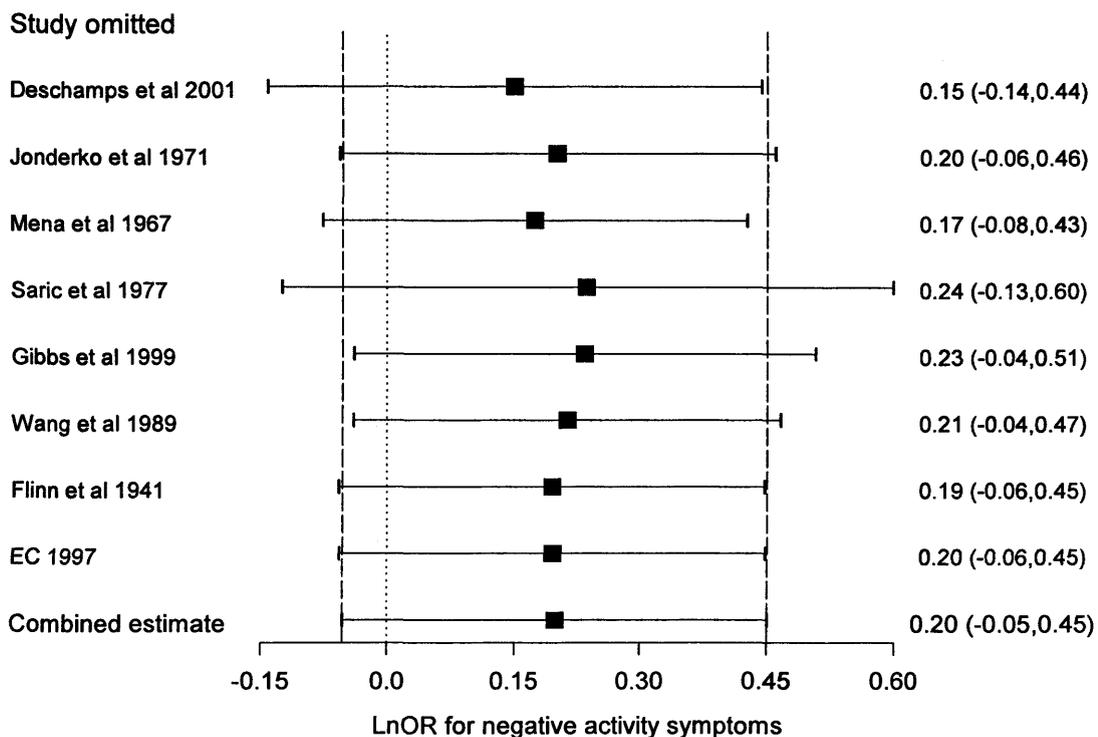
Table 6.8 Impact of Chandra *et al.* (1981) data on the pooled rat standardised mean difference estimate

	Pooled estimate (95% CI)
Initial analysis (Chandra <i>et al.</i> data included)	-4.02 (-7.29, -0.74)
Sensitivity analysis (Chandra <i>et al.</i> data excluded)	-4.70 (-9.15, -0.25)

Inclusion of data from Chandra *et al.* decreases the pooled standardised mean difference estimate; however, the overall conclusion is the same, i.e. the exposed rats have higher activity counts than unexposed rats. Moreover inclusion of the Chandra *et al.* data provides a more precise estimate than that obtained when these data are excluded.

The impact of each study in the meta-analysis can easily be assessed to determine whether one particular study is greatly influencing the pooled estimate. This is illustrated using the available epidemiological data on the $\ln(\text{OR})$ of reporting a negative activity symptom. Each study is systematically excluded from the meta-analysis and the remaining data re-analysed. Figure 6.5 shows the combined estimate from a classical random effects meta-analysis when each study is excluded.

Figure 6.5 Impact of each study on the pooled estimate



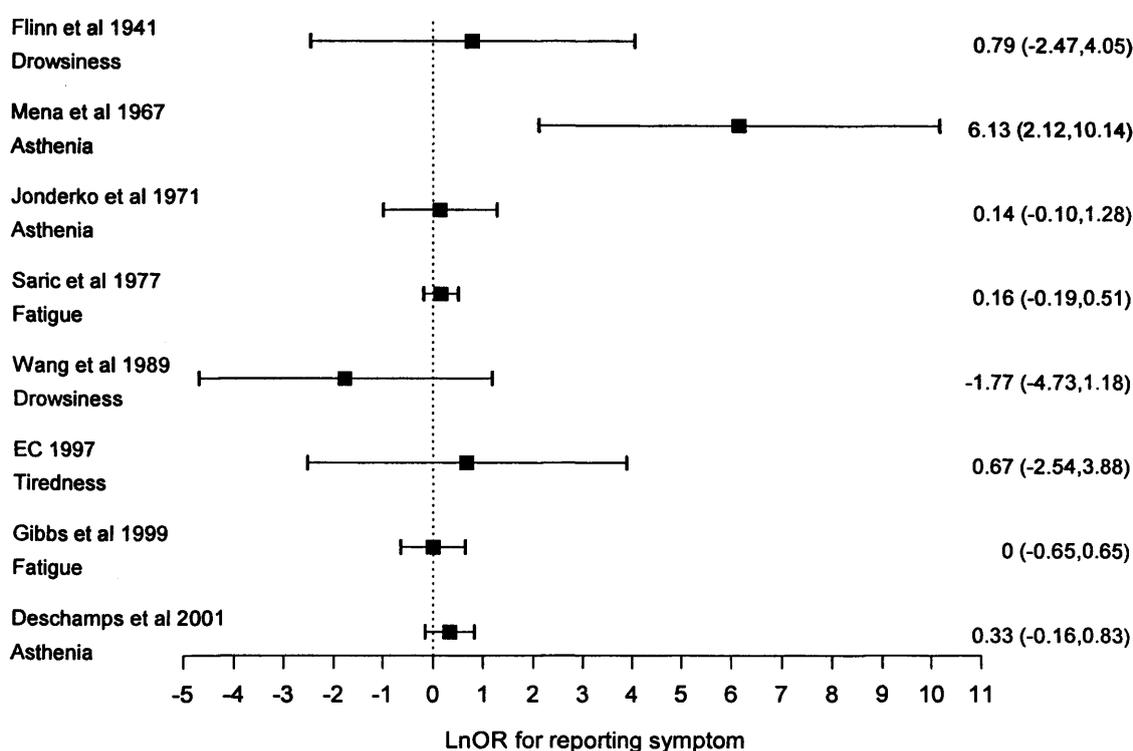
No one study has a huge influence on the combined estimate. However attention is drawn to the Saric *et al.* (1977) study, which when excluded, leaves a combined estimate with a larger 95% CI suggesting the Saric *et al.* study is quite precise.

Meta-regression and between-study heterogeneity

In Section 3.3.2 meta-regression techniques were described. These techniques can be used to explain heterogeneity between studies. For instance, year of publication may explain some heterogeneity between the eight epidemiology studies where a $\ln(\text{OR})$ for reporting a fatigue-like symptom was calculated. Although year of

publication itself is unlikely to explain heterogeneity, it may be a surrogate for other, unmeasured characteristics, such as a change in occupational practice affecting both the exposure and health outcome. Figure 6.6 is a plot of the eight studies in chronological order.

Figure 6.6 Plot of the $\ln(\text{OR})$ for reporting a fatigue-like symptom from eight epidemiology studies by year of publication



It is difficult to see if there is an association between $\ln(\text{OR})$ for reporting a fatigue-like symptom and year of publication. Using meta-regression techniques year of publication is regressed on to $\ln(\text{ORs})$. Results suggest there is little evidence of a linear relationship between the $\ln(\text{OR})$ and year of publication (estimated year of publication regression coefficient -0.002; 95% CI -0.022, 0.019). However meta-regression should be looked upon as an explanatory analysis and should therefore not be over-interpreted (Sutton *et al.*, 2000).

Publication bias

A further aspect of meta-analysis that should be considered is the possibility of publication bias, as described in Section 3.3.3. This relates to the idea that studies concluding significant results are more likely to be published. If data are subject to publication bias, the evidence available for the basis of the exposure limit may be biased, resulting in the possibility of a biased estimate for the exposure limit.

Hence, it is necessary for a thorough search of the literature to be carried out, including areas of the grey literature. However, if a study has been conducted, but not reported anywhere, clearly it cannot be identified. It is therefore important to assess whether publication bias is an issue for a particular set of evidence and what impact this may have on any analyses. There has been very little investigation of publication bias of animal experiments as shown in Section 4.4.3.

In many meta-analyses, the number of studies being combined is quite low, and so a funnel plot is often difficult to interpret. Figure 6.7 is a funnel plot of the 14 studies investigating Mn exposure and activity levels, illustrating quite clearly that the human studies are much more precise (a higher value for $1/\text{standard error}$) than the animal experiments. There is some evidence to suggest that small studies concluding a decrease in activity levels in Mn exposed groups compared to unexposed groups (a positive standardised mean difference) are missing (bottom right-hand corner of Figure 6.7). Although findings from the often used rank correlation and regression tests to detect publication bias (Begg and Mazumdar, 1994; Egger *et al.*, 1997) suggest there is little evidence of publication bias ($p=0.555$ and $p=0.511$, respectively). However, evidence from four different species is being assessed here and as Table 6.7 shows, a species effect is evident. Thus an assumption of the use of the funnel plot, and subsequent tests based on the funnel plot, is not met: that all studies should come from the same underlying distribution (Light and Pillemer, 1984). As is further discussed in Section 8.2.2 of Chapter 8, assessment of publication bias within species is one way to proceed, but this leads to fewer studies to be assessed, reducing the statistical power of these methods. However, as noted by a number of authors and demonstrated in Chapter 7, the power of methods to detect publication bias is poor as the number of studies in the meta-analysis decreases (Begg and Mazumdar, 1994; Sterne *et al.*, 2000; Macaskill *et al.*, 2001). In fact, there is evidence to suggest that current commonly

used methods for the detection and adjustment of publication bias are not ideal (Begg and Mazumdar, 1994; Sterne *et al.*, 2000; Macaskill *et al.*, 2001; Schwarzer *et al.*, 2002; Terrin *et al.*, 2003).

In human health risk assessments, species effects are likely to be observed and so investigation of how to appropriately assess publication bias in the presence of species effect, or more generally between-study heterogeneity, is required. There has been some suggestion that publication bias is more of an issue for epidemiology studies than for randomised controlled trials as some epidemiology studies may have been done in a more explorative manner, and have therefore been selectively published (Blettner *et al.* 1999). However the presence and impact of publication bias of animal experiments has not been investigated. In Chapters 7 and 8 assessment and adjustment for publication bias in meta-analyses of human studies and animal experiments is explored.

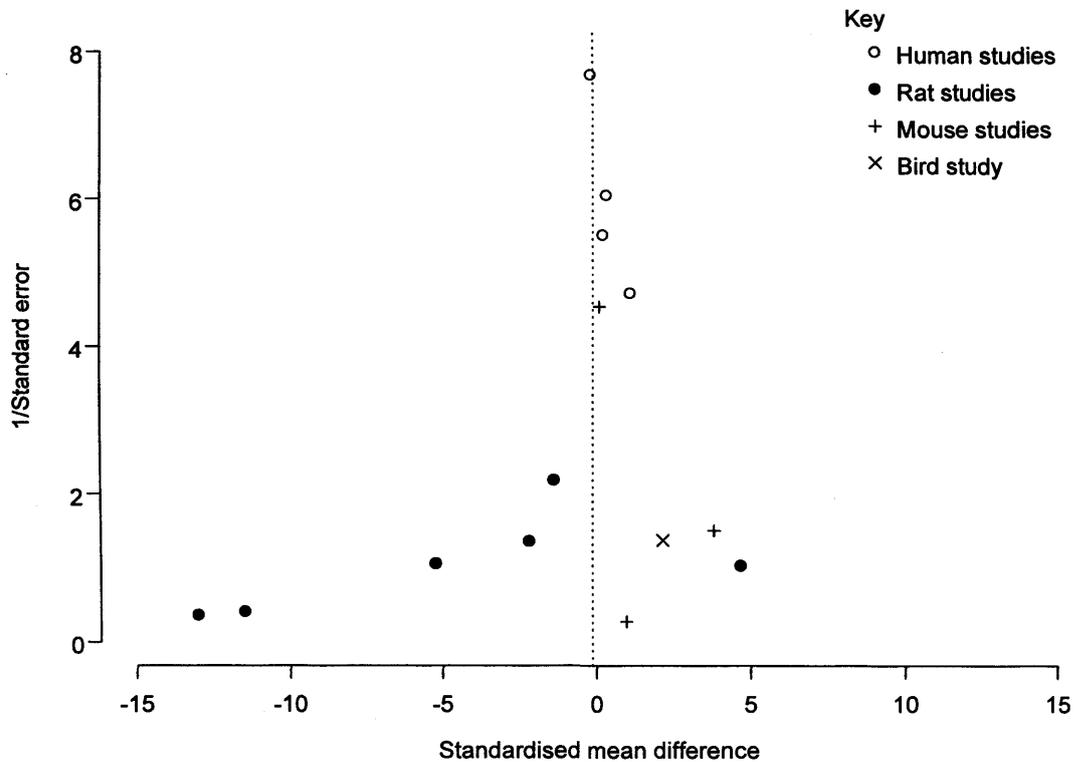
6.7 Summary

In this chapter a second, quite different example to that used in Chapter 5, has been described: neurobehavioural effects from occupational exposure to Mn. A systematic review of the evidence for such a risk assessment was carried out. From this systematic review, the relevant evidence from humans and animals was found to be quite diverse in terms of the population used, the exposure and the measurement of an adverse effect. Due to this diversity, an evaluation of the evidence is quite complex, however systematic review and meta-analysis methods have allowed for a transparent and structured summary of the relevant evidence. Synthesis of evidence *across species* has been illustrated in this chapter (Section 6.6.1), but the diversity of the evidence means that such summary estimates may be difficult to interpret. Synthesis of evidence *within species* has allowed straightforward examination of the consistency of estimated effects across the different species (e.g. Table 6.7), however this has still been limited by the diversity in the relevant experiments and studies, and by their reporting. Such an examination of consistency across species using common risk assessment methods may not be so straightforward. Synthesising similar study results means that

decisions regarding the pivotal study (as used in general risk assessment approaches, see Chapter 2) are redundant, leading to more efficient use of the evidence. Of course, decisions on the quality and relevance of studies are still needed when systematic review and meta-analysis methods are employed, but this can be achieved in a more transparent manner. For example, relevance and quality can be set out in the inclusion and exclusion criteria.

By attempting to take a formal approach to the review and evaluation of evidence regarding the human neurobehavioural health effects from occupational exposure to Mn, the many problems of reviewing such evidence have been highlighted and described in this chapter. The aim has not been to develop a synthesis model which can incorporate and account for all sources of evidence and uncertainty that would replace current risk assessment approaches. Instead, systematic reviews and meta-analysis methods have been applied to this example in order to demonstrate where and how these methods may be more advantageous than current methods of risk assessment, and to point out that regardless of the approach taken, many issues remain in how the human and animal evidence can be used to inform human health risk assessments. For instance, one particular problem highlighted in this chapter has been the varied reporting of the measured exposures within and across different study/experiment types. Regardless of the approach taken to risk assessment, this is an important limiting factor of the review and evaluation process. Identification of this highlights areas where reporting of individual studies and experiments need to be improved, and suggests that some form of harmonization would be advantageous. Moreover, this systematic approach to identifying and evaluating the evidence potentially relevant to a risk assessment has illustrated areas where evidence are available that may help improve current practice. For example, although different types of experiments used different routes of exposure (oral, inhalation and injection – see Table 6.1), evidence from two animal strains (rhesus monkeys and Sprague-dawley rats) covered all three exposure routes. This allows for an examination of findings across exposure routes for these animals that may potentially be useful when considering different strains or species.

Figure 6.7 Funnel plot of 14 studies investigating exposure to manganese and differences in activity levels between exposed and unexposed subjects



Evidence concerning just one of the many neurobehavioural outcomes has been synthesised in this chapter (i.e. activity level). To form the basis of a human health risk assessment, meta-analysis methods need to be applied to all relevant outcomes to help summarise the evidence and inform decision-makers as to the form of the exposure limit required for occupational exposure to Mn. As demonstrated in this chapter and in Chapter 5, such an approach has many attractive advantages over the more narrative methods of risk assessment described in Chapter 2.

Performance of tests and adjustments for publication bias

7.1 Chapter overview

Regardless of the application, possible publication bias and between-study heterogeneity are important features of a meta-analysis that must be considered. In this chapter, methods for the detection of, and adjustment for, potential publication bias are assessed in terms of their performance under a number of simulated meta-analysis scenarios. These include the rank correlation test (Begg and Mazumdar, 1994), a number of regression tests (Egger *et al.*, 1997; Macaskill *et al.*, 2001) and the trim and fill method (Duval and Tweedie, 2000a; Duval and Tweedie, 2000b). Publication bias and its potential consequences are introduced in Section 7.2. In Section 7.3 these tests and methods and their use in practice are described. An overview of the performance of the rank correlation test, the regression test and the trim and fill methods is presented in Section 7.4 based on the findings of six published simulation studies. Results from these studies do not fully answer questions on the performance of the tests and methods in certain scenarios and so, in Section 7.5 details of new simulation analyses are given. The characteristics of the meta-analyses and the tests and methods investigated in the simulations carried out in this chapter are defined, in addition to the analyses undertaken. The results of these simulation analyses are presented in Section 7.6 followed by some sensitivity analyses of the parameters defined in the simulated meta-analyses (Section 7.7). The results of the simulations and further analyses are discussed in Section 7.8. The simulations and results presented in this chapter include and go beyond those reported in Peters *et al.* (2006) (given in Appendix H).

7.2 Introduction

Publication bias describes the tendency for smaller studies reporting a non-significant or unfavourable effect, to be less widely disseminated as reports of larger studies concluding a significant and/or favourable effect (Easterbrook *et al.*, 1991; Scher *et al.*, 1994; Stern and Simes, 1997; Ioannidis, 1998; Song *et al.*, 2000). A number of processes may lead to publication bias. They include authors being less likely to write-up and submit studies with unfavourable or uninteresting results and editors of prominent journals (where space is limited) being less likely to accept and publish such papers. Evidence also suggests that larger, significant, more interesting studies are more likely to be submitted to English language journals, regardless of the first language of the authors, but the evidence regarding the impact of this on a systematic review or meta-analysis is mixed (Egger *et al.*, 1999; Song *et al.*, 2000; Egger *et al.*, 2003). Outcome reporting bias may also affect the findings of a systematic review or meta-analysis (Song *et al.*, 2000). This occurs when researchers only publish findings for outcomes where a significant effect was observed, regardless of the number of outcomes actually analysed. Much research has, and is currently, being done to investigate the effects of, and adjustments for, outcome reporting bias (Hahn *et al.*, 2002; Chan *et al.*, 2004; Chan and Altman, 2005; Williamson *et al.*, 2005; Williamson and Gamble, 2005).

The consequences of ignoring the possibility of publication bias may be quite severe. If publication bias is present, the subsequent systematic review and/or meta-analysis will then be based on a biased set of evidence which may in turn lead to misleading results (often an inflation of the true effect) and conclusions, affecting decision-making and policy.

When conducting a systematic review the most appropriate way to limit the likelihood of publication bias is to comprehensively search the literature. A review should not just concentrate on studies obtained from a search of electronic databases (such as Medline and Embase), but should involve searching numerous sources of evidence, e.g. the reference lists of relevant studies, conference proceedings, grey literature, internet, hand-searching of journals, contacting researchers or

manufacturers for any unpublished material (Sutton *et al.*, 2000; Deeks *et al.*, 2001; Egger *et al.*, 2001; Cochrane Collaboration, 2005).

However, even a comprehensive systematic review of the literature will not be able to identify those studies that have not even been written up, and so a meta-analysis may still be subject to publication bias, regardless of how thorough the systematic review. It is therefore important that all meta-analyses are investigated for the possible presence of publication bias. A number of methods have been developed to assist in determining whether a review is affected by publication bias. These include the funnel plot (Light and Pillemar, 1984), a rank correlation test (Begg and Mazumdar, 1994), a regression test (Egger *et al.*, 1997) and a non-parametric method called 'trim and fill' (Duval and Tweedie, 2000a; Duval and Tweedie, 2000b). In the review of the application of systematic reviews and meta-analyses to animal experiments (Chapter 3) it was noted that 17 out of the 46 meta-analyses (37%) mentioned the issue of publication bias. Of these meta-analyses, only six report carrying out some assessment. A number of authors gave reasons for not assessing possible publication bias in their meta-analysis, including the comment that "available techniques lack validity" (Kelley, 1996). In this chapter the performance of the rank correlation test, Egger's regression test and a number of alternative regression tests in addition to the trim and fill method is investigated using simulation analyses. The most commonly used methods to investigate publication bias are now described.

7.3 Methods for detecting publication bias

The simplest method is the funnel plot (Light and Pillemar, 1984). This is a scatter plot of the estimate of effect, e.g. the natural logarithm of the odds ratio ($\ln(\text{OR})$), against a measure of precision, usually $1/(\text{standard error of } \ln(\text{OR}))$, for each study in the meta-analysis. There has however been some discussion as to what form the axes should take (e.g. precision vs. sample size (Sterne and Egger, 2001)) and this will be returned to in Section 7.8. On the assumption that smaller, less precise studies are more subject to random variation than the more precise larger studies when estimating an effect, the scatter plot should resemble a funnel in the absence

of publication bias, with the highly precise studies at the top of the funnel and the smaller, less precise studies varying around the true effect (Figure 7.1). In the presence of publication bias, smaller studies reporting less favourable or non-significant effects are likely to be missing from the meta-analysis, resulting in an asymmetric plot (see Figure 7.2).

It is important to note that publication bias is only one reason why asymmetry may be observed in a funnel plot. Alternative reasons for this asymmetry include poorer methodological quality of smaller studies, between-study heterogeneity, chance and inadequate analyses (Sterne *et al.*, 2000; Ioannidis, 2005). For instance, Ioannidis (2005) discusses the possibility that level or quality of treatment received by an individual depends upon the study size, thus the perceived effectiveness of the treatment differs by size of study. For example, large multicentre trials require many centres to be involved, many of which will be less experienced compared to those few involved in smaller studies (Ioannidis, 2005).

Although very simple to use, the funnel plot is a subjective means for detecting possible publication bias. Statistical tests have therefore been developed to provide a more formal assessment. The following two tests (Begg and Mazumdar, 1994; Egger *et al.*, 1997) are based on the idea that when publication bias is present, an association can be observed between effect size and precision; so that the more extreme effect sizes will have lower precision (as is observed in a funnel plot).

Figure 7.1 Funnel plot of a meta-analysis in the absence of publication bias

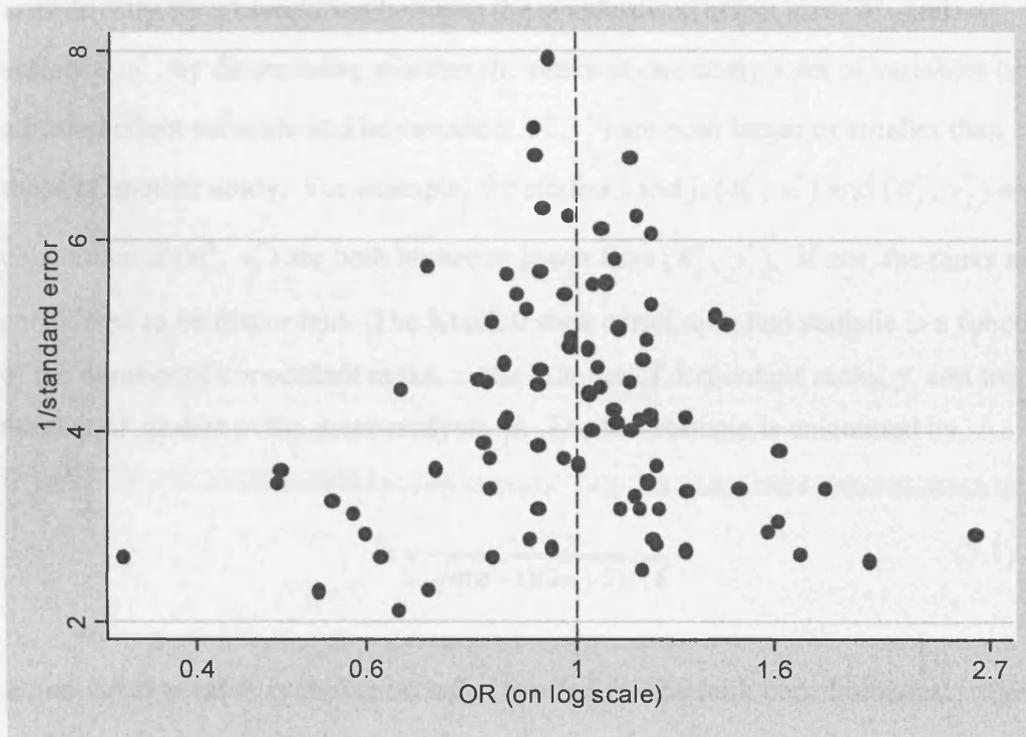
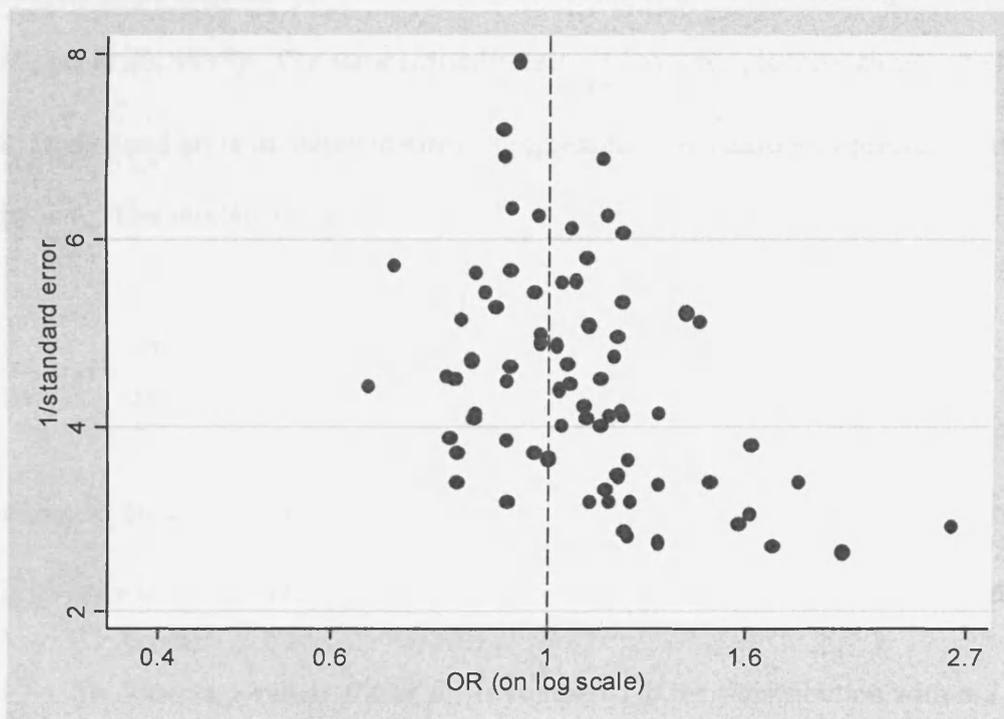


Figure 7.2 Funnel plot of a meta-analysis in the presence of publication bias



The rank correlation test is based on Kendall's tau (Begg and Mazumdar, 1994). It tests directly for a correlation between the standardised effect size, θ^* , and variance, v^* , by determining whether the ranks of one study's set of variables (the adjusted effect estimate and its variance: θ^*, v^*) are both larger or smaller than those of another study. For example, for studies i and j , (θ_i^*, v_i^*) and (θ_j^*, v_j^*) are concordant if (θ_i^*, v_i^*) are both higher or lower than (θ_j^*, v_j^*) . If not, the ranks are considered to be discordant. The Kendall rank correlation test statistic is a function of the number of concordant ranks, x , the number of discordant ranks, y , and the number of studies in the meta-analysis, n . The test statistic is calculated by

$$z = \frac{x - y}{\sqrt{n(n-1)(2n+5)/18}} \quad (7.1)$$

A two-sided p-value is conventionally reported for the rank correlation test, where evidence of a correlation between the ranks therefore suggests evidence of funnel plot asymmetry.

The second test commonly used to help detect publication bias is a regression test (Egger *et al.*, 1997). The standardised effect $\frac{y_i}{se_i}$, where y_i is the estimate of effect in study i and se_i is its standard error, is regressed on a measure of precision given by $\frac{1}{se_i}$. The model is

$$\frac{y_i}{se_i} = \alpha + \frac{\beta}{se_i} + \varepsilon_i \quad (7.2)$$

where ε_i is random error ($\varepsilon_i \sim N(0, \sigma^2)$). The null hypothesis is $\beta = 0$, i.e. there is no association between the precision of a study $\frac{1}{se_i}$ and its standardised effect

$\frac{y_i}{se_i}$. To obtain a p-value, $\beta / se(\beta)$ is compared to the t-distribution with $n-2$

degrees of freedom, where n is the number of studies in the meta-analysis (Egger *et al.*, 1997). Equation 7.2 is equivalent to

$$y_i = \alpha + \beta \cdot se_i + \varepsilon_i \cdot se_i \text{ weighted by } \frac{1}{se_i^2} \quad (7.3)$$

The rank correlation test and the regression test are both assessed at the 10% level of significance because of low statistical power (Egger *et al.*, 1997).

The funnel plot, rank correlation test and regression test are only useful to help identify whether publication bias is an issue, they do not solve the problem or help to assess the impact of any suspected publication bias. The trim and fill method, however, provides an adjusted (for publication bias) pooled estimate giving some idea of the likely effect of publication bias (Duval and Tweedie, 2000a; Duval and Tweedie, 2000b). It is an iterative non-parametric method based on the one-sided asymmetry of a funnel plot. The key assumption is that studies with the most extreme effect sizes are suppressed. Knowing the ranks of the absolute effect sizes and their signs around the pooled estimate, μ (which is obtained from either a fixed effects meta-analysis (Equation 3.1) or a random effects meta-analysis (Equation 3.2)), an estimate of k_o , the number of studies with the most extreme effect sizes that are 'missing', is obtained using one of three different estimators of k_o (Duval and Tweedie, 2000a; Duval and Tweedie, 2000b). They are

$$\begin{aligned} R_o &= \gamma^* - 1 \\ L_o &= \frac{4S_{rank} - n(n+1)}{2n-1} \\ Q_o &= n - \frac{1}{2} - \sqrt{2n^2 - 4S_{rank} + \frac{1}{4}} \end{aligned} \quad (7.4)$$

where n is the number of studies in the meta-analysis, γ^* is the length of the right most run of ranks and S_{rank} is the Wilcoxon statistic (Duval, 2005). Findings from simulations suggest that R_o and L_o are the preferred estimators to be used (Duval and Tweedie, 2000a; Duval and Tweedie, 2000b). Since μ is likely to be subject to

any publication bias that exists, the trim and fill method iteratively estimates k_o and μ . Thus, based on the initially estimated μ , k_o is estimate (using R_o or L_o) and ‘trimmed’ from one side of the funnel plot so that a more symmetrical plot is obtained. μ is then re-estimated on this ‘trimmed’ set of studies (using a fixed effects meta-analysis or a random effects meta-analysis), leading to a second estimate of k_o . This process is repeated until estimates of μ and k_o are stable. Given the final estimates of μ^* and k_o^* , the meta-analysis is ‘filled’, where the k_o^* ‘trimmed’ studies are replaced and the ‘missing’ studies have the estimates of effect and variance as their reflected ‘trimmed’ studies. A worked example of the trim and fill process using R_o or L_o is given in Duval (2005).

Selection models have also been used to model publication bias and provide an estimate of the pooled effect adjusting for the selection process (Iyengar and Greenhouse, 1988; Hedges and Vevea, 1996; Copas, 1999; Copas and Shi, 2001; Preston *et al.*, 2004). However, use of these models often requires there to be many studies in the meta-analysis and a great deal of computation (there is no specific software available for the general application of these models), and so they are not commonly used to assess publication bias (Hedges and Vevea, 2005).

Table 7.1 Availability of publication bias tests and methods in statistical software (Borenstein, 2005; Sterne *et al.*, 2001)

	Rank correlation test	Regression test	Trim & fill
Stata	Y	Y	Y
Comprehensive meta-analysis	Y	Y	Y
Metawin	Y	Y	N
RevMan	N	N	N
EasyMA	Y	N	N
StatsDirect	Y	Y	N

The rank correlation test, regression test, and the trim and fill method are commonly used to help identify publication bias in meta-analyses. To illustrate the frequent use of these methods, meta-analyses published (between May 1997 and January 2006) in the BMJ and the JAMA were surveyed. It was found that 42 out of 83

(51%) meta-analyses, in which an assessment of publication bias was made, report using the rank correlation test and/or the regression test. Only 8 (10%) reported using the trim and fill method. The popularity of the rank correlation and regression tests is also seen in the number of citations of the two papers introducing these tests. As of 27th March 2006, Web of Science holds 350 articles that cite Begg and Mazumdar's (1994) paper on the rank correlation test, and 908 articles citing Egger *et al.*'s (1997) paper on the regression test. The papers in which the trim and fill method is described (Duval and Tweedie, 2000a; Duval and Tweedie, 2000b) have 76 and 64 citations, respectively, in Web of Science (as of 27th March 2006). No doubt the common use of these tests and the trim and fill method is in part due to their availability in statistical software packages (see Table 7.1).

Although evidence suggests these methods are commonly used to help identify and adjust for possible publication bias, reports of simulation studies indicate that the performance of these methods is not as high as one would hope. Furthermore, a number of features of Egger's regression test require closer consideration. Firstly, Egger's regression test has a multiplicative error term (seen in Equation 7.3). This feature is not consistent with usual regression models where the error is additive (McCullagh and Nelder, 1996). Secondly, the test has been criticised for violating an assumption of regression models: that the independent variable, standard error in Equation 7.3, is estimated, and so subject to random error (Irwig *et al.*, 1998; Macaskill *et al.*, 2001). Thirdly, when the study summary estimates are ORs, there is a correlation between the $\ln(\text{OR})$ and its standard error, since the variance is a function of $\ln(\text{OR})$ (Irwig *et al.*, 1998; Macaskill *et al.*, 2001).

When the $\ln(\text{OR})$ is calculated from the usual 2x2 table

$$\begin{array}{c|c} a & b \\ \hline c & d \end{array} \quad \ln(\text{OR}) = \ln\left(\frac{ad}{bc}\right), \quad \text{and} \quad se(\ln(\text{OR})) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \quad (7.5)$$

Use of standard error as the independent variable in Egger's regression test (Equation 7.3) is likely to result in increased type I error rates, since this correlation is induced, particularly for large underlying ORs. Each study is also weighted by

the inverse of standard error, further magnifying effects from this correlation (Macaskill *et al.*, 2001).

In the following section papers using simulation analyses to assess the performance of the rank correlation test, the regression test or the trim and fill method are reviewed.

7.4 Review of related simulation studies

7.4.1 Rank correlation and regression tests

In their proposal for the rank correlation test, Begg and Mazumdar (1994) simulated meta-analyses of 25 and 75 primary studies to investigate the performance of the test. The underlying effect in the meta-analyses varied from 0 (the null hypothesis) through to three standard deviations from the null. The standardised variances of each primary study were either large (variance = 0.1, 1, 10) or small (variance = 0.5, 1, 2), so that the studies in the meta-analysis were of varying sizes. To induce funnel plot asymmetry, a smooth exponential weight function was used, with publication bias defined as 'strong' or 'moderate'. A number of different selection models were used based on the assumption that bias is dependent either only on the p-value or only on the effect estimate. Their findings were based on 5000 repeats of these combinations. Begg and Mazumdar found that the performance of the rank correlation test varies with the characteristics of the meta-analysis and the selection model used to induce dissemination bias. They concluded that although the test is powerful to detect publication bias for large meta-analyses (75 studies) and has moderate power for smaller meta-analyses (25 studies), the type I error rates (percentage of simulations where publication bias is incorrectly indicated) were lower than expected in all scenarios (they do not present these results).

Sterne *et al.* (2000) assessed the power of both the rank correlation test and Egger's regression test. The simulated meta-analyses are based on characteristics of 78 meta-analyses identified from a review of meta-analyses published between 1993 and 1997 from four general medical journals and four specialist medical journals. The number of primary studies in the meta-analyses was 5, 10, 20 or 30 with

underlying ORs of 1, 0.5 and 0.25. The underlying proportion of events in the control group was 5%, 10% and 20%. To induce publication bias a linear model was assumed, such that:

$$\ln(OR) = \text{true treatment effect} + (\text{bias coefficient} * \text{standard error of } \ln(OR))$$

The bias coefficient took the values 0 (i.e. no bias), -0.5 and -1. Sterne *et al.* describe three sets of simulations:

1. 1,000 repeats of each of the 108 combinations of the above parameters (4 sizes of meta-analysis, 3 OR sizes, 3 control event proportions and 3 degrees of publication bias)
2. 1,000 repeats of no bias and an extreme OR of 0.1, combined with 4 meta-analysis sizes and 3 control event rate proportions
3. For each of the 78 reviewed published meta-analyses the overall treatment effect, number of subjects in the treatment and control group and the observed control event proportions were obtained. 1,000 sets of simulations were based on each of these reviewed meta-analyses

From the findings of these simulations, the authors noted that as the number of primary studies in a meta-analysis increased, so did the power of both tests. They found that, in particular, when there is an underlying effect (i.e. $OR \neq 1$), the rank correlation test was generally less powerful than Egger's regression test. The rank correlation test had anywhere between 4% and 88% of the power of Egger's regression test for all scenarios (except where $OR=0.25$, the control group event rate = 5% and there was severe bias when the rank correlation test was more powerful than Egger's regression test). However, the type I error rates from Egger's regression test were too high when i) the OR was large, ii) there were few events in the study and iii) the primary studies were all of similar sizes.

Macaskill *et al.* (2001) assessed the power of the rank correlation test and the Egger regression test, in addition to an alternative regression test with sample size as the independent variable in order to overcome use of a random variable as the independent variable and help reduce the correlation between $\ln(OR)$ and its

standard error observed in Egger's regression test. The alternative regression model is given as

$$y_i = \alpha + \beta.size_i + \varepsilon_i \quad (7.6)$$

where y_i is the estimate of effect from study i , and $size_i$ is the total sample size for that study. In one form of this model, each estimate of effect is weighted by its precision. In the second form (referred to as the funnel pooled variance (FPV) model), each estimate is weighted by $\left(\frac{1}{a+b} + \frac{1}{c+d}\right)^{-1}$. Where a is the number of subjects in group 1 in which an event was observed, b is the number of subjects in group 2 in which an event was observed, and c and d are the number of subjects for which no events were observed in group 1 and group 2, respectively. This weighting is based on the assumption that the null hypothesis is true, i.e. the underlying OR is one.

The underlying OR in each of the simulated meta-analyses was 1, 0.67, 0.5 or 0.25, with an underlying probability of an event in the control group given by a uniform distribution (0.1, 0.5). The proportion of events in the treatment group is calculated from this via the OR. A selection model is used to induce 'severe' publication bias (as given in Begg and Mazumdar 1994). The probability of selection is based on the p-value, as opposed to the effect size. 10,000 repeats were generated for every combination.

Macaskill *et al.* report that the power to detect publication bias was low for all four tests. In particular, Egger's regression test exceeded the expected type I error rates and the authors report a marked imbalance in the tail probability areas for the 2-tailed test. Macaskill *et al.* conclude that the FPV regression model is the preferred method because of favourable type I error rates.

More recently, Schwarzer *et al.* (2002) assessed the type I error rates of the rank correlation test and Egger's regression test in the presence of unexplained between-study heterogeneity. They do not look at the power of these tests. The simulated

meta-analyses contained 10, 20 or 50 primary studies. The size of these studies was based upon findings of an empirical review of eight German medical journals; only trials with more than 30 patients were included in the simulation analyses. Both ORs and relative risks (RRs) were considered as the estimate of effect, but only those based on ORs are reported, as the results for the RRs were similar. The underlying OR was 0.5, 0.67, 1, 1.5 and 2 and between-study variance was defined as 0%, 25% and 50% of the within study variance of a study with 100 subjects. 10,000 repeats of each combination of parameter values were generated. The authors also considered the case for meta-analyses with larger primary studies.

Schwarzer *et al.* found that the type I error rates for both the rank correlation test and Egger's regression test exceeded the expected 10% level. They noted that the rank correlation test appears to perform well when OR=1, but for both tests, as the treatment effect and the number of studies increased, the type I error rates also increased. As between-study heterogeneity increases, the type I error rates increase for Egger's regression test, but are acceptable for the rank correlation test. However, the simulations including between-study heterogeneity are based on characteristics of a meta-analysis of thrombolytic therapy, and the authors point out that the high type I error rates could be due to a feature of that meta-analysis.

From these four articles, there is evidence that the rank correlation test is less powerful than Egger's regression test (Sterne *et al.*, 2000; Macaskill *et al.*, 2001), but that the type I error rates for Egger's regression test are generally too high (Sterne *et al.*, 2000; Schwarzer *et al.*, 2002). The power of these tests tends to increase as i) the number of primary studies increase (Sterne *et al.*, 2000), ii) the control group event rate increases (Sterne *et al.*, 2000) and iii) the further the OR is from the null (Sterne *et al.*, 2000; Macaskill *et al.*, 2001). Unfortunately it seems that in situations where such methods would be most appealing (e.g. small number of primary studies, moderate levels of bias), the power of the rank correlation test and Egger's regression test is low (Begg and Mazumdar, 1994; Sterne *et al.*, 2000; Macaskill *et al.*, 2001). On the other hand, the FPV regression test in Macaskill *et al.* (2001) had the appropriate type I error rates and power to detect induced publication bias. Further research is needed to assess the performance of

Macaskill's regression test compared to the rank correlation and Egger's regression test in different meta-analysis scenarios.

7.4.2 Trim and fill method

Assessment of the performance of the trim and fill method has been limited. Duval and Tweedie (Duval and Tweedie, 2000a; Duval and Tweedie, 2000b) defined the total number of studies conducted and relevant to the meta-analysis as $N = n + k$, where N is the total number of studies carried out ($N = 25, 50$ and 75), defined by n , the number of observed studies and k , the number of unobserved studies ($k = 0, 5$ and 10). Results are based on 1000 repeats using a fixed effects model. In the paper, publication bias is based on the assumption that studies with the most extreme estimates of effect (at one end of the effect size scale) have been censored, rather than based on the p-value as in Begg and Mazumdar (1994), Hedges and Vevea (1996) and Macaskill *et al.* (2001). Inducing publication bias on the basis of effect size corresponds to the key assumption of the trim and fill methods, that studies with the most extreme effect sizes are censored (Duval and Tweedie, 2000a; Duval and Tweedie, 2000b).

When there is no bias (i.e. $k = 0$), the error rate is slightly less than that expected; when the error rate of 2.5% is expected, the error rate obtained is at best 2% and at worse 1%. The authors explain that this is because a random effects model has been used to analyse data generated from a fixed effects model and is therefore conservative. Duval and Tweedie (2000a) conclude that the trim and fill method performs well in estimating the number of 'missing' studies when publication bias is induced by effect size, but stress that it is a useful tool for sensitivity analyses only.

Terrin *et al.* (2003), however, conclude that in some situations the trim and fill method incorrectly adjusts for studies that are not 'missing'. In their simulations, both fixed and random effects models are used to generate meta-analyses. The parameter values are based on the findings of a review of 125 meta-analyses. The number of primary studies was set to be either 10 or 25 and the size of the primary studies were taken from the following i) a uniform distribution ($\ln(50), \ln(500)$), ii) a uniform distribution ($\ln(100), \ln(1500)$), iii) a uniform distribution

($\ln(100)$, $\ln(10000)$) or iv) chosen to attain 80% power. The underlying ORs were defined as large, moderate and no effect (0.5, 0.8 and 1, respectively), with underlying probability (and variance for random effects model) of an event in the control group of 0.15 (0.005) or 0.3 (0.02) from which the true probability in the treatment group was calculated. The between-study variance for the random effects was either 0.01 or 0.15. All combinations of the above parameter values were repeated 1000 times. Terrin *et al.* (2003) do not induced publication bias in any of their simulated scenarios, and therefore only assess the performance of the trim and fill method in the absence of publication bias.

When the primary studies were from a fixed effects model, the trim and fill method performed well in that the coverage probabilities were similar to the expected coverage probability of 0.95. However, when the studies were from a random effects model and the between-study heterogeneity was large ($\tau^2 = 0.15$), the coverage probabilities from the trim and fill method were low. In conclusion Terrin *et al.* caution that in the presence of between-study heterogeneity, the trim and fill method inappropriately adjusts for publication bias when none exists.

Generally, it would appear that methods to detect publication bias sound appealing and intuitive, but it is not necessarily the case that their properties are good enough to allow safe use in practice. Only two of the papers described here (Schwarzer *et al.*, 2002; Terrin *et al.*, 2003), consider another important aspect of meta-analysis; between-study heterogeneity. As mentioned in Section 3.3.2, between-study heterogeneity can seriously affect the inference and conclusions of a meta-analysis and its sources should be investigated within the meta-analysis framework, particularly as evidence of between-study heterogeneity is found in many meta-analyses as Engels *et al.* (2000) and Villar *et al.* (2001) conclude.

Engels *et al.* (2000) reviewed 125 meta-analyses of randomized controlled trials using binary data (meta-analyses had to have > 6 studies to be included) from 7 major journals from 1990 to 1996, and the 1994 Cochrane pregnancy and childbirth database. Risk differences and ORs were calculated for each study and were then meta-analysed. A test for between-study heterogeneity was applied to all meta-

analyses. Results indicated that meta-analyses of ORs were less likely to show significant between-study heterogeneity compared to meta-analyses of risk difference estimates. Nevertheless, at the usual 10% level of significance for this test, 35% of meta-analyses indicated between-study heterogeneity when ORs were combined and 47% when risk differences were combined. In addition to this, Villar *et al.* (2001) found that 25% of the 84 meta-analyses published in Issue 3 of the Cochrane Library's pregnancy and childbirth module (1998) had evidence of between-study heterogeneity. Between-study heterogeneity is a feature common to many meta-analyses and so must be considered alongside an assessment of publication bias. As yet there has been very little work on assessing publication bias in heterogeneous meta-analyses. Terrin *et al.* (2003) investigated the impact between-study heterogeneity had on the performance of the trim and fill test and Schwarzer *et al.* (2002) have looked at the type I error rates of the rank correlation test and the regression test in the presence of between-study heterogeneity. Summary characteristics of the published simulation studies described above are given in Table 7.2

Table 7.2 Summary of characteristics of published simulation studies assessing performance of tests and adjustments for publication bias

Study	Test(s) and Method(s) of interest	Publication bias	How publication bias induced	Number of studies in the meta-analysis	Underlying effect(s)	Between-study heterogeneity
Begg and Mazumdar (1994)	Rank correlation test	Yes	p-value or effect size	25, 75	0 (null hypothesis) + 3 standard deviations	Yes
Sterne <i>et al.</i> (2000)	Rank correlation test Egger's regression test	Yes	Function of standard error	5, 10, 20, 30	ORs of 1, 0.5 and 0.2	No
Macaskill <i>et al.</i> (2001)	Rank correlation test Egger's regression test FIV sample size regression FPV sample size regression	Yes	p-value	21, 63	ORs of 1, 0.67, 0.5 and 0.25	No
Schwarzer <i>et al.</i> (2002)	Rank correlation test Egger's regression test	No	-	10, 20, 30	ORs of 0.5, 0.6, 1, 1.5 and 2	Yes
Duval and Tweedie (2002a)	Trim and fill method	Yes	Effect size	25, 50, 75	Not specified	No
Terrin <i>et al.</i> (2003)	Trim and fill method	No	-	10, 25	ORs of 0.5, 0.8 and 1	Yes

From Table 7.2, one can see that there are gaps in the evidence concerning the performance of these tests and the trim and fill method. For instance, the performance of Egger's regression test in comparison to the rank correlation test has not been assessed in the presence of unexplained between-study heterogeneity; the trim and fill method has not been compared to the usual unadjusted meta-analysis methods (see Equations 3.1. and 3.2) in the presence of publication bias induced other than on the basis of effect size, nor in the presence of both publication bias and between-study heterogeneity. Furthermore, the promising results of the FPV regression model in Macaskill *et al.* (2001) requires further examination. Building on the published work described above and in Table 7.2, simulation analyses have been carried out to assess the performance of the rank correlation test, Egger's regression test and a number of alternative regression tests, in addition to the trim and fill method. Under conditions of varying levels of publication bias and unexplainable between-study heterogeneity, the power and type I error rates of these tests are assessed and compared, and the bias in estimates from the trim and fill method is investigated. In Chapter 8, the performance of these tests is assessed in the presence of unexplainable between-study heterogeneity, and that which can be explained to some extent by a measured covariate (explainable between-study heterogeneity).

7.5 Simulations

7.5.1 Parameters

Both fixed and random effects models have been used to simulate the meta-analyses. The fixed effects model is given by

$$y_i = \theta + \varepsilon_i, \quad (7.7)$$

where θ is the true underlying effect, $\ln(\text{OR})$.

The random effects model is

$$y_i = \theta_i + \varepsilon_i, \quad \theta_i \sim N(\mu, \tau^2) \quad (7.8)$$

where θ_i is the true effect in study i , μ is the true underlying effect, $\ln(\text{OR})$, and τ^2 is the estimate of between-study variance.

The simulated meta-analyses were based on characteristics of the meta-analyses of animal experiments identified from the systematic review in Chapter 3. However, findings can be applied generally. The characteristics took the following values:

- The number of primary studies in a meta-analysis was 6, 16, 30 or 90 studies.
- The probability of an adverse event in the control group was sampled from a uniform distribution (0.3, 0.7).
- The natural logarithm of the number of control subjects within each primary study was taken from the distribution $N(5, 0.3)$.
- For simplicity the ratio of exposed to control subjects was one.
- The underlying ORs were 1, 1.2, 1.5, 3 or 5.

Given the known number of subjects in the control group, the probability of an adverse event in the control group and the underlying OR, values for the individual cells of the usual 2x2 table for calculation of an OR (see Equation 7.5) were simulated.

For random effects meta-analyses, the between-study variance is defined to be 20%, 150% and 500% of the average within-study variance for studies from the corresponding simulation. This can be compared in terms of I^2 , the percentage of total variation across studies that is due to between-study heterogeneity rather than chance (Higgins and Thompson, 2002). Here, the between-study variation defined to be 20%, 150% and 500% of the within-study variation corresponds to an I^2 of 16.7%, 60% and 83.3%, respectively.

Publication bias was induced in two ways: 1) on the assumption that studies are excluded from the meta-analysis as a result of the *one-sided p-value* associated with the effect estimate of interest (Begg and Mazumdar, 1994; Hedges and Vevea, 1996; Macaskill *et al.*, 2001), 2) on the assumption that the *size* of the effect

estimate determines whether a study is included in the meta-analysis or no, so that studies with the most extreme estimates of effect are excluded (Duval and Tweedie, 2000a).

Inducing publication bias by p-value

Two levels of publication bias were simulated based on the one-sided p-value. These two levels are termed 'moderate' and 'severe', and correspond to levels specified in Hedges and Vevea (1996). The probability for inclusion into the meta-analysis depends on the p-value from the study. This probability is defined by a step function and is shown in Table 7.3. Thus, for moderate publication bias, a study with a p-value between 0.2 and 0.5 has a probability of 0.5 for inclusion in the meta-analysis.

Table 7.3 *Specification of publication bias severity based on one-sided significance*

Severity of publication bias	p-value from study	Probability for inclusion
Moderate	<0.05	1
	0.05 – 0.2	0.75
	0.2 – 0.5	0.5
	>0.5	0.25
Severe	<0.05	1
	0.05 – 0.2	0.75
	> 0.2	0.25

All simulations were repeated until the desired number of studies in each meta-analysis (6, 16, 30 or 90) was obtained. For the scenario where the underlying OR is 1 and the number of studies included in the meta-analysis is 30, the studies included and excluded from the meta-analysis by these two levels of publication bias severity are given in Figure 7.3.

As Figure 7.4 shows, more studies need to be generated for severe bias than for moderate bias, until the specified number of studies is obtained. This is as expected, since fewer studies are excluded for moderate bias than for severe bias.

Figure 7.3 Funnel plot of studies from simulation of 'moderate' and 'severe' publication bias; included and excluded studies are indicated

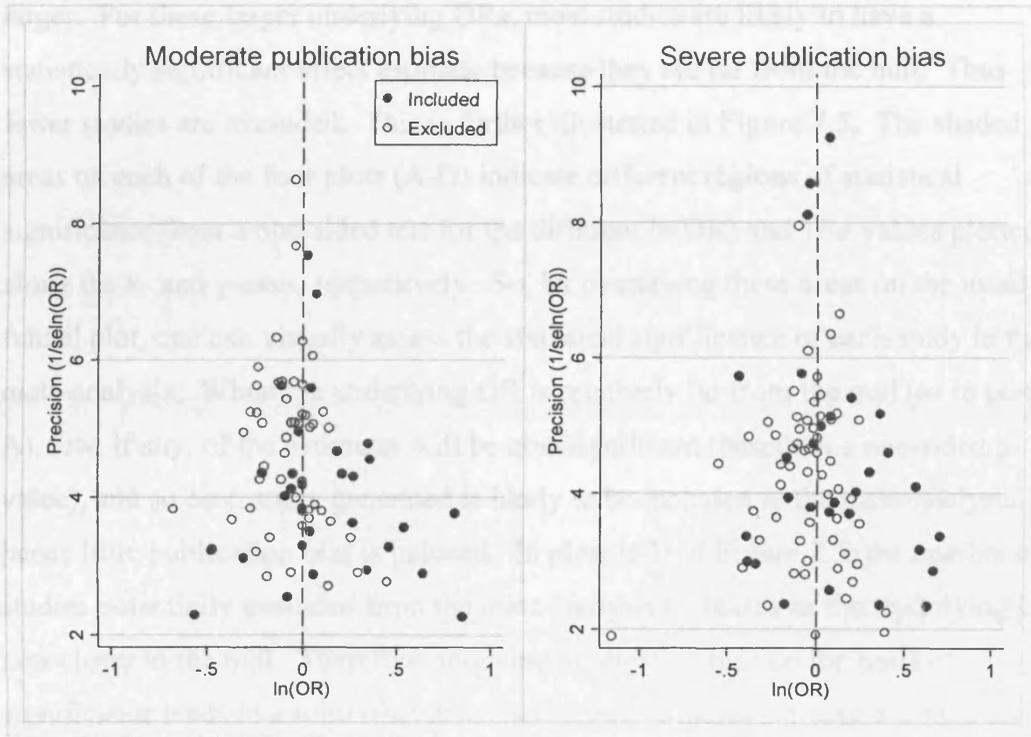
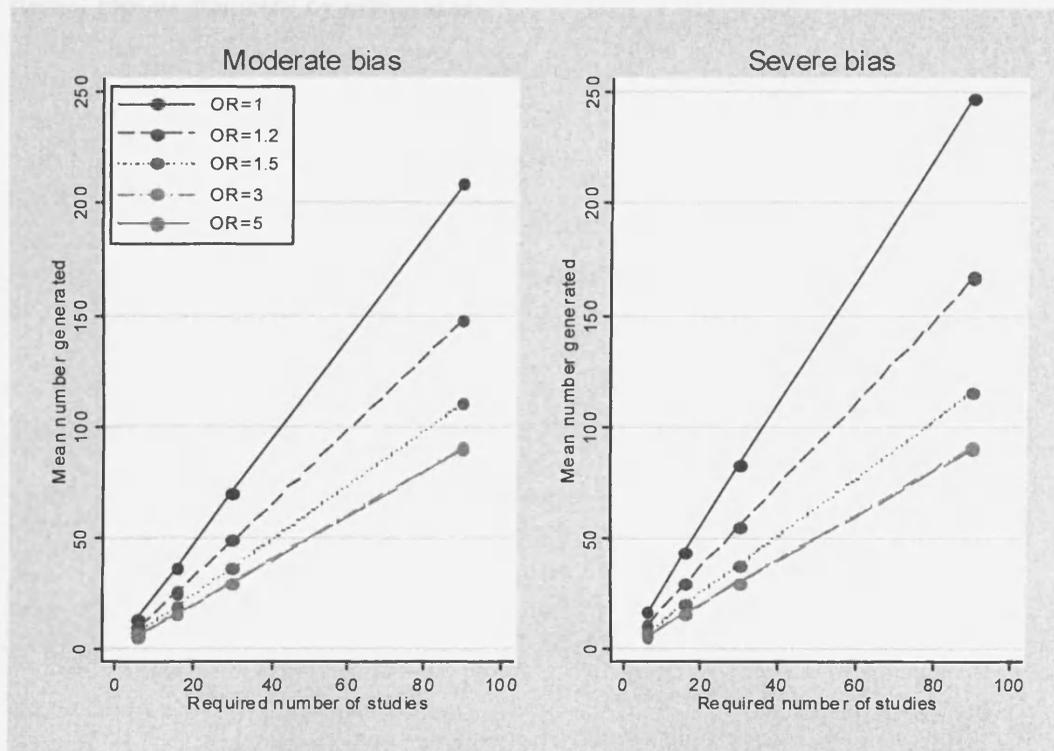
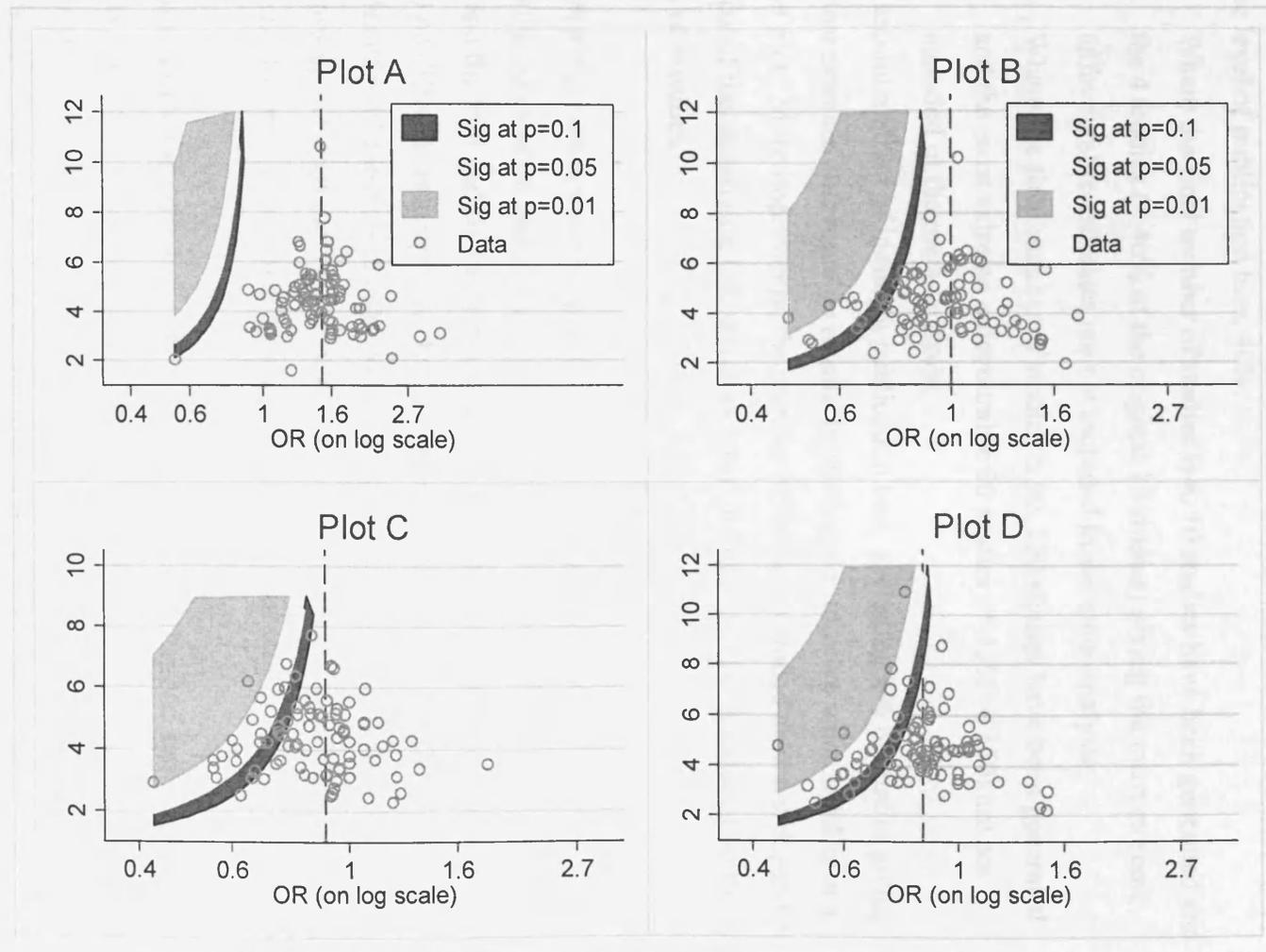


Figure 7.4 Mean number of studies generated to obtain the number of studies required under 'moderate' and 'severe' bias for each specified underlying OR



Of particular interest in Figure 7.4 is that the mean number of studies generated to obtain the required number of included studies reduces as the underlying OR gets larger. For these larger underlying ORs, most studies are likely to have a statistically significant effect estimate because they are far from the null. Thus fewer studies are excluded. This is further illustrated in Figure 7.5. The shaded areas on each of the four plots (A-D) indicate different regions of statistical significance from a one-sided test for the different $\ln(\text{OR})$ and $1/\text{se}$ values plotted along the x- and y-axes, respectively. So, by overlaying these areas on the usual funnel plot, one can visually assess the statistical significance of each study in the meta-analysis. When the underlying OR is relatively far from the null (as in plot A), few, if any, of the estimates will be non-significant (based on a one-sided p-value), and so each study generated is likely to be included in the meta-analysis, hence little publication bias is induced. In plots B-D of Figure 7.5, the number of studies potentially excluded from the meta-analysis increases as the underlying OR gets closer to the null. Therefore, inducing publication bias on the basis of significance leads to a somewhat distorted picture. Studies simulated from a meta-analysis with a large underlying OR are less likely to be excluded from the meta-analysis than studies from a meta-analysis where the underlying OR is close to the null. This phenomenon leads to consideration of an alternative method for inducing funnel plot asymmetry by effect size.

Figure 7.5 Examples of censoring by level of statistical significance of the effect estimate



Plot A: underlying OR is far from the null (OR = 1.5); Plot B: underlying OR = 1
 Plot C: underlying OR = 0.9; Plot D: underlying OR = 0.8

Inducing publication bias by effect size

Again, two levels of bias were induced to represent ‘moderate’ and ‘severe’ publication bias. Either the 14% or 40% most extreme studies showing an unfavourable effect were excluded from the meta-analysis such that the final number of studies in a meta-analysis was still 6, 16, 30 or 90. For example, with the severe level of publication bias, 40%:

- Where the final number of studies is 6, 10 studies have been generated and the 4 studies (= 40% of the original 10 studies) giving the most extreme unfavourable estimates are not included in the meta-analysis.
- Where the final number of studies is 90, 150 studies have been generated and the most extreme unfavourable 60 studies (= 40% of 150) are not included in the meta-analysis.

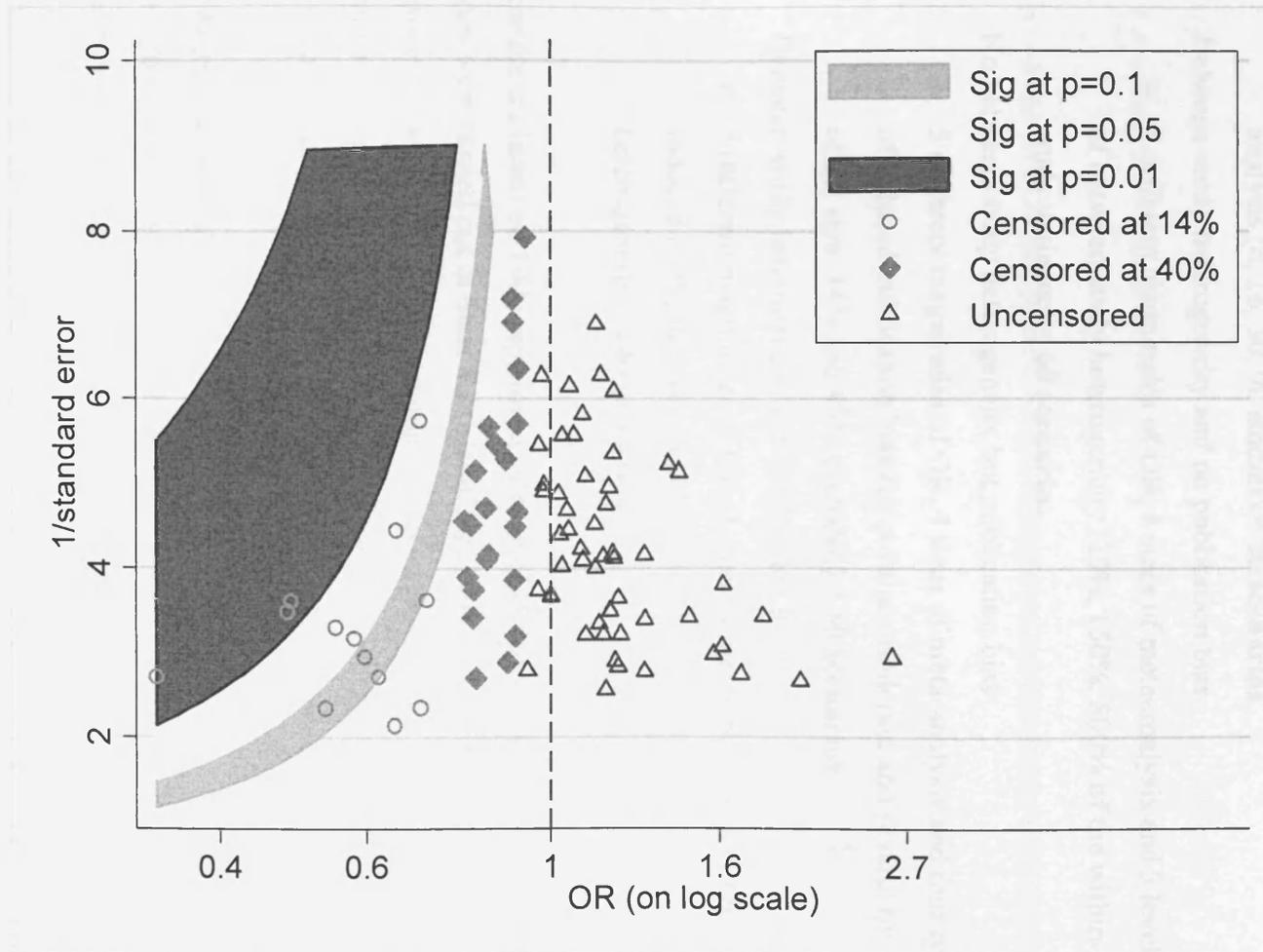
This second method of inducing publication bias, excluding x% of studies giving the most extreme unfavourable results, is much more intuitive when looking at a funnel plot. More importantly, the number of studies excluded does not depend on the size of the underlying OR as it does when publication bias is induced on the basis of p-values.

Relationship between publication bias induced by p-value and effect size

Inducing publication bias in these simulations may well impact on how the different tests and the trim and fill method are seen to work because the tests and methods are based on different assumptions. For instance, the trim and fill method is based on the idea of publication bias induced by effect size rather than p-value. The relationship between the p-value of an estimate effect and its size is not immediately obvious. Figure 7.6 may help in understanding the connection.

In this example the underlying OR is 1. The circle and diamond symbols represent studies that would be excluded using the 14% and 40% cut-off points, respectively, for the most extreme effect sizes. The shaded areas show the level of (one-sided) significance of the effect estimate, illustrating the studies that are more likely to be excluded (i.e. those found in the shaded areas).

Figure 7.6 *Inducing publication bias by p-value and by size of effect*



In summary, the simulations used in this chapter to assess the performance of the rank correlation test, a number of regression tests, and the trim and fill method take the following form:

- No between-study heterogeneity and no publication bias
 - » 5 different sizes of OR (ORs of 1, 1.2, 1.5, 3, 5) and 4 sizes of meta-analysis (6, 16, 30, 90 studies) = **20 scenarios**
- Between-study heterogeneity and no publication bias
 - » 5 different magnitudes of OR, 4 sizes of meta-analysis and 3 levels of between-study heterogeneity (20%, 150%, 500% of the within-study variance) = **60 scenarios**
- No between-study heterogeneity but publication bias
 - » 5 different magnitudes of OR, 4 sizes of meta-analysis and four types of induced publication bias (by p-value: moderate and severe; by effect size: 14% and 40% excluded) = **80 scenarios**
- Between-study heterogeneity and publication bias
 - » 5 different magnitudes of OR, 4 sizes of meta-analysis, four types of induced publication bias and three levels of between-study heterogeneity = **240 scenarios**

The results are based on 1000 repetitions of each of the above 400 situations. All analyses were carried out in Stata 8.2 (StataCorp, 2004). In the next section Egger's regression test and alternative regression tests assessed here are specified. These alternative regression tests defined here allow for different criticisms of Egger's regression test (given at the end of Section 7.3) to be investigated and compared.

7.5.2 Regression models

Egger's fixed effects regression on the standard error

As defined in Equations 7.2 and 7.3, Egger's regression test is given by

$$\frac{y_i}{se_i} = \beta + \frac{\alpha}{se_i} + \varepsilon_i, \text{ which is equivalent to}$$

$$y_i = \alpha + \beta \cdot se_i + \varepsilon_i \cdot se_i \text{ weighted by } \frac{1}{se_i^2} \quad \text{(Model 1)}$$

where y_i is the $\ln(\text{OR})$ from study i and se_i is the standard error of y_i .

To avoid violating an assumption of regression analysis, a model is defined where the inverse of sample size is used in place of the standard error in a form of Egger's regression test (Model 2).

Egger's fixed effects regression on the inverse of sample size

$$y_i \cdot size_i = \alpha + \beta \cdot size_i + \varepsilon_i \quad \text{(Model 2)}$$

where $size_i$ is the total sample size of study i .

To address the issue of multiplicative error in Egger's test (Model 1), Model 3 is defined as the additive error version of Model 1.

Linear fixed effects regression on standard error

$$y_i = \alpha + \beta \cdot se_i + \varepsilon_i, \quad \text{weighted by } \frac{1}{se_i^2} \quad \text{(Model 3)}$$

Macaskill *et al.* (2001) proposed a linear model with additive error as that in Model 3, but where the independent variable, standard error, is replaced by the total sample size. As described in Section 7.4.1 Macaskill *et al.* define two forms of this model depending on the weighting used and these are defined as Models 4a and 4b.

Linear fixed effects regression on sample size (FIV model)

$$y_i = \alpha + \beta \cdot size_i + \varepsilon_i, \quad \text{weighted by } \frac{1}{se_i^2} \quad \text{(Model 4a)}$$

Linear fixed effects regression on sample size (FPV model)

$$y_i = \alpha + \beta \cdot size_i + \varepsilon_i, \quad \text{weighted by } \left(\frac{1}{a_i + b_i} + \frac{1}{c_i + d_i} \right)^{-1} \quad \text{(Model 4b)}$$

In Models 1 and 3, standard error is used as a measure of precision for each study and is then tested for an associated with effect size. Precision can also be thought of in terms of the inverse of sample size for each study; studies with large sample sizes

tend to have larger precision. Thus Model 4c is defined as a version of Model 4b, where the inverse of sample size is used in place of sample size.

Linear fixed effects regression on inverse of sample size

$$y_i = \alpha + \frac{\beta}{size_i} + \varepsilon_i, \quad \text{weighted by } \left(\frac{1}{a_i + b_i} + \frac{1}{c_i + d_i} \right)^{-1} \quad \text{(Model 4c)}$$

Finally, two random effects linear models are defined (Model 5 and Model 6). Since between-study heterogeneity is induced in some simulations used here, it is of interest to assess how well random effects models for publication bias will fair in such simulations. Models 5 and 6 can be thought of as random effect versions of Models 3 and 4a.

Linear random effects regression on standard error

$$y_i = \alpha + \beta.se_i + \mu_i + \varepsilon_i, \quad \text{weighted by } \frac{1}{se_i^2} \quad \text{(Model 5)}$$

Linear random effects regression on sample size

$$y_i = \alpha + \beta.size_i + \mu_i + \varepsilon_i, \quad \text{weighted by } \frac{1}{se_i^2} \quad \text{(Model 6)}$$

The eight regression models described above are summarised in Table 7.4.

Table 7.4 Summary of the regression models assessed in these simulations

Model based on...	Egger's fixed effects model	Linear fixed effects model	Linear random effects model
Some transformation of standard error	Model 1	Model 3	Model 5
Some transformation of sample size	Model 2	Model 4 *	Model 6

* three different weightings for each study are implemented (Models 4a, 4b and 4c)

All of these regression models test an association between the (standardized) effect size from each study in the meta-analysis, y_i , and a measure of its precision (whether a transformation of standard error or sample size). Evidence of an association may suggest that the meta-analysis is subject to publication bias if the smaller, less powerful studies have larger effect sizes than the more precise studies.

7.5.3 Estimate of effect

When publication bias is suspected one should go beyond simply testing for it in an attempt to determine if this publication bias is likely to have a significant impact on the conclusions one may draw from the meta-analysis. The trim and fill method allows some examination of the possible effect of publication bias on the pooled estimate. Given that a fixed effects or random effects meta-analysis model can be used in both the iterative part of the trim and fill process and the calculation of the adjusted pooled effect from the 'filled' meta-analysis (see Section 7.3), there are four possible fixed and random effects versions of the trim and fill method: fixed-fixed effects, fixed-random effects (*fixed effects* model for the *iterative process* and *random effects* for the *pooled estimate*), random-fixed effects, random-random effects. The random-random effects version is advocated by Duval and Tweedie (2000a; 2000b), however, more recently discussion of the reporting of results from the fixed-fixed effects model has received attention (Sutton, 2005). In these simulation analyses only the performance of the fixed-fixed effects and random-random effects trim and fill models are considered and are thus referred to as the trim and fill fixed effects model and the trim and fill random effects model in the rest of the thesis. There has, in the past, been some discussion as to alternative ways to proceed in a meta-analysis when publication bias is suspected. One method is the 'best evidence synthesis' approach proposed by Slavin (1986 and 1995). This approach requires strict inclusion criteria to be set so that only the best evidence, in terms of quality, are included and synthesised. This approach has many benefits, but although quality is likely to be affected by sample size, such an approach may still be subject to publication bias as key factors leading to publication bias include the effect size and statistical significance, in addition to sample size. However, Berlin *et al.* (1989) suggest a form of best evidence synthesis wherein sample size is the key factor for inclusion into the meta-analysis. They argue that since large

studies are usually published and it is the effect estimates of smaller studies that are subject to more random variation, a meta-analysis of only large studies should not be subject to publication bias. There is therefore a case for looking at results from individual studies rather than using results from a meta-analysis especially if there is evidence of heterogeneity and/or publication bias. With this in mind, the performance of using the estimate from the largest study or the study that is most precise in place of the pooled estimate is examined in these simulation analyses in the presence and absence of publication bias and between-study heterogeneity. To summarise, the estimate of effect from six different methods are compared:

1. Usual unadjusted fixed effects meta-analysis estimate (see Equation 3.1)
2. Usual unadjusted random effects meta-analysis estimate (see Equation 3.2)
3. Trim and fill fixed effects meta-analysis estimate (see Equation 7.4)
4. Trim and fill random effects meta-analysis estimate (see Equation 7.4)
5. Estimate from the largest study in the meta-analysis
6. Estimate from the study with the greatest precision in the meta-analysis

7.5.4 The analyses

The *type I error rates* (percentage of simulations where publication bias is incorrectly indicated at $p < 0.1$) for the rank correlation test, Egger's regression test (Model 1) and the alternative regression tests (Models 2, 3, 4a, 4b, 4c, 5 and 6) are investigated in a range of scenarios described in Section 7.5.1.

In the presence of publication bias, the *power* (percentage of simulations where publication bias is correctly indicated at $p < 0.1$) of these tests is explored. An ideal test would have type I error rates of 10% and good power to detect publication bias when it is present, regardless of the size of the underlying effect, the number of primary studies in the meta-analysis and the amount of between-study heterogeneity. The corresponding maximum standard error for these estimates of the type I error rates and power are also given.

The two-tailed p-value for the coefficient of interest in each regression model is calculated in two ways: i) from the usual t-test and ii) from a permutation test (or

randomization test). The permutation test has been proposed by Higgins and Thompson (2004) as an alternative in meta-regression to temper high type I error rates. This approach does not rely on an assumed distribution of the regression coefficients when calculating statistical significance, unlike the usual approach. The permutation test has not before been used to calculate p-values from regression tests to detect publication bias.

Performance of the six methods to estimate an effect outlined above are assessed through measures of

- *relative bias*, the difference between the underlying effect and the estimated effect as a percentage of the underlying effect, where a negative bias indicates an underestimate of the underlying OR, a positive bias indicates an overestimate, and
- *coverage probabilities*, the proportion of simulations in which the underlying effect lies within the 95% confidence intervals of the estimate, in the absence and presence of publication bias.

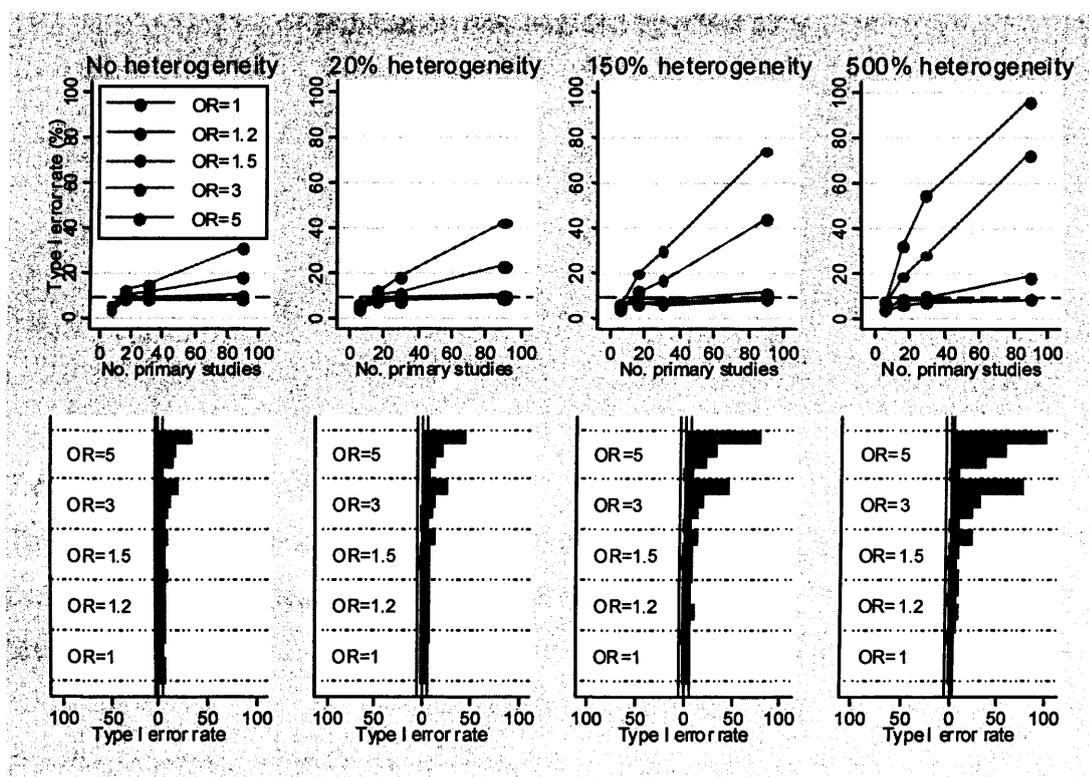
When looking at the performance of these methods it must be borne in mind that when publication bias is induced on the basis of the p-value associated with the effect estimate from a study, a meta-analysis with a large underlying OR is unlikely to contain many studies with a non-significant estimate (as discussed in Section 7.5.1). Because of this, few studies from such a meta-analysis will be excluded and so little publication bias will be induced. Hence, for meta-analyses where the underlying $OR \geq 3$, and publication bias has been induced by p-value, any effect of publication bias will be minimal. This must be taken into account when interpreting the performance of these methods to detect, and adjust for, possible publication bias. The performance of the tests and the trim and fill method in these simulated meta-analyses is reported in the following section. Only results for 'severe' publication bias are shown in this chapter, however relevant results for 'moderate' publication bias are referred to and given in Appendix I (these follow the same trend as for 'severe' bias, but results for 'moderate' bias are not as pronounced).

7.6 Results

7.6.1 The rank correlation test

The type I error rates of the rank correlation test for different underlying ORs, meta-analysis size (the number of primary studies in the meta-analysis) and level of between-study heterogeneity are given in Figure 7.7. The maximum standard error in estimates of type I error rates and power for the rank correlation tests is 1.6%, so that estimates of the type I error rates and power have a 95% CI that does not exceed $\pm 1.96 * 1.6\%$ in width. For small meta-analyses ($n=6$), the type I error rates are lower than expected, however for large ORs and reasonably sized meta-analyses ($n \geq 30$), the error rates are larger than expected, especially when there is a great deal of heterogeneity. Imbalance in the tail probability areas can also be seen (bottom row of Figure 7.7).

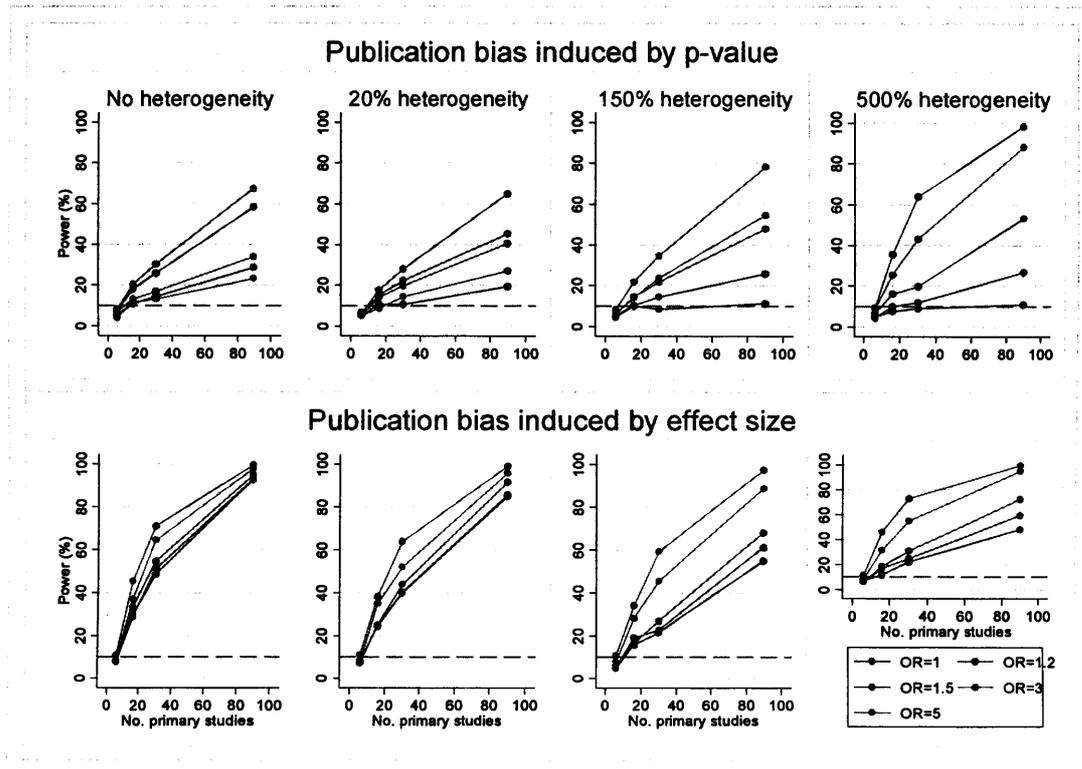
Figure 7.7 Type I error rates for the rank correlation test



Results of the simulation analyses suggest that the rank correlation test has good power to detect 'severe' publication bias, especially when there is a small or moderate underlying effect, but that this power tends to decrease as the amount of between-study heterogeneity increases (Figure 7.8). When the underlying effect is

large ($OR \geq 3$), it is difficult to distinguish between type I error rates and power, because of the large type I error rates observed in Figure 7.7. This trend in levels of power for 'severe' bias is also seen when 'moderate' publication bias is induced, although the observed power is not as high (see Figure I.1 in Appendix I).

Figure 7.8 Power of the rank correlation test to detect 'severe' publication induced by p-value (top row) and effect size (bottom row)



7.6.2 The regression model tests

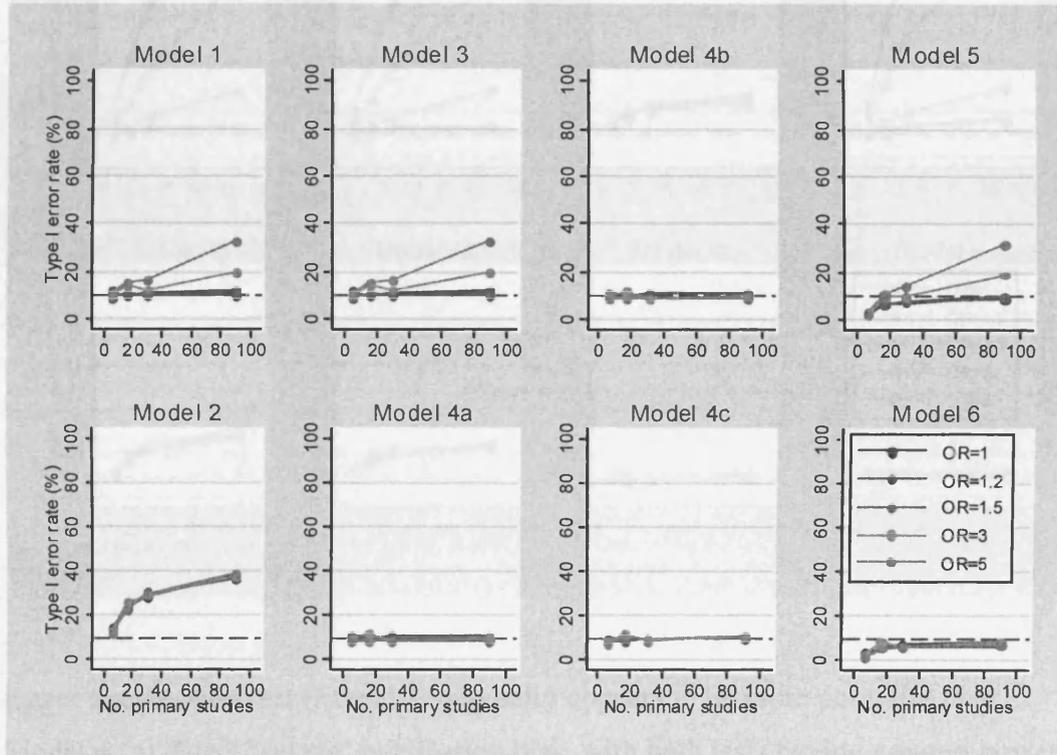
The maximum standard error in the estimates of power and type I error rates from the eight regression tests using p-values from the usual t-test and the permutation test is 1.6%.

P-values obtained from the usual t-test

The first set of results for the regression model tests are based on the use of the usual t-test to obtain p-values for coefficients of a regression model. Results from the permutation tests are given later. Unlike the rank correlation test, when applied to small meta-analyses Egger's regression test (Model 1) has the expected 10% type I error rates (Figure 7.9). However, as with the rank correlation test, the expected

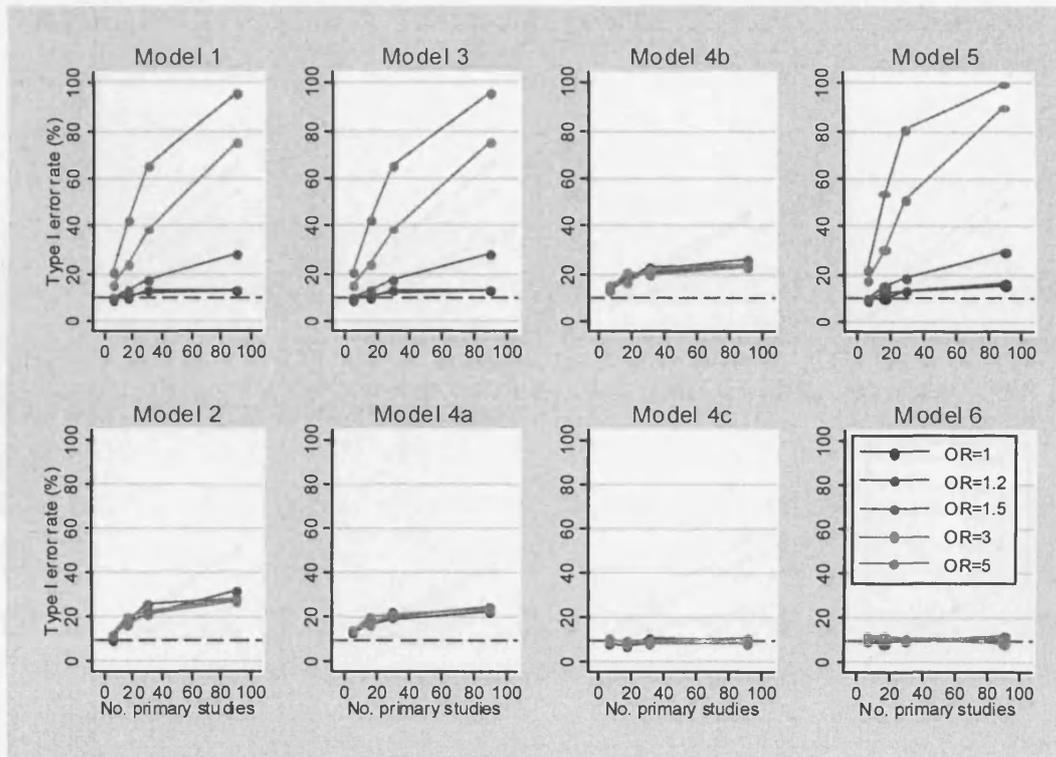
type I error rate is exceeded for meta-analyses with large underlying effects ($OR \geq 3$).

Figure 7.9 Type I error rates for all regression models when there is no between-study heterogeneity



Three tests (Models 4a, 4b and 4c) appear to perform well in terms of attaining the expected type I error rates regardless of the number of primary studies and the size of the underlying effect. Apart from having very low type I error rates for meta-analyses with a small number of primary studies, Model 6 appears to perform reasonably well when there are a large number of studies in the meta-analysis (although the expected type I error rate of 10% is not quite achieved). When there is a great deal of between-study heterogeneity (see Figure 7.10), Model 6 attains the expected type I error rates, however Model 4c is the only model to consistently achieve the expected 10% type I error rate regardless of the amount of between-study heterogeneity, the number of primary studies in the meta-analysis and the size of the underlying effect.

Figure 7.10 Type I error rates for all regression models when between-study heterogeneity is 500% of the average within-study heterogeneity



Egger's regression test (Model 1) generally appears to be more powerful than Model 4c to detect 'severe' publication bias, with both tests having greatest power when 'severe' publication bias is induced on the basis of effect size (compare Figures 7.11 and 7.12). However, this observed power must be interpreted in light of the type I error rates, and clearly a trade-off between the two is required. Because of the high type I error rates of Egger's regression test, it is difficult to distinguish power from these error rates. On the other hand, since Model 4c has the expected type I error rates in all scenarios, there is more confidence in the observed power actually representing power to detect publication bias. For these reasons Model 4c appears superior to Egger's regression test and the other six regressions tests to detect publication bias when the p-values are calculated from the usual t-test.

Figure 7.11 Power to detect 'severe' publication bias induced by p-value when there is no between-study heterogeneity

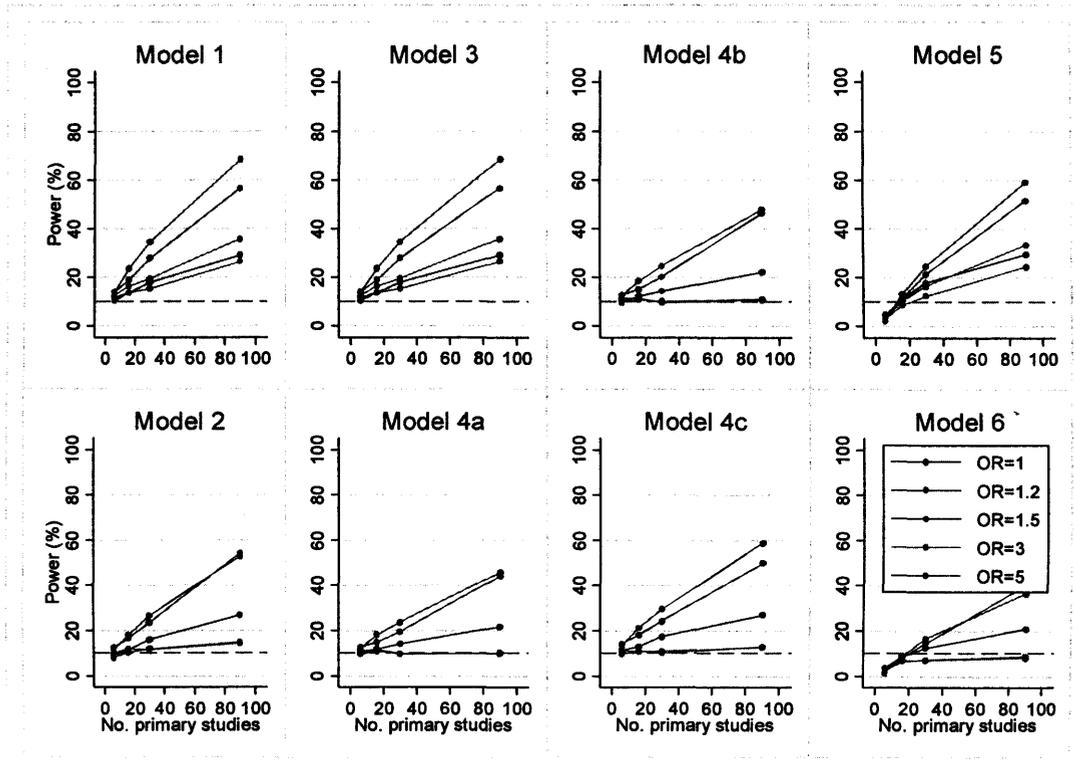
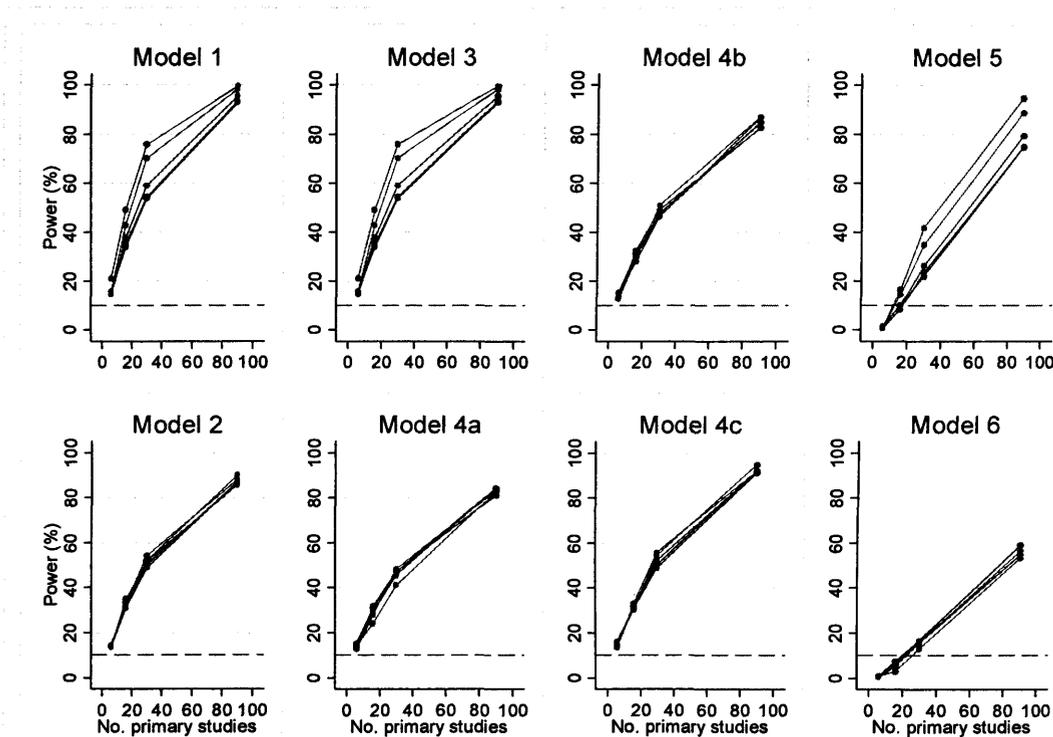


Figure 7.12 Power to detect 'severe' publication bias induced by effect size when there is no between-study heterogeneity (see Figure 7.11 above for legend)



None of the models with appropriate type I error rates have power to detect ‘severe’ publication bias, regardless of how it is induced, when there is a great deal of between-study heterogeneity. The same general trend in power is seen when ‘moderate’ publication bias is induced, but the levels of power are smaller (Figures I.2 and I.3 in Appendix I).

P-values obtained from the permutation test

In Figure 7.13, the type I error rates from all eight regression models are shown for the case when there is no between-study heterogeneity. Type I error rates of Egger’s regression test exceed the expected 10% level and increase as the underlying ORs and the number of studies in the meta-analysis increases. Type I error rates for Model 4c are higher than the expected 10% level for all underlying ORs and the number of studies in the meta-analysis. Model 6, the random effects regression model using sample size is the only model seen to have the expected type I error rates regardless of OR size or meta-analysis size, when the permutation test is used to calculate the p-value. However, there is an imbalance in the tail probability areas for Model 6 (see Figure I.4 in Appendix I).

In the presence of substantial between-study heterogeneity, Models 4a and 4b, in addition to Model 6, are the only ones to have the appropriate type I error rates. Those from Egger’s regression test (Model 1) well exceed 10%, especially when $OR \geq 3$ (Figure 7.14).

Figure 7.13 Type I error rates of the eight regression models when the permutation test is used and there is no between-study heterogeneity

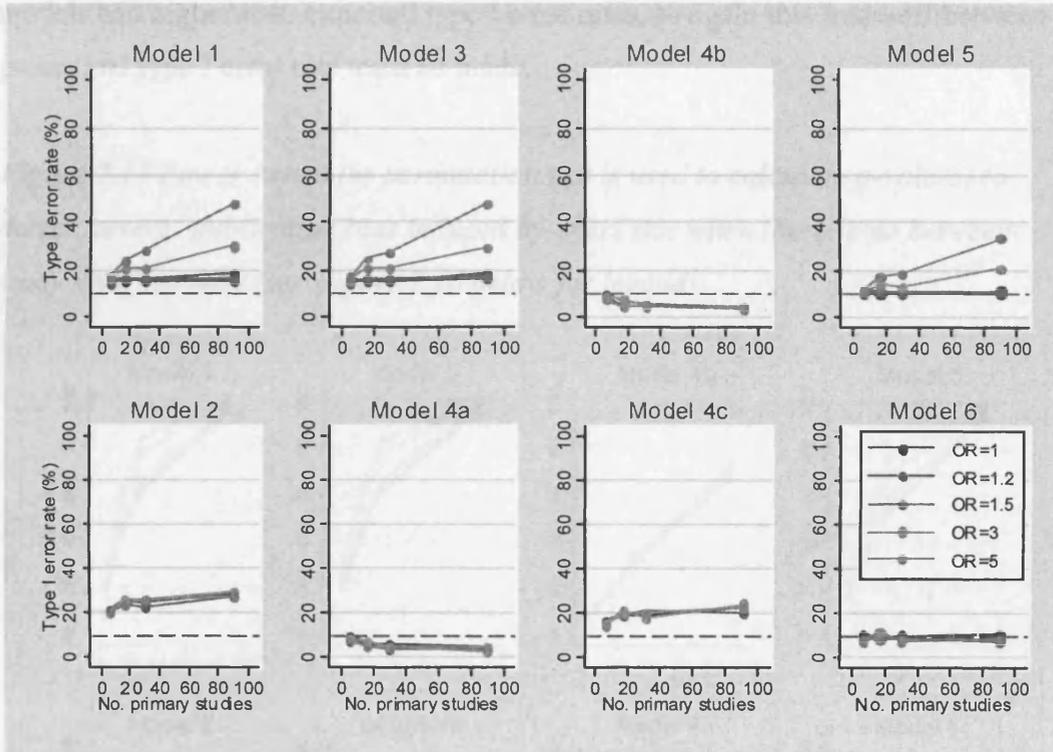
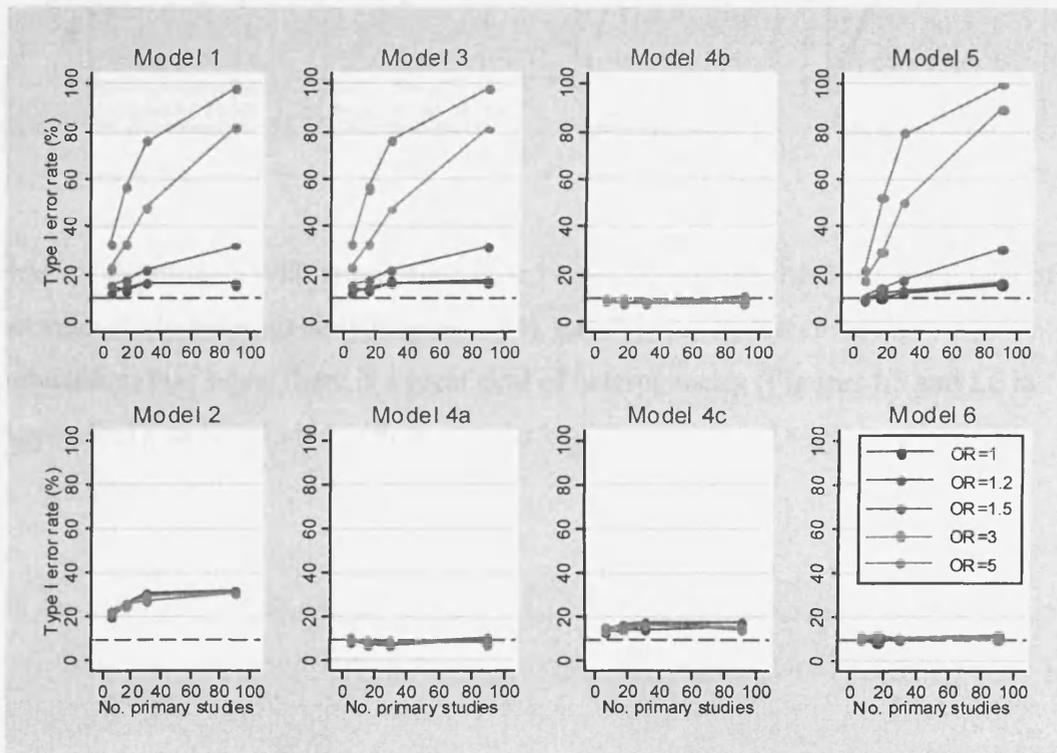
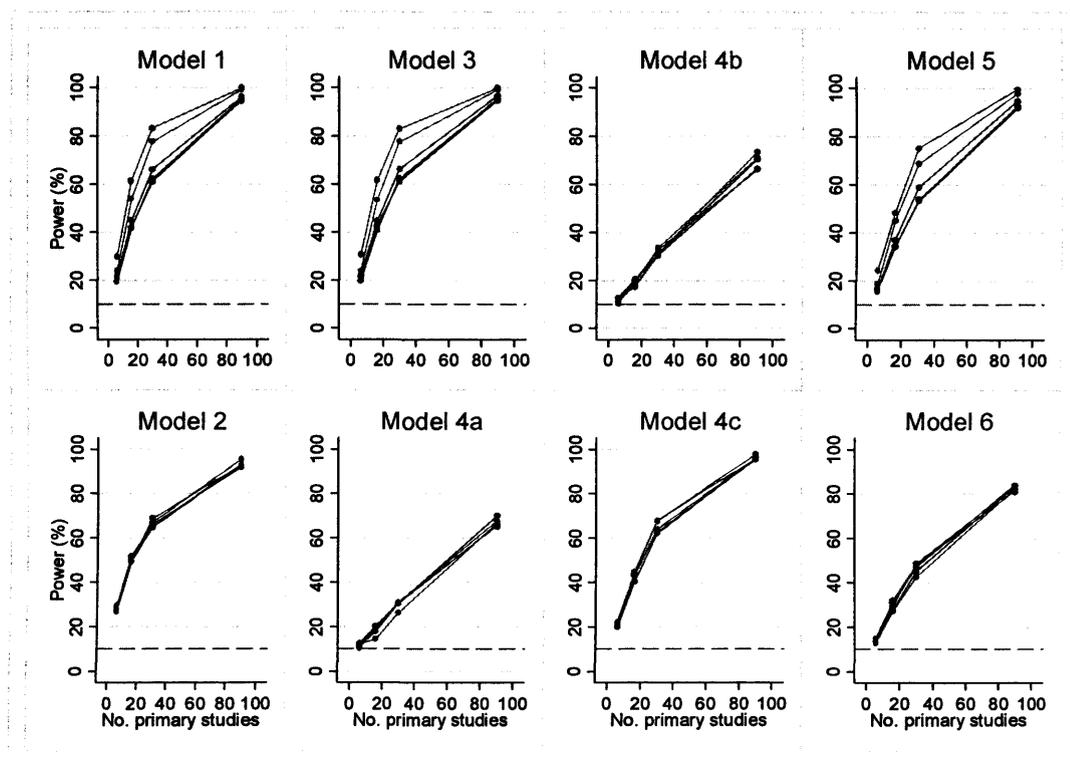


Figure 7.14 Type I error rates of the eight regression models when the permutation test is used and between-study heterogeneity is 500% of the within-study variation



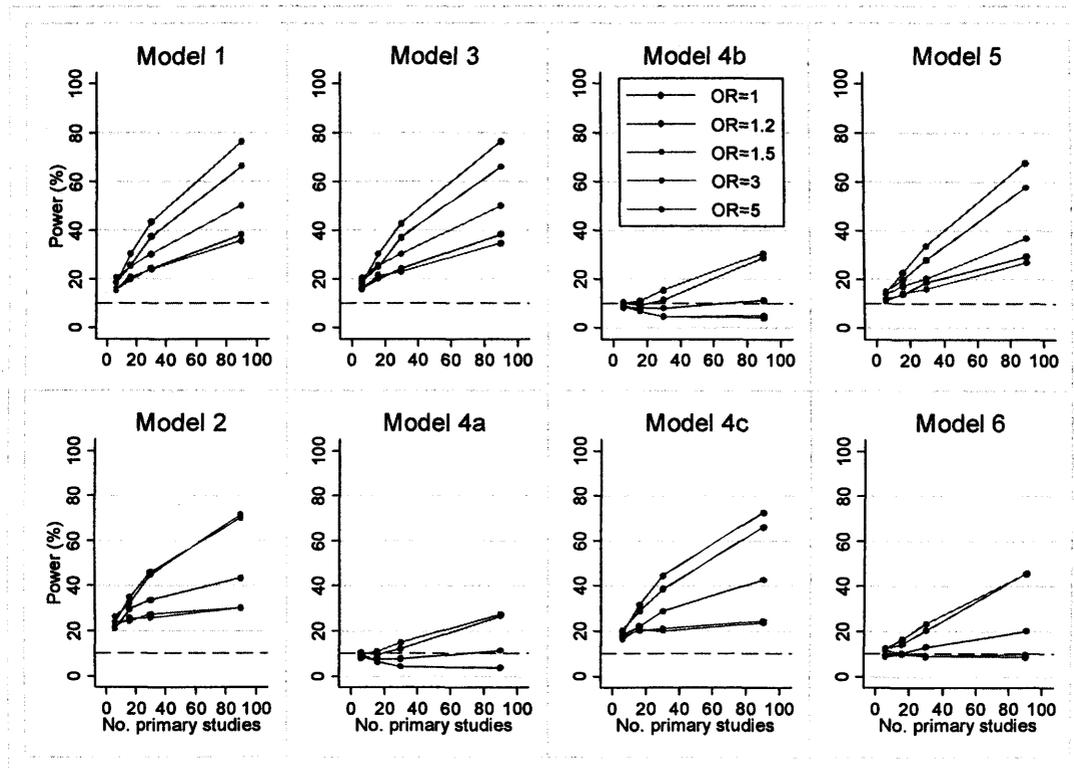
Models 1, 2, 3, 4c and 5 all have higher levels of power to detect 'severe' publication bias than Models 4a, 4b and 6 (see Figures 7.15 and 7.16), but these models had higher than expected type I error rates, so again this trade-off between power and type I error rate must be made.

Figure 7.15 Power (when the permutation test is used to calculate p-values) to detect 'severe' publication bias induced by effect size when there is no between-study heterogeneity (see Figure 7.16 below for legend)



None of the models with appropriate type I error rates when there is a great deal of between-study heterogeneity (Figure 7.14), have power to detect 'severe' publication bias when there is a great deal of heterogeneity (Figures I.5 and I.6 in Appendix I); as concluded with the results based on the usual t-test.

Figure 7.16 Power of the eight regression models to detect 'severe' publication bias induced by p-value when there is no between-study heterogeneity



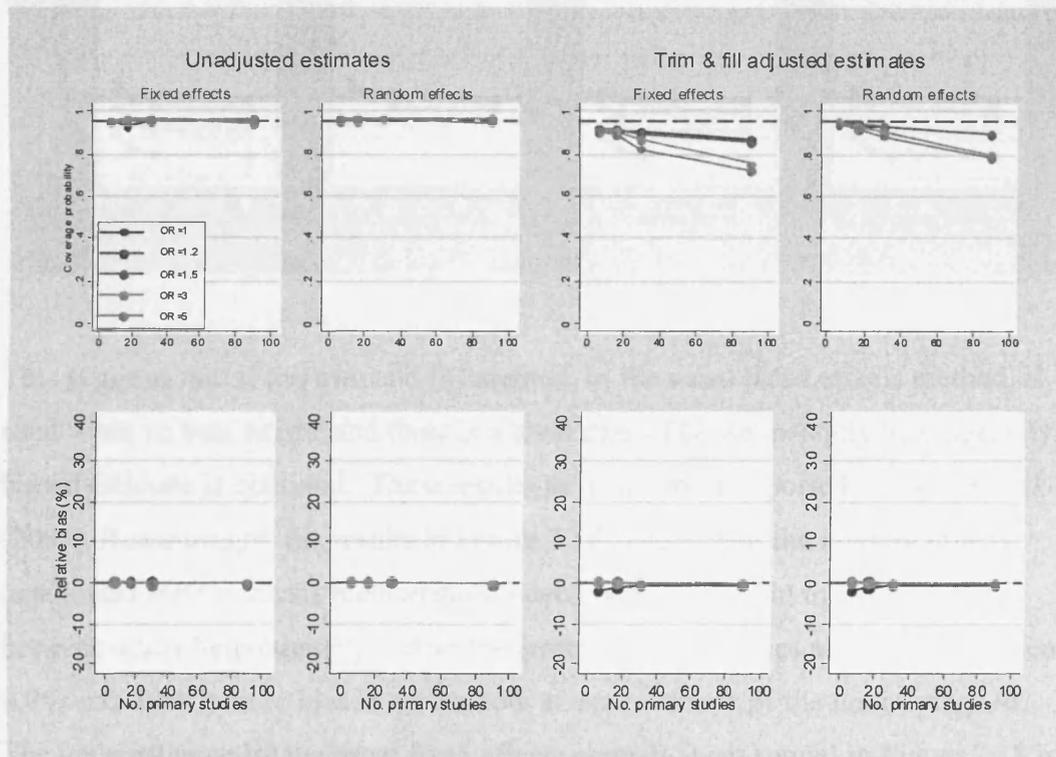
In summary Model 4c appears superior to both Egger's regression test (Model 1) and Macaskill's regression test (Model 4b) when the usual t-test is used for calculation of p-values. Apart from lower than expected type I error rates when there is no between-study heterogeneity, Model 6 also seems to perform reasonably well. More so when the permutation test is used to calculate the p-values. None of the tests (whether using the t-test or permutation test to calculate p-values) performs well to detect publication bias when a great deal of unexplained between-study heterogeneity is present.

7.6.3 Estimates of effect – the trim and fill method

In the absence of publication bias, the trim and fill method has slightly lower coverage probabilities than expected, i.e. the underlying OR lies within the 95% CI of the estimated OR fewer times than expected (top row of Figure 7.17). There is very little relative bias in the (fixed and random effects) trim and fill adjusted estimate of the underlying OR (bottom row of Figure 7.17). This suggests that although the underlying OR lies outside the 95% confidence interval (calculated

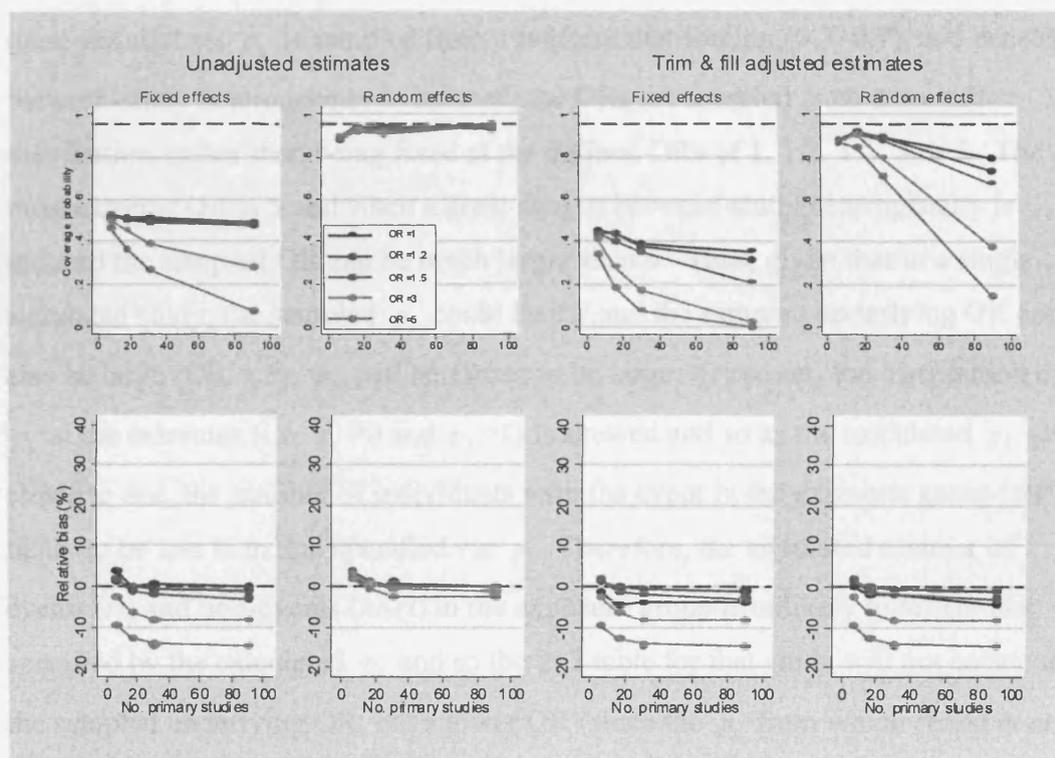
from the trim and fill method) more often than expected, on average, the fixed and random effects trim and fill estimate is similar to the underlying OR when there is no publication bias. Therefore indicating that the 95% CIs from the trim and fill method are too narrow. The corresponding coverage probabilities and relative bias in the usual (unadjusted) inverse-variance weighted estimates (from Equations 3.1 and 3.2) are also given in Figure 7.17. There is very little departure from the expected 0.95 coverage probabilities and 0% relative bias for both the fixed and random effects model.

Figure 7.17 Coverage probabilities (top row) and relative bias (bottom row) of methods for pooling estimates in the absence of publication bias and between-study heterogeneity



As between-study heterogeneity increases, the coverage probabilities for the fixed and random effects trim and fill estimates and the usual fixed effects unadjusted estimates do not perform well (top row Figure 7.18). As the underlying OR increase, these estimates increasingly *under-estimate* the underlying OR (bottom row Figure 7.18).

Figure 7.18 Coverage probabilities (top row) and relative bias (bottom row) of methods for pooling estimates in the absence of publication bias, when between-study heterogeneity is 500% of the within-study heterogeneity



This suggests that if the trim and fill method, or the usual fixed effects method, is used when no bias exists, and there is a great deal of between-study heterogeneity, a biased estimate is obtained. These results have also been reported by Terrin *et al.* (2003). Reassuringly, the results in Figure 7.18 suggest that the random effects unadjusted meta-analysis method does indeed perform well in the presence of between-study heterogeneity, where the coverage probabilities are as expected (i.e. 0.95) and there is little bias in the random effects estimate of the underlying OR. The underestimate by the usual fixed effects meta-analysis model in Figure 7.18 is an interesting finding. This underestimate is not seen for i) ORs closer to the null, ii) when there is no between-study heterogeneity (see Figure 7.17) or iii) when the random effects meta-analysis model is used. This last point suggests that it is the most precise studies that are being underestimated since more relative weighting is given to these studies in a fixed effects meta-analysis compared to a random effects meta-analysis. This leads to discussion as to why studies are being underestimated when the OR is very large and there is a great deal of between-study heterogeneity.

It is likely to be a factor in how the values in the 2x2 tables for the primary studies are generated in this simulation analysis. The exposure group event rate, p_i , is calculated by fixing the control group event rate, p_c , and the underlying OR. In these simulations p_c is sampled from a uniform distribution (0.3, 0.7), and because between-study heterogeneity is induced, the ORs are sampled from a normal distribution, rather than being fixed at the defined ORs of 1, 1.2, 1.5, 3 or 5. The most extreme OR is 5 and when a great deal of between-study heterogeneity is induced the sampled OR can be much larger than 5. Thus, given that in a single simulated study, the sampled p_c could be 0.7 and the sampled underlying OR could also be large (OR > 5), p_i will be forced to be large. However, the distribution of p_i at the extremes (i.e. $p_i=0$ and $p_i=1$) is skewed and so as the calculated p_i gets closer to one, the number of individuals with the event in the exposure group (rt) is likely to be less than that specified via p_i . Therefore, the simulated number of events (rt) and non-events ($nt-rt$) in the exposure group is unlikely to reflect that specified by the calculated p_i and so the 2x2 table for that study will not equate to the sampled underlying OR, but a lower OR (since the p_i from which rt and $nt-rt$ have been sampled is smaller than the p_i calculated via the sampled p_c and OR). For instance for Study A in Figure 7.19, p_i is 0.92 with the OR calculated to be 5, however if one assumes that the calculated p_i was in fact 0.97, for instance, (but the skew in the distribution of p_i meant a lower value of p_i , 0.92, was sampled) it can be seen how the OR is hugely underestimated, i.e. OR = 5 in Study A vs OR = 14 in Study B.

The studies where the underlying OR has been underestimated are likely to be more precise for the reason discussed earlier in this chapter (Section 7.3, page 173): the correlation between the OR and its standard error (Macaskill et al., 2001). This is seen in an extreme case in Figure 7.20 where a meta-analysis is simulated to have 90 studies (all having 400 subjects, 200 in each arm). The underlying $\ln(\text{OR})$ in each study is drawn from $N(1.61, 0.45)$, such that the mean OR is 5 with 95% CI 1.34, 18.62. In this extreme case it can be seen that studies with large $\ln(\text{ORs})$ are estimated with a great deal of variability and studies with small $\ln(\text{ORs})$ are estimated very precisely.

Figure 7.19 Hypothetical studies

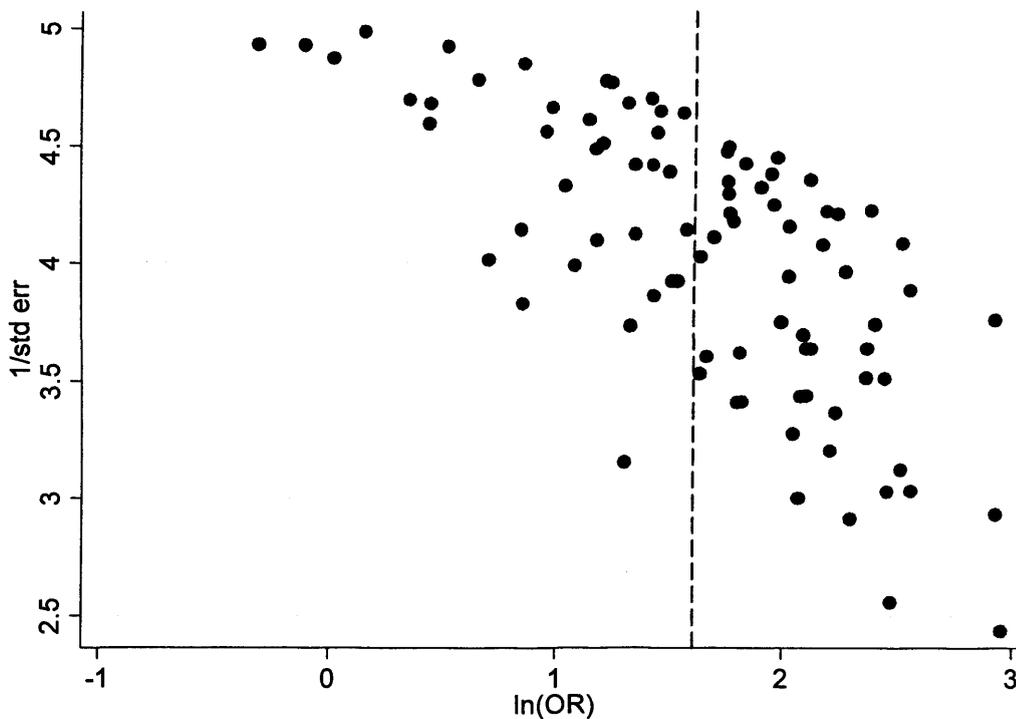
Study A

	Control group	Exposure group	OR = 5 ln(OR) = 1.61
Event	109	146	variance of ln(OR) = $\frac{1}{109} + \frac{1}{50} + \frac{1}{146} + \frac{1}{13} = 0.11$
No event	50	13	
	159	159	pc = 0.685 pt = 0.92

Study B

	Control group	Exposure group	OR = 14 ln(OR) = 2.65
Event	109	154	variance of ln(OR) = $\frac{1}{109} + \frac{1}{50} + \frac{1}{154} + \frac{1}{5} = 0.24$
No event	50	5	
	159	159	pc = 0.685 pt = 0.97

Figure 7.20 Funnel plot of meta-analysis simulated under extreme conditions



The pooled OR from the fixed effects model is 4.5 (95% CI: 4.27, 4.73) – an underestimate, while the pooled OR from the random effects model is closer to that defined (5.04 with 95% CI: 4.36, 5.82) since studies are given relatively more equal weighting in a random effects model than in a fixed effects model. When the trim and fill model is applied to meta-analyses simulated from such extreme scenarios underestimates are also obtained from both the fixed effects and random effects trim and fill models as shown in Figure 7.18. The trim and fill adjusted estimates for the extreme meta-analysis simulated above (where all sample sizes are equal) are OR = 3.52 (95% CI 3.39, 3.67) for the fixed effects trim and fill model and OR = 4.10 (95% CI 3.56, 4.76) for the random effects trim and fill mode.

An alternative approach to simulating the binary data could help to alleviate some of this bias. For instance, Schwarzer et al. (2002) do not fix the control group event rate, p_c , instead they fix p_A such that

$$\log it(p_c) = \log it(p_A) - \frac{1}{2} \ln(OR)$$

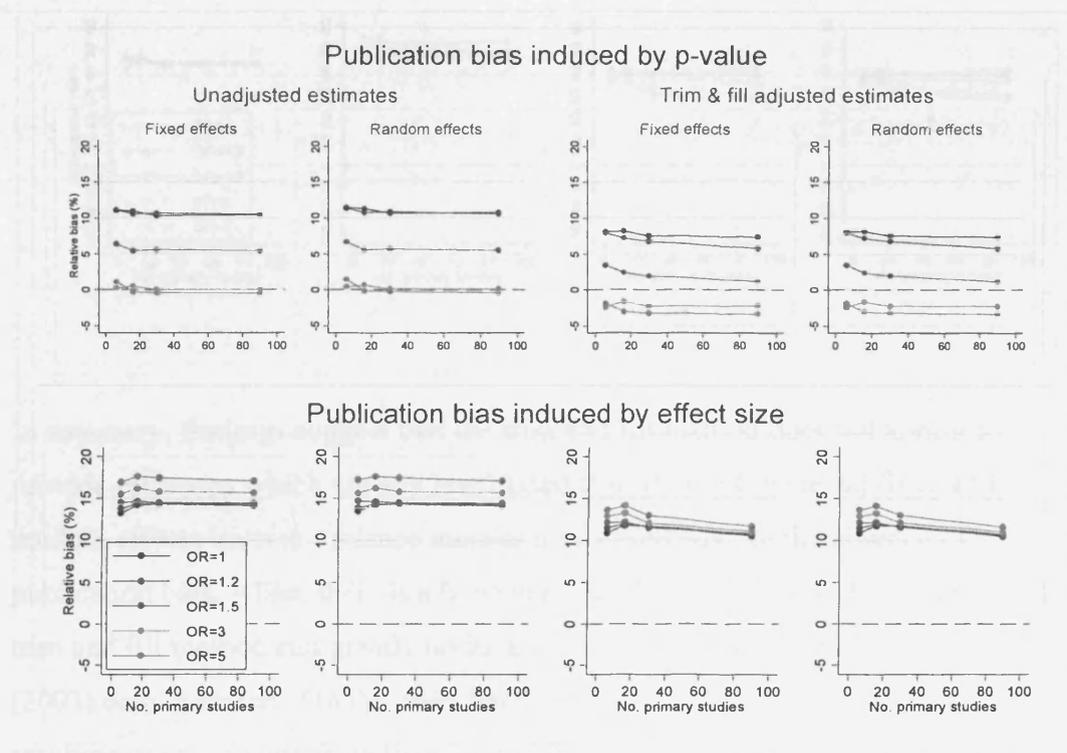
$$\log it(p_t) = \log it(p_A) + \frac{1}{2} \ln(OR)$$

where OR is the underlying OR. Using this approach less extreme values for p_t may be obtained that do not force the OR in a study to be an underestimate of the sampled OR.

The method used in these simulations to calculate the individual level data within a study (i.e. by fixing the control group event rate and the underlying OR to then calculate the exposure group event rate) is the standard approach taken in simulations of binary data (Sterne et al., 2000; Macaskill et al., 2001; Terrin et al., 2003; Harbord et al., 2006). However, as discussed above the results suggest that this may lead to a bias in meta-analyses of large ORs when there is a great deal of between-study heterogeneity. Clearly this bias needs to be investigated further than has been discussed above, with more simulations being required to assess the bias and factors contributing to it. This is, however, beyond the scope of this thesis, although is likely to form further work in this area.

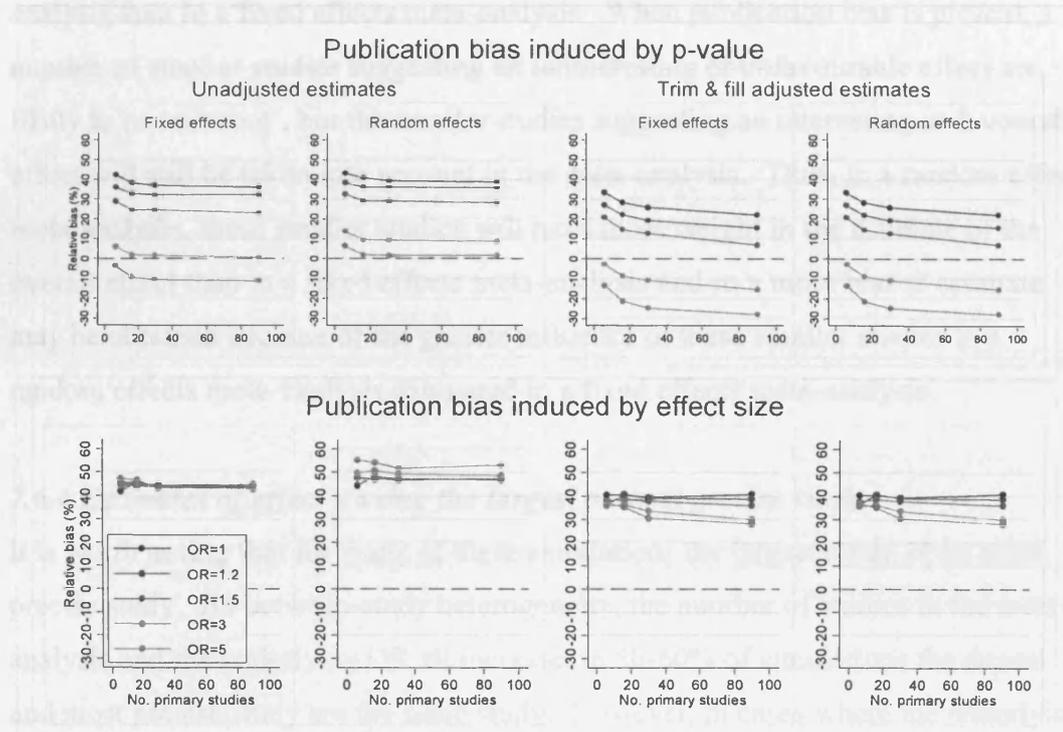
In the presence of 'severe' publication bias and no between-study heterogeneity, the fixed and random effects trim and fill method over-estimates the underlying OR by up to 10% when bias is induced by p-value, and up to 15% when induced by effect size (see Figure 7.21). (Note that the 0%, -5% relative bias seen in Figure 7.21 for the larger underlying ORs when publication bias is induced by p-value is an artefact, since little publication bias has been induced as discussed in Sections 7.5.1 and 7.5.3.) The trim and fill method appears to perform slightly better than the unadjusted method since less biased estimates are obtained.

Figure 7.21 Relative bias of four methods in estimating the true OR when there is no between-study heterogeneity in the presence of severe publication bias



As between-study heterogeneity increases, these estimates become more biased. When there is 500% between-study heterogeneity, all four methods give an over-estimate of the true OR of approximately 30-50% (Figure 7.22).

Figure 7.22 Relative bias of four methods in estimating the true OR in the presence of severe publication bias and a great deal of between-study heterogeneity (500% of the within-study heterogeneity)



In summary, findings suggest that the trim and fill method does not appear to provide estimates which are any less biased than those of the usual fixed and random effects inverse-variance meta-analysis methods. In the absence of publication bias, where there is a large amount of between-study heterogeneity, the trim and fill method can greatly under-estimate the underlying OR as Terrin *et al.* (2003) demonstrated. On the other hand, when publication bias is present, the resulting over-estimate from the trim and fill method can also be misleading in many situations.

Related to this is the point that use of a random effects meta-analysis rather than a fixed effects meta-analysis in the presence of between-study heterogeneity can give a more biased estimate of the effect when publication bias is present. This is seen to some extent in Figure 7.22 where the unadjusted random effects estimates are slightly more biased than the unadjusted fixed effects estimates. The reason for this is that when a random effects meta-analysis is used, because of the addition of τ^2 to

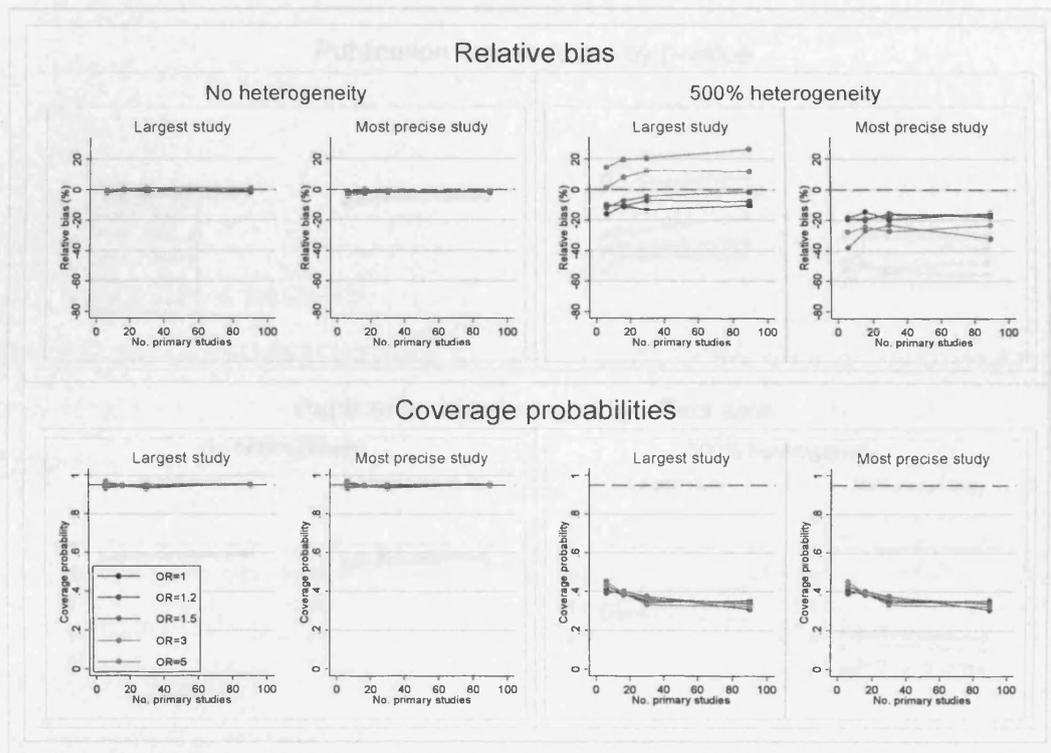
the calculation of the weighting given to each study (see Equation 3.2), individual study weightings become smaller than if a fixed effects meta-analysis is used. In particular, smaller studies get relatively more weight in a random effects meta-analysis than in a fixed effects meta-analysis. When publication bias is present, a number of smaller studies suggesting an *uninteresting* or *unfavourable* effect are likely to be 'missing', but the smaller studies suggesting an interesting or favourable effect will still be taken into account in the meta-analysis. Thus, in a random effects meta-analysis, these smaller studies will have more weight in the estimate of the overall effect than in a fixed effects meta-analysis and so a more biased estimate may be obtained because of the greater influence of these smaller studies in a random effects meta-analysis compared to a fixed effects meta-analysis.

7.6.4 Estimates of effect – using the largest or most precise study

It is worth noting that for many of these simulations the largest study *is* the most precise study. As between-study heterogeneity, the number of studies in the meta-analysis and the underlying OR all increase, in 50-60% of simulations the largest and most precise study are the same study. However, in cases where the underlying OR is close to the null, the number of studies in the meta-analysis is small and there is no between-study heterogeneity, the largest study and the most precise study are the same for 80-90% of the simulations. Hence, in these situations one would expect little difference in performance where the largest study or the most precise study was used, since they are the same studies.

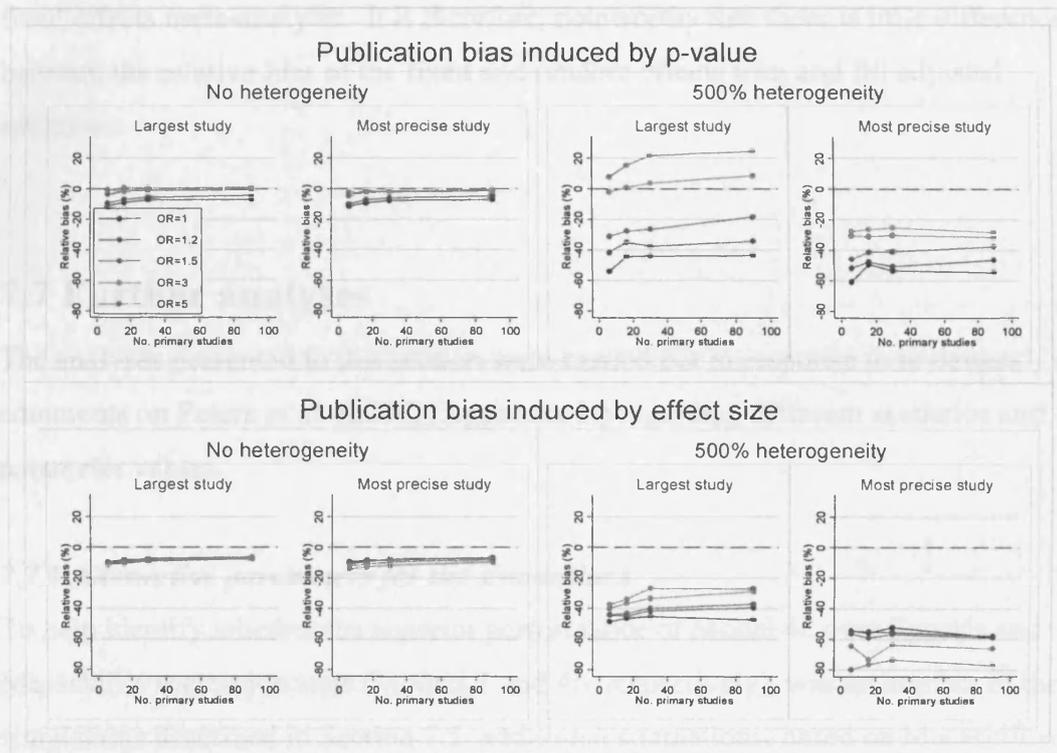
When there is no publication bias or between-study heterogeneity, taking the estimate from the largest study or the most precise study gives very good estimates of the underlying effect in terms of the relative bias and the coverage probabilities. However, when there is a great deal of heterogeneity between studies (500% of the within-study variance) these approaches tend to give biased estimates of the underlying effect (Figure 7.23).

Figure 7.23 Relative bias and coverage probabilities of using the observed effect from the largest study and the most precise study to estimate the underlying effect when there is no publication bias



In the presence of ‘severe’ publication bias, estimates from the largest or most precise study are biased by about 10% of the underlying effect (Figure 7.24). When there is a great deal of between-study heterogeneity this increases up to 80% when the estimate from the study with the largest sample size is used and up to 50% when the estimate from the most precise study is used (Figure 7.24).

Figure 7.24 Relative bias and coverage probabilities of using the observed effect from the largest study and the most precise study to estimate the underlying effect when publication bias is induced by p-value or effect size



7.6.5 Estimates of effect – summary

In the presence of simulated ‘severe’ publication bias all approaches assessed here – the trim and fill method, taking estimates from the largest study and the most precise study – do not perform particularly well when used to estimate the underlying effect. These findings confirm those in Terrin *et al.* (2003) that when the trim and fill method is applied to a meta-analysis with no publication bias, under-estimates of effect are obtained when between-study heterogeneity exists. However, in the presence of publication bias, regardless of the amount of between-study heterogeneity, the fixed and random effects trim and fill estimates generally give the least biased estimates in these simulation analyses compared to the usual fixed and random effects meta-analysis models and use of either estimate from the largest study or the most precise study. Compared to the fixed effects trim and fill method, the random effects trim and fill method gives coverage probabilities closer to the expected 0.95 level. However, the random effects meta-analysis tends to

have a wider 95% CI around the pooled estimate (since the between-study heterogeneity parameter is accounted for), thus there is more opportunity for the true effect to be included in the 95%CI from a random effects meta-analysis, than a fixed effects meta-analysis. It is therefore, noteworthy that there is little difference between the relative bias of the fixed and random effects trim and fill adjusted estimates.

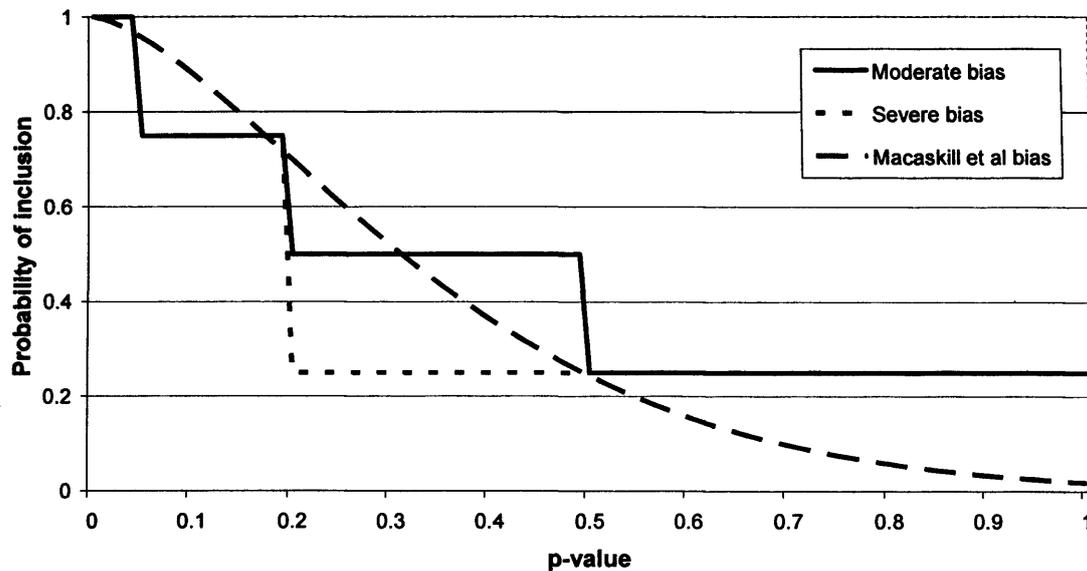
7.7 Further analyses

The analyses presented in this section were carried out in response to reviewers' comments on Peters *et al.* (2006) (Appendix H) regarding different scenarios and parameter values.

7.7.1 Alternative parameters for the simulations

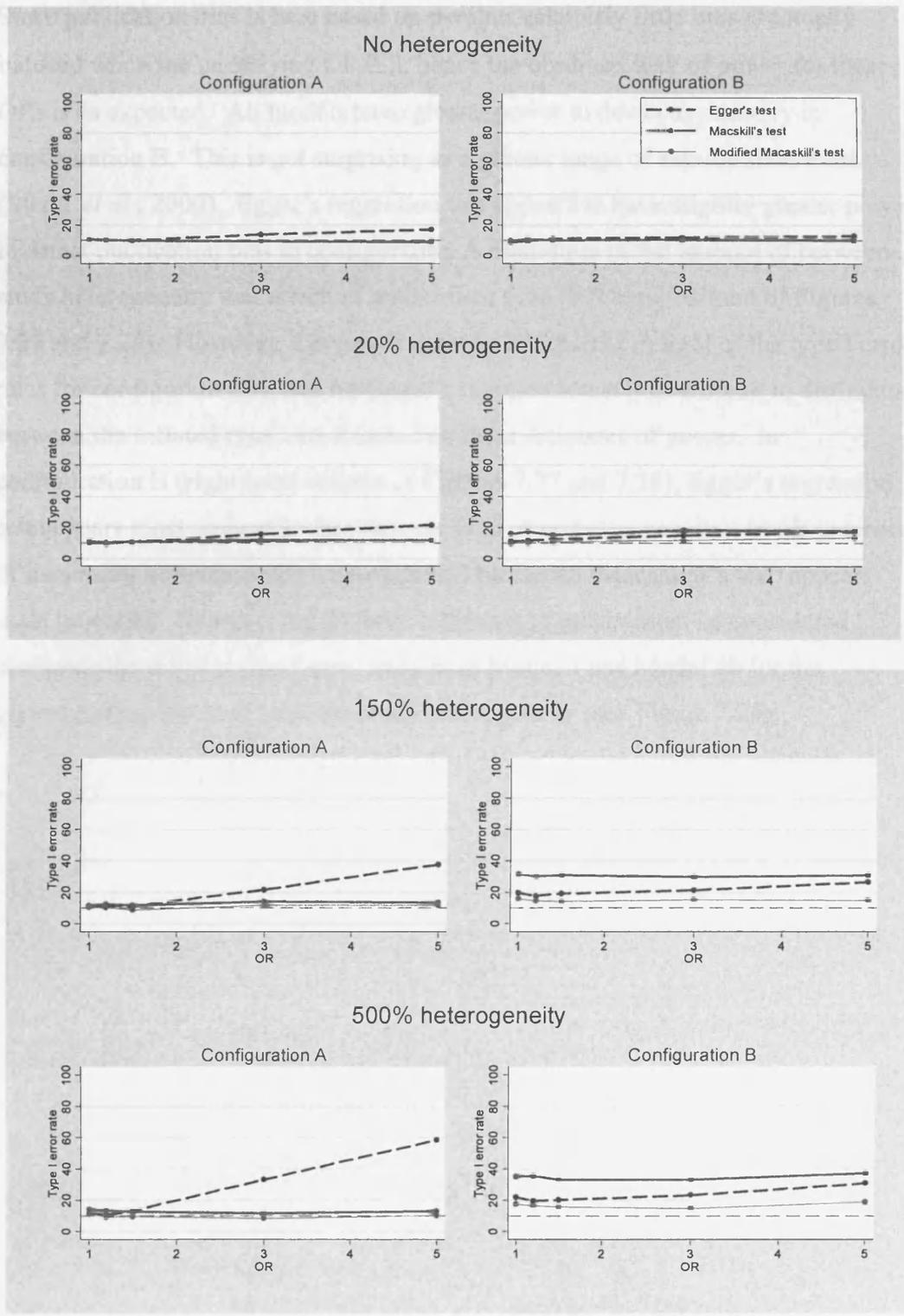
To help identify whether the superior performance of Model 4c over Egger's and Macaskill's regression tests (Models 1 and 4b, respectively), was an artefact of the simulations described in Section 7.5, additional simulations, based on Macaskill *et al.* (2001), were carried out. The first set of additional simulations (1) are based on the two primary study sample size configurations given in Macaskill *et al.* Both configurations, A and B, have 21 studies in the meta-analysis. Configuration A consists of 11 studies with 100 subjects per group, 6 with 200 subjects per group and 4 studies with 300 subjects per group. Configuration B has 10 studies with 100 subjects per group, 5 with 200 subjects per group, 3 with 300 subjects per group, 2 with 500 subjects per group and 1 study with 1000 subjects per group. Publication bias is defined as severe such that the probability of selection into the meta-analysis, $p(w)$, is 1 if the study's p-value is < 0.05 ; $p(w) = 0.75$ if $0.05 < \text{p-value} < 0.2$; $p(w) = 0.25$ if $\text{p-value} > 0.2$.

The second set of simulations (2) have the sample size configurations A and B above and the selection mechanism $p(w) = \exp(-4 * (\text{p-value}^{1.5}))$ as used in Macaskill *et al.* The different p-value based selection mechanisms used are shown in Figure 7.25.

Figure 7.25 '*p-value based*' selection mechanisms used

Simulations (1) and (2) were carried out under conditions of 0%, 20%, 150% and 500% between-study heterogeneity as defined in Section 7.5.1. In the absence of publication bias, Model 4c appears to have better performance in terms of the appropriate type I error rates, regardless of the amount of heterogeneity, the size of the underlying OR and the sample size configuration, over Models 1 and 4b (Figure 7.26 – note that the format of these figures is different to those before: the x-axis represents the size of the underlying OR and each line represents the type I error rate or power of Egger's regression test (Model 1), Macaskill's regression test (Model 4b) and the modified Macaskill's test (Model 4c)).

Figure 7.26 Type I error rates of Models 1, 4b and 4c in sample size configurations A and B (number of primary studies in meta-analysis is 21 in each simulation)



In the presence of publication bias, the findings are slightly harder to interpret (Figure 7.27 for simulation (1) results and Figure 7.28 for simulation (2) results). Since publication bias is here based on p-value, relatively little bias is actually induced when the underlying $OR \geq 3$, hence the observed lack of power for these ORs is as expected. All models have greater power to detect asymmetry in configuration B. This is not surprising as a greater range of sample sizes exist (Sterne *et al.*, 2000). Egger's regression test appears to have slightly greater power to detect publication bias in configuration A regardless of the amount of between-study heterogeneity and selection mechanism used (left hand column of Figures 7.27 and 7.28). However, this power must be interpreted in light of the type I error rates for configuration A, and for Egger's regression test it is difficult to distinguish between the inflated type I error rates and these estimates of power. In configuration B (right hand column of Figures 7.27 and 7.28), Egger's regression test appears most powerful when there is little or no heterogeneity. In the presence of increasing between-study heterogeneity, Model 4b (Macaskill's test) appears most powerful. However, again these estimates of power must be considered alongside the inflated type I error rates from Model 1 and Model 4b for the corresponding levels of between-study heterogeneity (see Figure 7.26).

Figure 7.27 Power of Models 1, 4b and 4c to detect publication bias in sample size configurations A and B for simulation set (1) (number of primary studies in meta-analysis is 21 in each simulation)

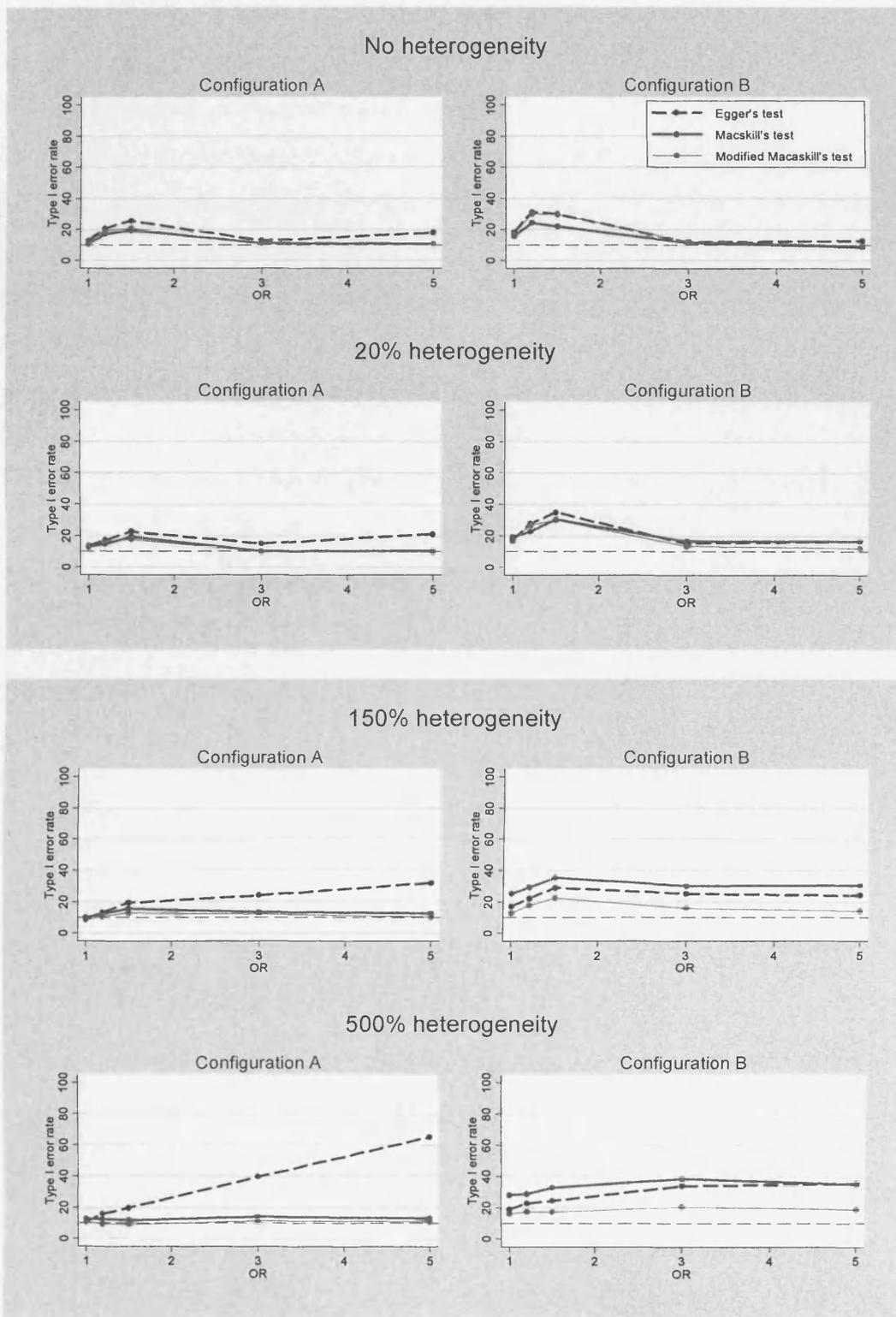
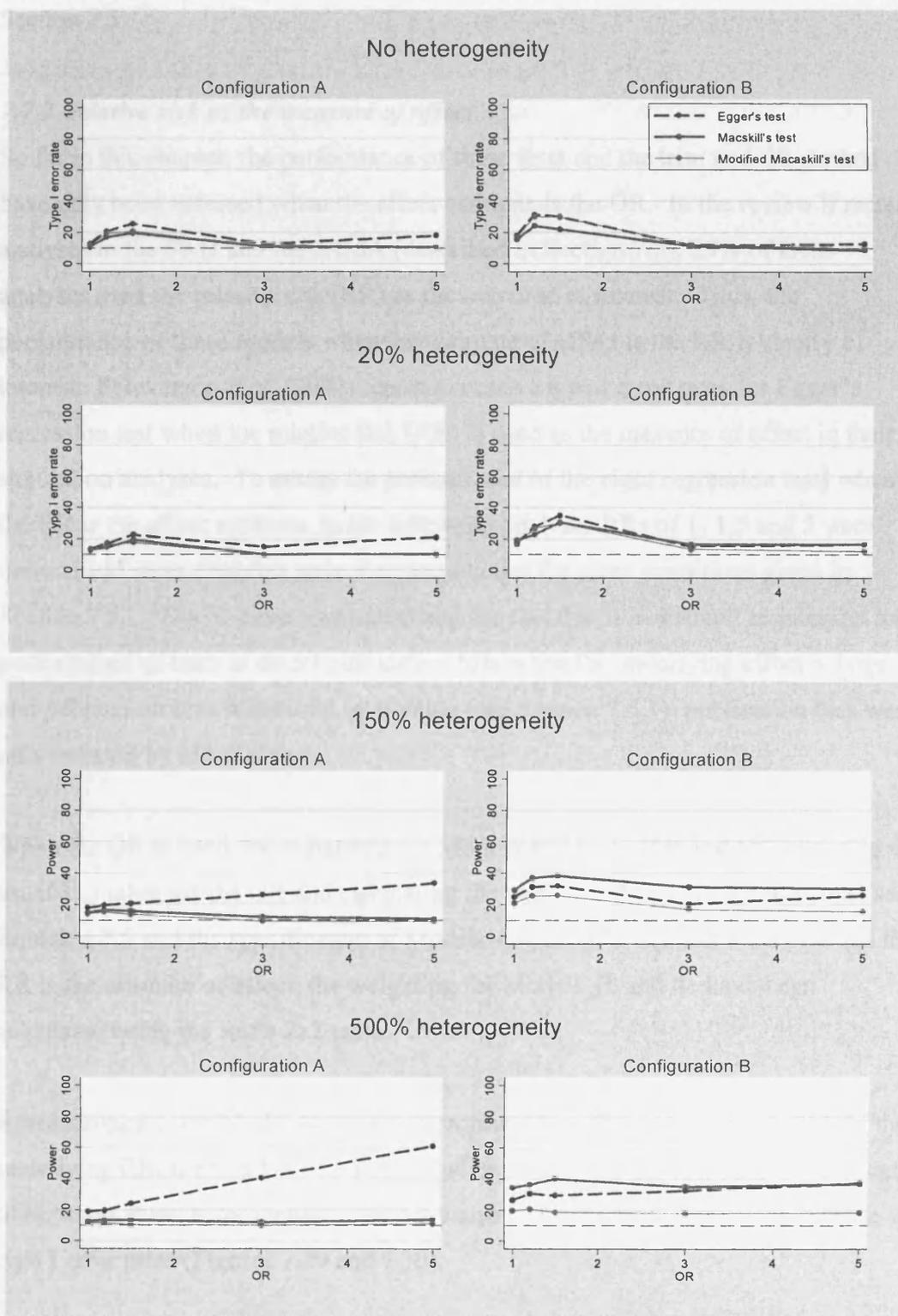


Figure 7.28 Power of Models 1, 4b and 4c to detect publication bias in sample size configurations A and B for simulation set (2) (number of primary studies in meta-analysis is 21 in each simulation)



These findings suggest that the superiority of Model 4c, in terms of its appropriate type I error rates and reasonable power relative to Models 1 and 4b, is not an artefact of the study size design defined in the initial set of simulations described in Section 7.5.1.

7.7.2 Relative risk as the measure of effect

So far in this chapter, the performance of these tests and the trim and fill method have only been assessed when the effect estimate is the OR. In the review of meta-analyses in the BMJ and the JAMA (described in Section 7.3), 29% of meta-analyses used the relative risk (RR) as the outcome of interest. Thus, the performance of these models when the estimate of effect is the RR is clearly of interest. Schwarzer *et al.* (2002) report excessive type I error rates for Egger's regression test when the relative risk (RR) is used as the measure of effect in their simulation analyses. To assess the performance of the eight regression tests when the RR is the effect estimate, in the following analyses RRs of 1, 1.5 and 5 were defined and meta-analyses were simulated under the same conditions given in Section 7.5.1. Due to time constraints and the fact that it is difficult to interpret the performance of tests to detect publication bias when the underlying effect is large and publication bias is induced by p-value (see Section 7.5.1), publication bias was only induced by effect size.

When the OR is used, the weighting for Models 4b and 4c is based on collapsing the usual 2x2 table for the OR and calculating the variance of this pooled log odds (see Equation 7.5 and the specification of Models 4b and 4c in Section 7.5.2). When the RR is the estimate of effect, the weighting for Models 4b and 4c have been calculated using the same 2x2 table.

Results suggest that Model 4c attains appropriate type I error rates regardless of the underlying RR, the number of primary studies in the meta-analysis and the amount of between-study heterogeneity. Model 6 also performs reasonably well in terms of type I error rates (Figures 7.29 and 7.30).

In the presence of 'severe' publication bias (induced by effect size), all models have reasonable power to detect the asymmetry (Figures 7.31 and 7.32), although as with previous results, power and type I error rates must be considered together.

RRs and the variance of the RR are correlated (as with ORs) and this is seen in these plots, since the sample size models (Models 4a, 4b, 4c and 6) perform better in terms of obtaining the appropriate type I error rates. In conclusion, model 4c appears to out-perform the other models when the effect is a RR as it does with ORs. However, further confirmatory work is required, particularly with the specification of the weighting for Models 4b and 4c, and an exploration of the performance of these models when other measures of effect are the summary estimates (such as risk difference) is needed.

Figure 7.29 Type I error rates when RRs are the measures of effect and there is no publication bias or between-study heterogeneity

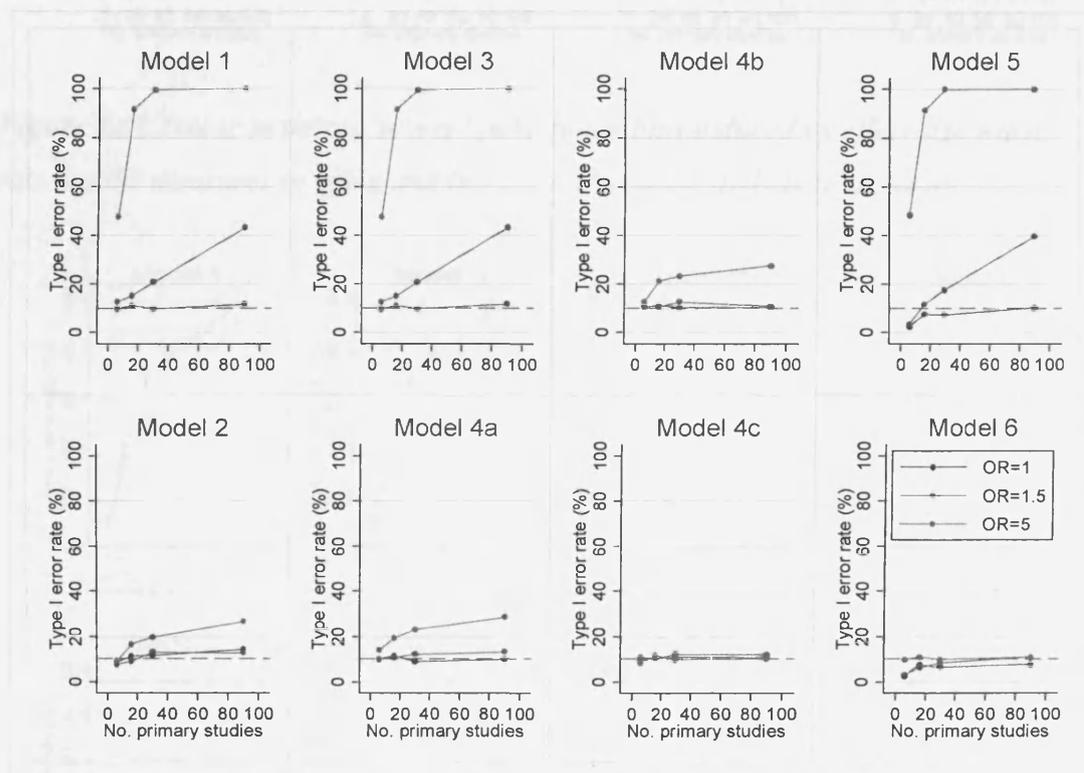


Figure 7.30 Type I error rates when RRs are the measures of effect when there is no publication bias but 500% between-study heterogeneity

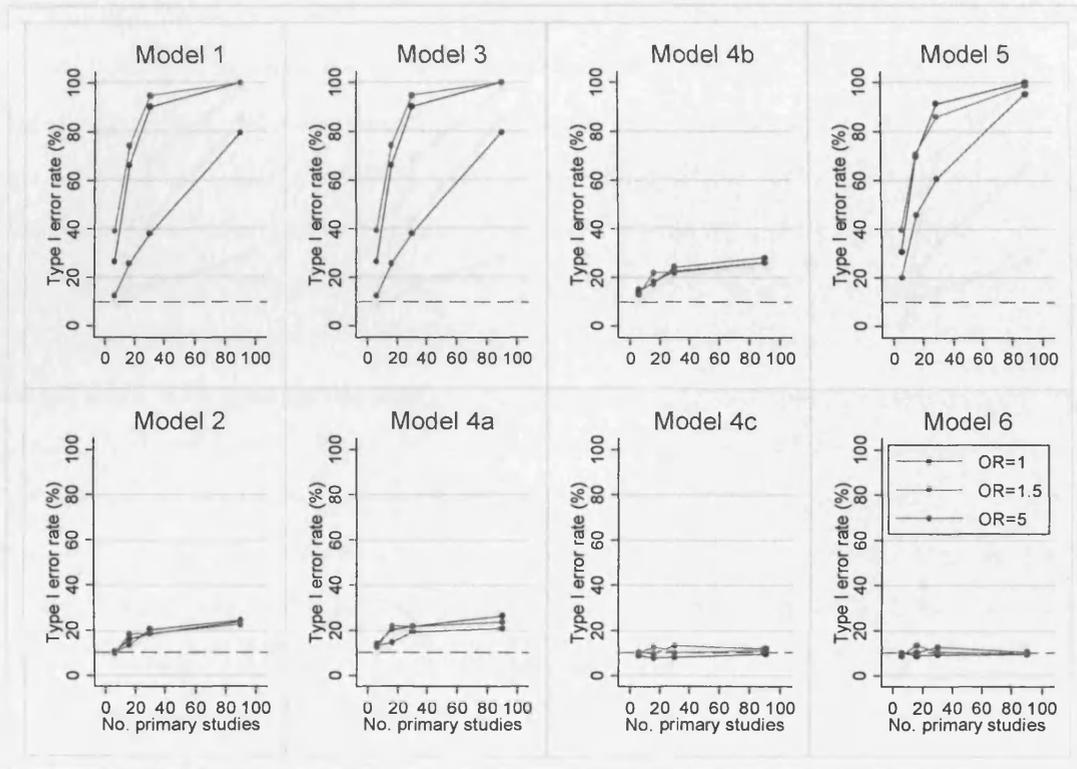


Figure 7.31 Power to detect 'severe' publication bias induced by effect size when RRs are the measures of effect and there is no between-study heterogeneity

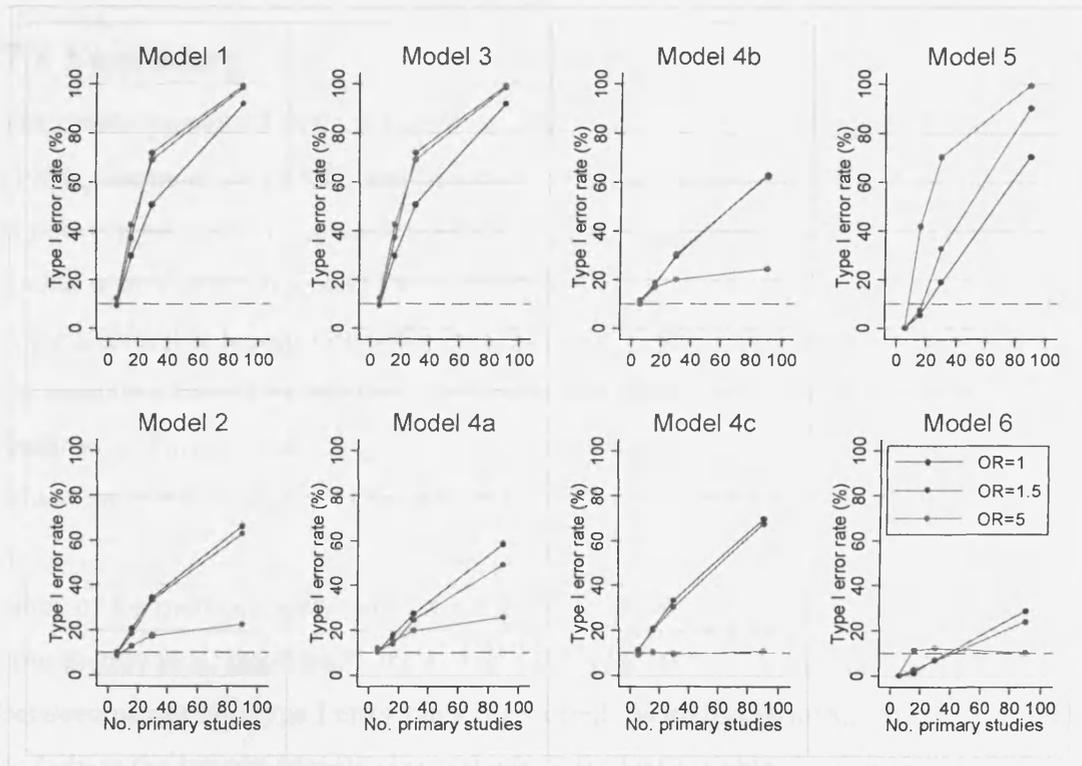
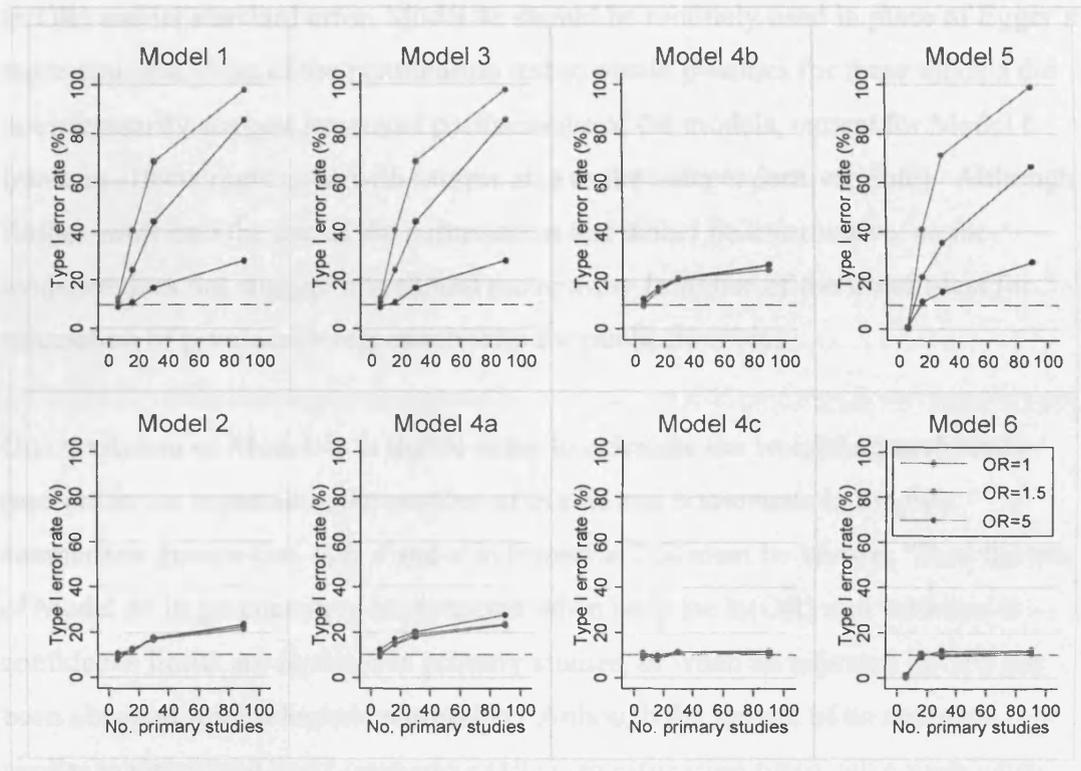


Figure 7.32 Power to detect 'severe' publication bias induced by effect size when RRs are the measures of effect and there is 500% between-study heterogeneity



7.8 Summary

The results presented in this chapter confirm 1) the findings of Begg and Mazumdar (1994), Sterne *et al.* (2000), and Macaskill *et al.* (2001) that the rank correlation test is not very powerful for meta-analyses with a small number of studies, 2) those of Sterne *et al.* (2000), Macaskill *et al.* (2001) and Schwarzer *et al.* (2002) that Egger's regression test is inappropriate for the detection of publication bias when ORs are the summary measures because of the excessive type I error rates and 3) the findings of Terrin *et al.* (2003) that the trim and fill method can inappropriately adjust for publication bias in the presence of between-study heterogeneity.

None of the methods assessed here to detect or adjust for publication bias has consistently performed well. Instead, for the publication bias tests, a trade-off between power and type I error rates is required. With this in mind, although Model 4c (where the inverse sample size is the independent variable) is no more powerful

than Egger's regression test (Model 1) and Macaskill's regression test (Model 4b), because of favourable type I error rates and a reduction in the correlation between $\ln(\text{OR})$ and its standard error, Model 4c should be routinely used in place of Egger's regression test. Use of the permutation test to obtain p-values for these models did not necessarily suggest improved performance of the models, except for Model 6 (random effects regression with sample size as the independent variable). Although further work into the use of the permutation test would be informative, so far evidence does not suggest one should move away from use of the usual t-test for calculation of p-values in regression tests for publication bias.

One limitation of Model 4c is that in order to calculate the weighting each study receives in the regression, the number of events and non-events in the two comparison groups (i.e. a , b , c and d in Equation 7.5) must be known. Thus the use of Model 4c in practice may be restricted when only the $\ln(\text{OR})$ and variance or confidence limits are reported in primary studies, or when an adjusted $\ln(\text{OR})$ has been obtained from a logistic regression. Although the impact of an approach similar to Greenland and Longnecker (1992) to estimating fitted cell counts when adjusted $\ln(\text{OR})$ s are reported, as described in Equation 5.7, could be investigated.

In general, the regression models including sample size as the independent variable were seen to be superior to those using standard error as the independent variable. This provides strong evidence for the use of sample size, rather than the inverse of standard error as an axis in the funnel plot. Sterne and Egger (2001) do not recommend use of sample size because of the difficulty in defining the shape of the funnel, however when the summary estimate and standard error are correlated (as for ORs), use of standard error as an axis on the funnel plot is likely to accentuate asymmetry. This is particularly likely when the underlying OR is believed to be far from the null.

Sensitivity analyses suggest that the performance of the regression tests when ORs are used is similar to that when RRs are used as the measures of effect. Again Egger's regression test has excessive type I errors, and although a trade-off is needed between these error rates and power, models using sample size generally

performed well in terms of appropriate type I error rates and reasonable levels of power.

Findings from the assessment of the trim and fill method are not particularly reassuring. In the absence of publication bias, when there is a large amount of between-study heterogeneity, the trim and fill method can under-estimate the underlying OR, as pointed out by Terrin *et al.* (2003). Similarly when publication bias is present, the resulting over-estimate from the trim and fill method can also be misleading. However, this approach was less biased in its estimate of the underlying effect than using the estimate from the largest or most precise study and the usual fixed and random effects unadjusted models. As is further discussed in Section 9.4, these methods do have a valuable role to play in assessing the sensitivity of the estimate of effect.

In these simulations, a relatively common event has been simulated, however, meta-analyses are generally also useful for rare effects and so investigation of this would be useful when recommended how to investigate publication bias. Evidence suggests the type I error rates for Egger's regression test are particularly high in these situations (Sterne *et al.*, 2000; Schwarzer *et al.*, 2002), but performance of all tests and the trim and fill method when the event is rare would be part of further work. A further limitation is that publication bias has been induced on the basis of either p-value or effect size, when in reality, it is likely to be a combination of these plus influences from other factors. However, the fact that these simulations suggested improved performance of Model 4c over all of the other tests regardless of whether publication bias was induced by p-value or effect size, gives confidence to the findings. Nevertheless, further simulations where publication bias is induced as a combination of p-value and effect size would be advantageous.

In the context of human health risk assessment these findings suggest that the usefulness of regression tests, in particular Model 4c, may be variable. For instance, the number of relevant studies to be synthesised in a risk assessment may differ; for example while Crump *et al.* (1999) estimate the proportion of liver carcinogens using almost 400 rat and mouse bioassays (see Appendix D), in the THMs example in Chapter 5 only 13 study results were synthesised, similarly Guth

et al. (1997) report the synthesis of just 12 experiments (see Appendix D). When the number of studies in the meta-analysis is small, the results in Section 7.6 suggest that none of the regression tests assessed have good power to detect publication bias. Furthermore, the relevant evidence included in a risk assessment may be quite heterogeneous due to species, exposure route and outcome differences, among others (as illustrated in Chapter 6 for the Mn example). In meta-analysis scenarios where unexplained between-study heterogeneity was present, the simulation analyses in this chapter found that all tests performed badly, regardless of the size of the underlying effect or the number of studies in the meta-analysis.

In some meta-analyses in human health risk assessments, and other contexts, between-study heterogeneity may be explained by a study-level covariate. For example, the different species or route of exposure may explain the between-study heterogeneity. In Chapter 4, only 14 of the reviewed meta-analyses (30%) did not report combining study estimates from different species. In the next chapter, the simulated analyses described in Section 7.5 are extended to include situations where between-study heterogeneity can be explained by a study-level covariate, a scenario likely to occur when different sources of evidence are combined for a human health risk assessment. These analyses will assess whether detection of publication bias in these scenarios is possible and whether it can be improved.

Assessment of publication bias in the presence of between-study heterogeneity

8.1 Chapter overview

In Chapter 7 the performance of tests to detect publication bias in the presence of unexplained between-study heterogeneity was investigated and shown to be problematic. In some situations between-study heterogeneity can be partly explained by measured study-level covariates, such as country of study, the type of population and, in meta-analyses of animal experiments, the different species used. When such predictable between-study heterogeneity exists, assessing publication bias is not straightforward. The aim of this chapter is to explore approaches for the assessment of publication bias in the presence of explainable between-study heterogeneity in an attempt to disentangle one from the other. In Section 8.2, a published meta-analysis of animal experiments where some of the between-study heterogeneity can be explained is introduced as the motivating example for this chapter: the meta-analysis by Mapstone *et al.* (2003). Details of simulation analyses used to assess the performance of methods for the detection and adjustment of publication bias when between-study heterogeneity exists and can be partially explained, are given in Section 8.3. In Section 8.4 their results are presented and applied to the Mapstone *et al.* meta-analysis. The practical use of these techniques is discussed in Section 8.5 in addition to some recommendations.

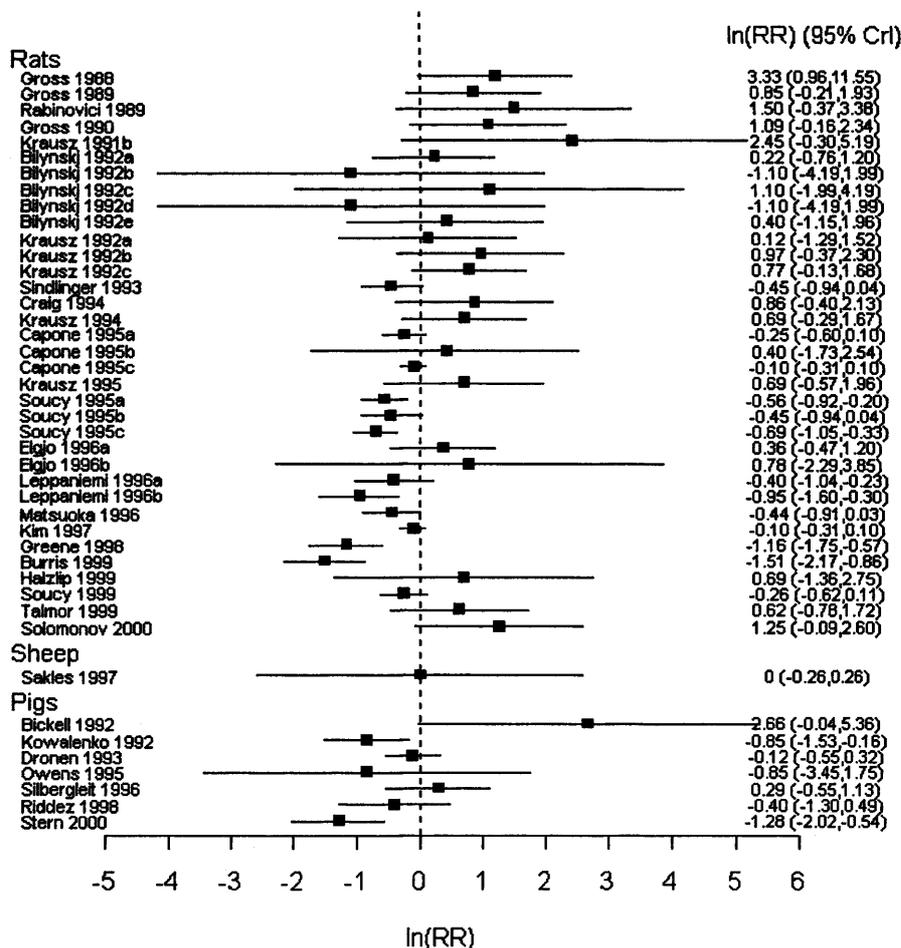
8.2 Motivating example

8.2.1 *The Mapstone et al. (2003) meta-analysis*

Between-study heterogeneity and publication bias are important aspects of meta-analyses (Sutton *et al.*, 2000; Egger *et al.*, 2001) and can be simultaneous features of a meta-analysis. Neglect of either one when carrying out a meta-analysis can have serious implications for the conclusions and subsequent decisions based on the meta-analysis. Although methods for the detection of between-study heterogeneity and publication bias exist, assessment of one in the presence of the other can be misleading. This is demonstrated by a re-analysis of a meta-analysis of animal experiments identified from the systematic review described in Chapter 4.

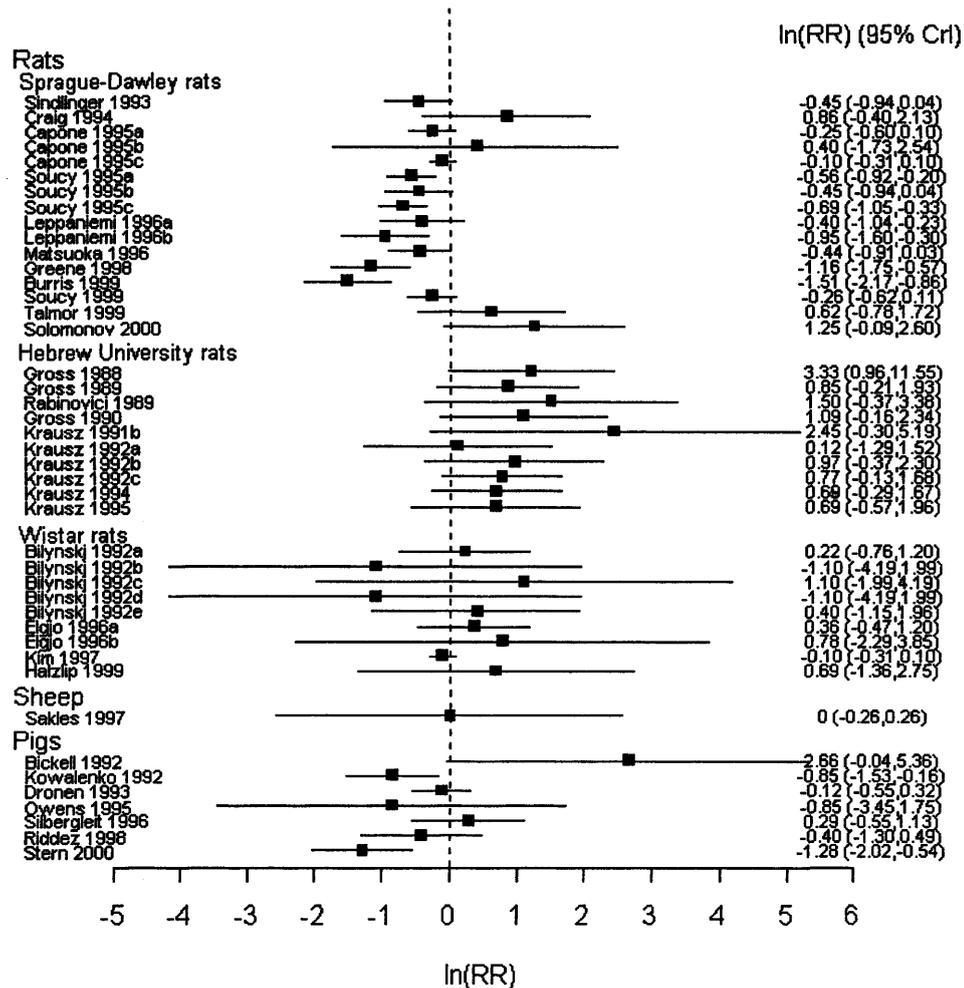
Mapstone *et al.* (2003) synthesise relative risks (RR) from experiments investigating fluid resuscitation and the risk of death in animals with haemorrhage. Using the quality assessment described in Section 4.5, the reporting of this meta-analysis was of reasonably high quality (reporting 81% of the specified items, compared to the mean of 53% of guideline items reported). Although possible sources of between-study heterogeneity are explored, Mapstone *et al.* (2003) do not assess publication bias.

RRs for mortality are obtained from 43 animal experiments on three species, rats ($n=35$), pigs ($n=7$) and sheep ($n=1$), and are combined using a weighted regression with random effects. Using the available data, a pooled RR of 0.88 (95% CI: 0.73, 1.07) was replicated in Stata with a random effects meta-analysis. Evidence suggests a great deal of between-study heterogeneity: I^2 of 67% (95% CI: 55%, 76%). Mapstone *et al.* report that type of haemorrhage model explained some of the heterogeneity between studies, in addition to follow-up time and volume of fluid infused. Consequently, they report RRs stratified by haemorrhage model, and adjusted for fluid used, follow-up time and species of animal, even though little difference is reported in the RRs with and without adjustment for species (and since 81% of the experiments are on the same species, a species effect would be difficult to identify). Figure 8.1 shows the $\ln(\text{RR})$ from all experiments by the species used.

Figure 8.1 Forest plot of experiments from Mapstone *et al.* (2003) by species used

The one sheep experiment (Sakles *et al.*, 1997) suggests no increase or decrease in the risk of mortality with fluid resuscitation compared to no fluid resuscitation. The data from the experiments on pigs generally suggests some evidence of a decrease in the risk of mortality, except for the experiment by Bickell *et al.* (1992). Although evidence from the rat experiments is mixed and quite heterogeneous, a number of different rat strains are used, including Sprague-Dawley, Hebrew university and Wistar rats. To investigate whether strain of rat explains any of the observed heterogeneity a forest plot indicating species and strain of rat is presented in Figure 8.2.

Figure 8.2 Forest plot of experiments from Mapstone et al. (2003) by species and strain used



Visual inspection of this plot suggests some evidence of a strain effect in the rat experiments, with experiments using Sprague-Dawley rats generally concluding a reduction in the risk of mortality with fluid resuscitation, while evidence from experiments using Hebrew university rats suggests an increased risk of mortality. The evidence from the Wistar rat strains is mixed and the $\ln(\text{RR})$ s from these experiments are generally less precise than those from experiments using other rat strains. The fixed and random effects models described in Equations 3.1 and 3.2, respectively, are used to combine the $\ln(\text{RR})$ estimates within each species and strain group indicated in Figure 8.2. These pooled estimates are presented in Table 8.1.

Table 8.1 Pooled RR estimates for mortality by species and strain

Species/strain	No. experiments	Pooled RR estimate		I ² (%) (95% CI)
		Fixed effects	Random effects	
Sprague-Dawley rats	16	0.68 (0.61, 1.31)	0.65 (0.52, 1.23)	68 (45, 81)
Hebrew university rats	10	2.34 (1.62, 3.52)		0 (0, 62)
Wistar rats	9	0.95 (0.79, 1.15)		0 (0, 65)
Sheep	1	1 (0.07, 13.3)		
Pigs	7	0.68 (0.51, 0.91)	0.69 (0.39, 1.20)	67 (27, 85)

These pooled RR estimates reflect the differences between rat strains observed in Figure 8.2. It is, however, plausible that the differences between the observed RRs from the rat strains are not necessarily reflecting true differences in the effect of fluid resuscitation between rat strains; rat strain may be a proxy for some other factor. For instance, the fact that all ten experiments using Hebrew university rats are carried out by the same research group should not be overlooked. Furthermore, six of the 16 experiments on Sprague-Dawley rats are carried out by the same group and five of the nine experiments using Wistar rats involve the same authors. It is therefore possible that the rat strain effect observed is a proxy for experimenter effect, environment effect or some other effect related to the laboratory used or the group of researchers.

A number of possible sources of between-study heterogeneity were explored by Mapstone *et al.*, (2003) such as haemorrhage model, fluid used and follow-up time, however for the purpose of demonstrating how ignoring between-study heterogeneity (explained by study-level covariates) can give misleading results when assessing publication bias, only species and strain differences are referred to in this chapter. Although it is very likely that more than one covariate may explain between-study heterogeneity, this is not addressed here but is discussed in Section 8.5. The aim of this chapter is to explore approaches for assessing publication bias when some, or all, of the between-study heterogeneity can be explained by a study-level covariate, e.g. species or strain of animal. Throughout this chapter, such

between-study heterogeneity is referred to as predictable or explainable.

Unexplained, or unpredictable, between-study heterogeneity refers to extra variation that cannot be explained by a study level covariate measured in the individual studies (i.e. has been generated by a random effects model in the simulations).

8.2.2 Assessment of publication bias

Even when possible sources of explainable between-study heterogeneity are identified in a meta-analysis, it is often not taken into account when assessing publication bias (e.g. Boffetta and Silverman, 2001; Fleischaur and Arab, 2001). Such an approach could be termed *naïve*, as a possible source of heterogeneity is ignored in this assessment which could distort the results. For example, funnel plot asymmetry and/or significant findings from a publication bias test may not be due to publication bias at all, but instead due to between-study heterogeneity distorting the funnel plot and/or publication bias test. A funnel plot of all 43 experiments in the Mapstone *et al.* (2003) meta-analysis based on the inverse of the standard error suggests evidence of publication bias (Figure 8.3), whereas a funnel plot based on the inverse of sample size appears less skewed (Figure 8.4). In both plots the dashed line represents the pooled RR of 0.88.

Results of Chapter 7 (Section 7.6.2) suggest that inflated type I error rates seen with Egger's regression test when ORs are used are also seen when RRs are used. The modified Macaskill test (based on the inverse of the sample size – Model 4c from Chapter 7; Peters *et al.*, 2006 (Appendix H)) however, was found to have appropriate type I error rates when RRs are used and reasonable power to detect publication bias. For completeness, results from the modified Macaskill test are presented alongside those from the commonly used regression and rank correlation tests to accompany the above funnel plots. The rank correlation test and Egger's regression test (both based on the standard error) reflect the asymmetry observed in Figure 8.3, with $p = 0.069$ and $p = 0.018$, respectively. The modified Macaskill test however, suggests little evidence of publication bias ($p = 0.647$) as Figure 8.4 suggests.

Figure 8.3 Funnel plot (where y-axis is the inverse of the standard error) of experiments in Mapstone et al. (2003) meta-analysis

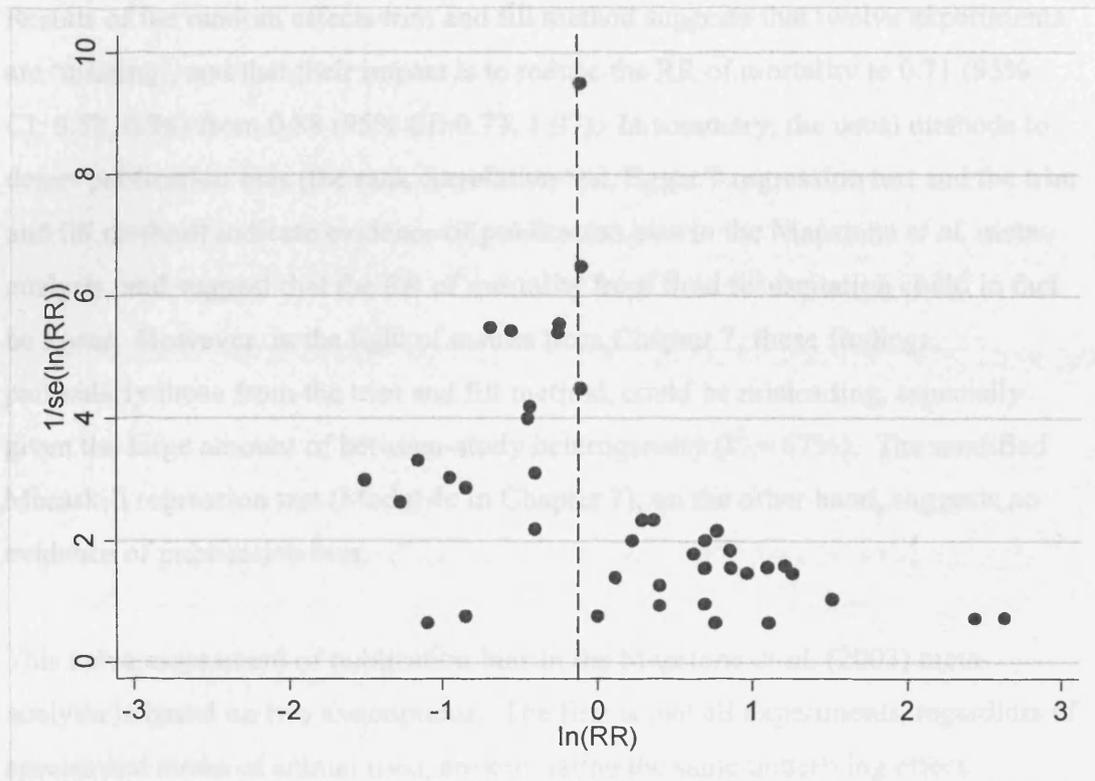
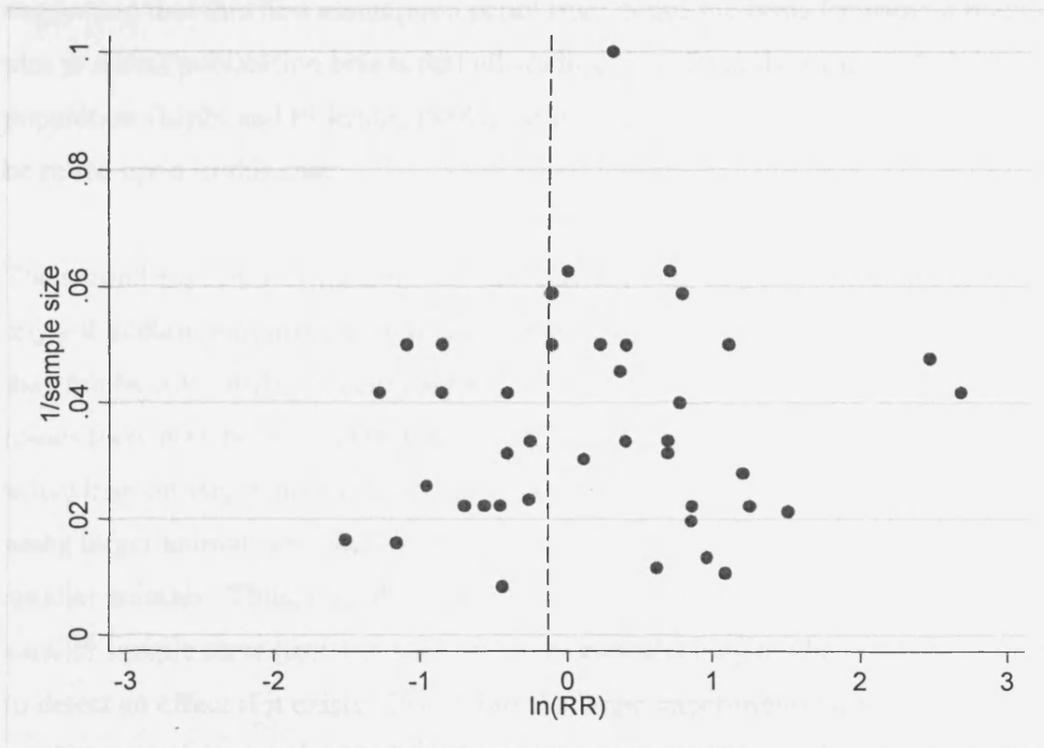


Figure 8.4 Funnel plot (where y-axis is the inverse of the sample size) of experiments in Mapstone et al. (2003) meta-analysis



As part of sensitivity analyses, one may wish to apply the trim and fill method to these data to investigate the impact of the asymmetry observed in Figure 8.3. Results of the random effects trim and fill method suggests that twelve experiments are 'missing', and that their impact is to reduce the RR of mortality to 0.71 (95% CI: 0.58, 0.86) from 0.88 (95% CI: 0.73, 1.07). In summary, the usual methods to detect publication bias (the rank correlation test, Egger's regression test and the trim and fill method) indicate evidence of publication bias in the Mapstone *et al.* meta-analysis, and suggest that the RR of mortality from fluid resuscitation could in fact be lower. However, in the light of results from Chapter 7, these findings, particularly those from the trim and fill method, could be misleading, especially given the large amount of between-study heterogeneity ($I^2 = 67\%$). The modified Macaskill regression test (Model 4c in Chapter 7), on the other hand, suggests no evidence of publication bias.

This naïve assessment of publication bias in the Mapstone *et al.* (2003) meta-analysis is based on two assumptions. The first is that all experiments, regardless of species and strain of animal used, are estimating the same underlying effect. However, as is seen from Table 8.1, the evidence indicates some difference in the effect of fluid resuscitation depending on the species or strain of animal used, suggesting that this first assumption is not true. Since the basis for using a funnel plot to detect publication bias is that all studies come from the same underlying population (Light and Pillemar, 1984), the funnel plot and tests based on it cannot be relied upon in this case.

The second assumption made by this naïve assessment of publication bias is that any publication bias affects all the experiments in the same way. This assumption may not be true either with animal experiments. For example, because of ethical issues there may be more experiments carried out on smaller animals (e.g. rats and mice) than on larger animals (e.g. dogs and monkeys). Furthermore, experiments using larger animals are likely to use fewer of them than the experiments using smaller animals. Thus, regardless of the underlying effect, experiments with smaller sample sizes (usually those on larger animals) may not be powerful enough to detect an effect if it exists. In contrast the larger experiments (usually those on smaller animals) are more likely to conclude a statistically significant effect if it

exists because of increased power. If different effects are observed in different species and species is related to the size of the experiment a funnel plot is likely to be skewed by this. When it comes to publication, experiments using larger animals are likely to be published regardless of their findings, because of the ethical issues of experimentation on larger animals. Thus, publication bias may well affect experiments on different animals in different ways, leading one to question the validity of the second assumption of this naïve assessment of publication bias. This issue of differential publication bias in studies of different designs has been discussed by Sutton *et al.* (2002) in relation to differing human study designs.

Since it is likely that both of the assumptions made in the naïve assessment of publication bias are untrue, a number of possible approaches could be taken, including the assessment of publication bias within the homogeneous groups and the simultaneous modelling of publication bias and the explainable between-study heterogeneity (more details on these models are given in Section 8.3.2). In Figure 8.5 (y-axis is the inverse of standard error) and Figure 8.6 (y-axis is the sample size) funnel plots are presented for assessment of publication bias within the different species and rat strain groups in the Mapstone *et al.* (2003) meta-analysis.

P-values from the rank correlation, Egger's regression and the modified Macaskill tests for publication bias within each species and strain subgroup (except for sheep where only one experiment exists) are presented in Table 8.2. The findings from these tests suggest little evidence for the presence of publication bias *within* the subgroups.

Table 8.2 *P-values for publication bias within species and strain of animal*

Species/strain	No. experiments	p-value			I ²
		Rank correlation test	Egger's regression test	Modified Macaskill test*	
Sprague-Dawley rats	16	0.79	0.87	0.83	68%
Hebrew university rats	10	0.83	0.20	0.93	0%
Wistar rats	9	0.21	0.07	0.36	0%
Pigs	7	0.55	0.73	0.87	67%

*Model 4c in Chapter 7

Figure 8.5 Funnel plots, with inverse standard error on the y-axis, of experiments in Mapstone et al. (2003) meta-analysis specified by species and rat strain*

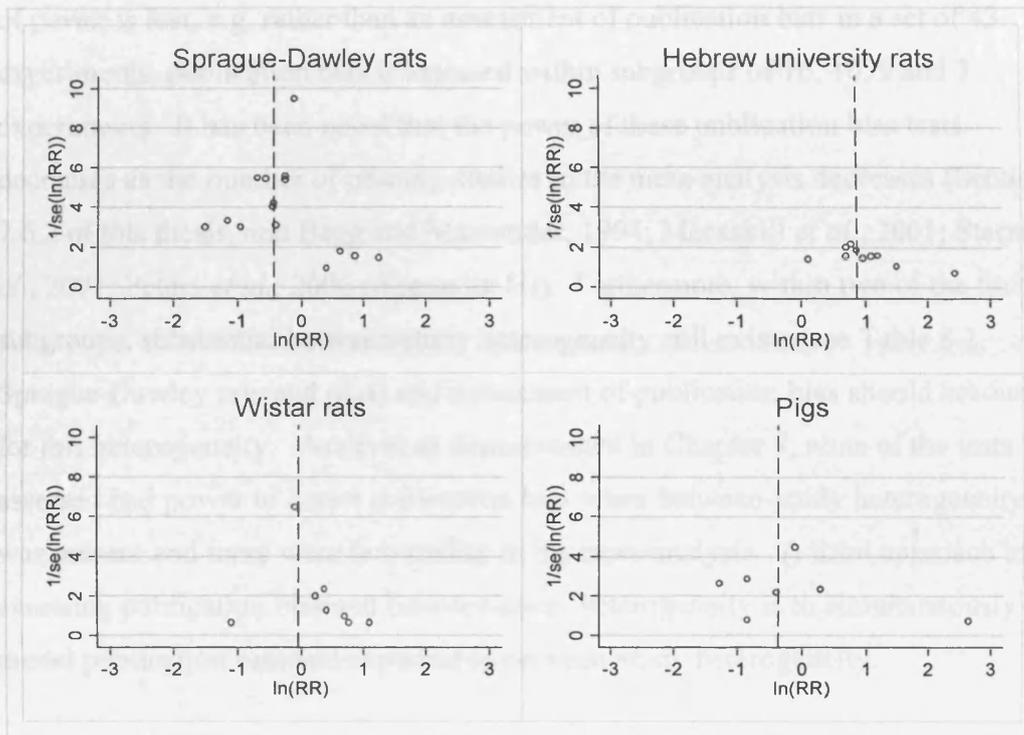
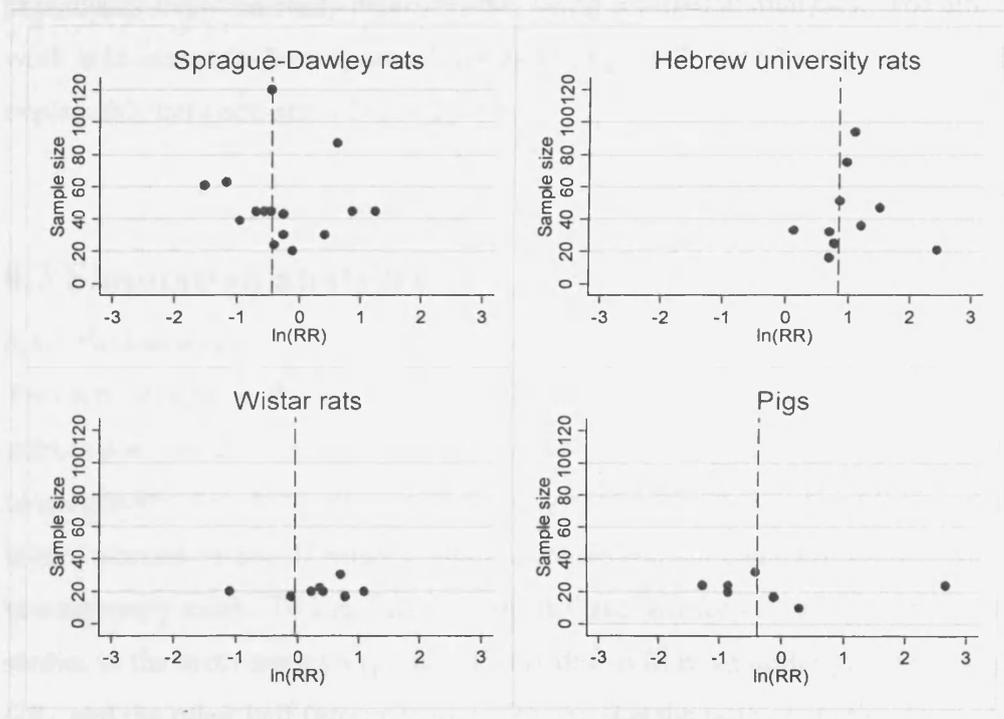


Figure 8.6 Funnel plots, with sample size on the y-axis, of experiments in Mapstone et al. (2003) meta-analysis specified by species and rat strain*



* Dashed line is the pooled random effects estimate of $\ln(RR)$ for each subgroup as in Table 8.1

This approach to assessing publication bias in the presence of between-study heterogeneity takes account of the different effects in the subgroups, but a great deal of power is lost, e.g. rather than an assessment of publication bias in a set of 43 experiments, publication bias is assessed within subgroups of 16, 10, 9 and 7 experiments. It has been noted that the power of these publication bias tests decreases as the number of primary studies in the meta-analysis decreases (Section 7.6.2 of this thesis, and Begg and Mazumdar, 1994; Macaskill *et al.*, 2001; Sterne *et al.*, 2001; Peters *et al.*, 2006 (Appendix H)). Furthermore, within two of the four subgroups, substantial between-study heterogeneity still exists (see Table 8.2, Sprague-Dawley rats and pigs) and assessment of publication bias should account for this heterogeneity. However as demonstrated in Chapter 7, none of the tests assessed had power to detect publication bias when between-study heterogeneity was present and there were few studies in the meta-analysis. A third approach to assessing publication bias and between-study heterogeneity is to simultaneously model publication bias and explainable between-study heterogeneity.

In the following sections naïve assessments of publication bias in the presence of explainable between-study heterogeneity are examined and compared to the performance of regression models which simultaneously model publication bias and explainable between-study heterogeneity using simulation analyses. The aim of this work is to assess the best approach for assessing publication bias in the presence of explainable between-study heterogeneity.

8.3 Simulation analyses

8.3.1 Parameters

Two sets of meta-analysis scenarios were simulated to allow an assessment of publication bias detection methods in the presence of explainable between-study heterogeneity: 1) where only between-study heterogeneity explainable by a study level covariate exists, 2) where both explainable and unexplained between-study heterogeneity exist. To simulate the first of these two scenarios half of the primary studies in the meta-analysis (group 1) were drawn from an underlying OR given by OR_1 , and the other half (group 2) from OR_2 , so that the ratio of studies drawn from

OR_1 to studies drawn from OR_2 is one. Table 8.3 details the three combinations of underlying OR that were used.

Table 8.3 *Combinations of $\ln(OR)$ used to simulate predictable between-study heterogeneity*

Combination	OR_1	OR_2
1	1	1.5
2	1	5
3	1.5	5

Within each group the fixed effects model given in Equation 7.8 was used to simulate the meta-analyses. This scenario may not, however, be particularly realistic, as even after adjustment for known covariates between-study heterogeneity may still exist (Sutton *et al.*, 2000). This is a reason why random effects meta-regression is often advocated over fixed effects meta-regression (Sutton *et al.*, 2000; Thompson and Higgins, 2002; Whitehead, 2002).

Meta-analyses in the second scenario (where both explained and unexplained heterogeneity is present) were simulated as above using combinations 1, 2 and 3 in Table 8.3, but the random effects model given in Equation 7.8 was used within each group to model the unexplained between-study heterogeneity. Three levels of unexplained between-study heterogeneity were applied as with the simulated meta-analyses in Chapter 7: 20%, 150% and 500% of the average within-study variance. In each meta-analysis, the same level of unexplainable between-study heterogeneity is simulated in groups 1 and 2.

Within each meta-analysis, the total number of primary studies is defined as in Section 7.51 to be 6, 16, 30 or 90, thus within each group (1 or 2) in a meta-analysis, there are 3, 8, 15 or 45 studies, respectively. The probability of an adverse event in the control group sampled from a uniform distribution (0.3, 0.7), the natural logarithm of the number of control subjects within each primary study taken from the distribution $N(5, 0.3)$ and the ratio of exposed to control subjects is one.

The meta-analyses were simulated in the presence and absence of publication bias. Publication bias was induced in the same way as described in Section 7.5.1, by p-value or effect size. It is assumed that publication bias occurs within each of the two groups of studies within a meta-analysis, so that when publication bias is induced by effect size, in each group studies in the left hand-side of each funnel are missing, as shown in Figure 8.7. When publication bias is induced by p-value and the underlying OR is large, the actual bias induced by this mechanism is likely to be small. Therefore the studies in group 2 with underlying OR_2 in a meta-analysis (see Table 8.3) are less likely to be subject to publication bias when it is induced by p-value. This is illustrated in Figure 8.8, where studies with an underlying OR closer to the null (solid points) are more likely to be subject to publication bias than the studies with an underlying OR further from the null (hollow points). This issue was described and discussed in Section 7.5.1 and should be borne in mind when interpreting power based on publication bias that has been induced by p-value. Only findings relating to 'severe' publication bias are presented here. Findings from 'moderate' publication bias follow the same general trend in power to detect 'severe' publication bias, but levels of power for 'moderate' publication bias are at lower levels.

Figure 8.7 Inducing publication bias (based on effect size) within each group in a meta-analysis

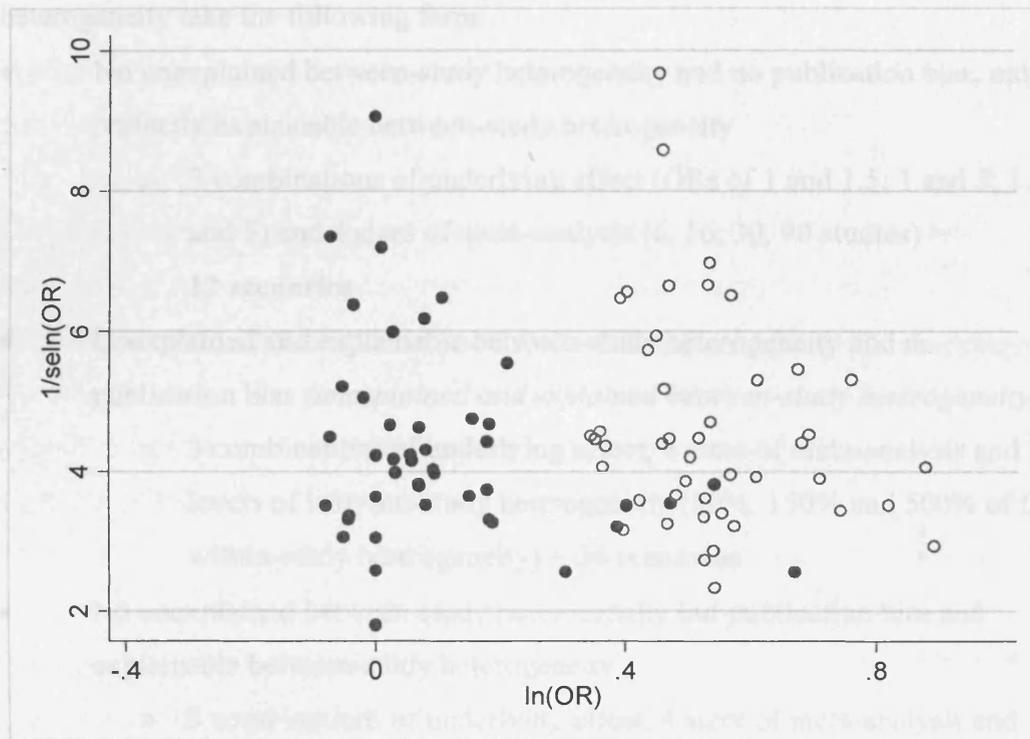
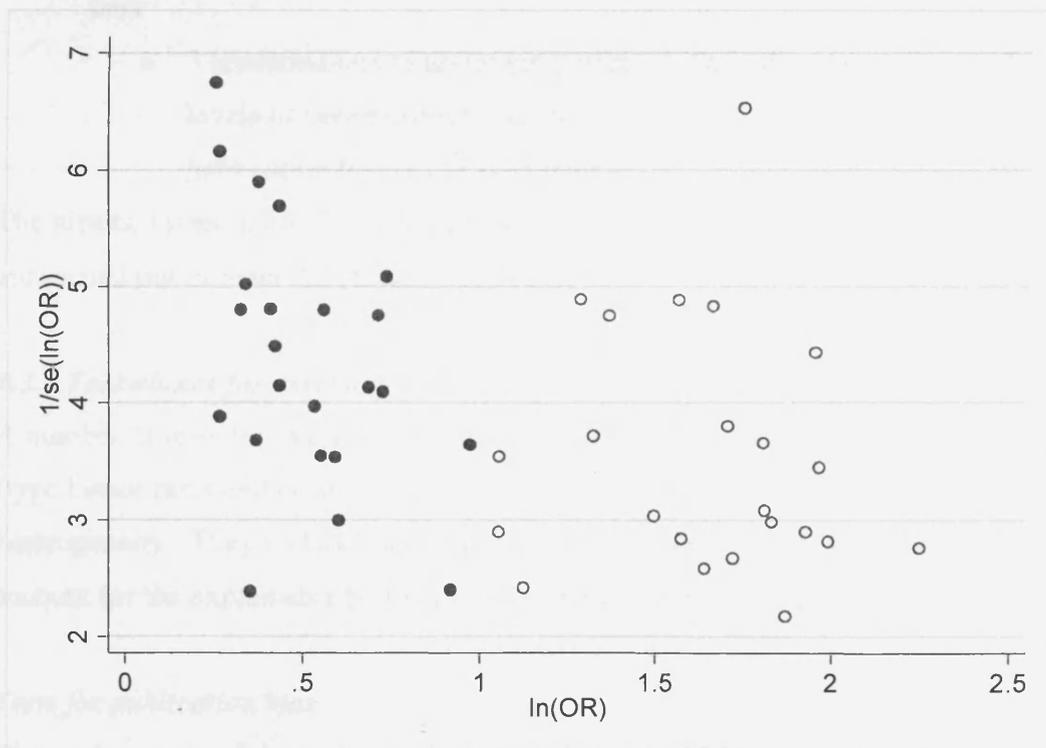


Figure 8.8 Inducing publication bias (based on p-value) within each group in a meta-analysis



In summary, the simulated meta-analyses used in the assessment of methods to detect and adjust for publication bias in the presence of explainable between-study heterogeneity take the following form:

- No unexplained between-study heterogeneity and no publication bias, only perfectly explainable between-study heterogeneity
 - » 3 combinations of underlying effect (ORs of 1 and 1.5; 1 and 5; 1.5 and 5) and 4 sizes of meta-analysis (6, 16, 30, 90 studies) = **12 scenarios**
- Unexplained and explainable between-study heterogeneity and no publication bias (*unexplained and explained between-study heterogeneity*)
 - » 3 combinations of underlying effect, 4 sizes of meta-analysis and 3 levels of between-study heterogeneity (20%, 150% and 500% of the within-study heterogeneity) = **36 scenarios**
- No unexplained between-study heterogeneity but publication bias and explainable between-study heterogeneity
 - » 3 combinations of underlying effect, 4 sizes of meta-analysis and four types of induced publication bias (by p-value: moderate and severe; by effect size: 14% and 40% censored) = **48 scenarios**
- Unexplained and explainable between-study heterogeneity and publication bias
 - » 3 combinations of underlying effect, 4 sizes of meta-analysis and 3 levels of between-study heterogeneity and four types of induced publication bias = **144 scenarios**

The results, given in Section 8.4, are based on 1000 repetitions of each scenario and are carried out in Stata 8.2 (StataCorp, 2004).

8.3.2 Techniques for assessing publication bias

A number of tests for publication bias are assessed in terms of their performance (type I error rates and power) in the presence of explainable between-study heterogeneity. They include regression models, with and without a covariate to account for the explainable between-study heterogeneity. These are now described.

Tests for publication bias

The performance of the rank correlation test (Begg and Mazumdar, 1994) and the

eight regression tests described in Section 7.5.2 (Models 1 – 6) are assessed where between-study heterogeneity, explainable by a study level covariate, is a feature of the simulated meta-analyses. This type of analysis reflects the naïve assessment of publication bias in the Mapstone *et al.* meta-analysis described in Section 8.2.2 and will be referred to as the *naïve* assessment throughout this chapter.

The performance of extended versions of the eight regression models is also assessed in these simulation analyses. These extended regression models (Models 1E – 6E) include a term to account for the study-level covariate explaining the between-study heterogeneity. The extension of Egger's regression test to account for between-study heterogeneity by including an extra term, has been suggested (Sterne *et al.*, 2000), but not carried out before.

The idea behind this approach is that if between-study heterogeneity is being mistaken for publication bias, as was shown in Mapstone *et al.* (2003) in Section 8.2, one can adjust for the explainable between-study heterogeneity (using regression), and then assess the meta-analysis using one of these tests for the presence of publication bias, all in one step.

These extended models are given below, where $group_i$ ($i = 1, 2$) specifies which underlying OR each study comes from in the meta-analysis. For example, in a meta-analysis simulated under combination 1 in Table 8.3, $group_1$ defines those studies with an underlying OR of 1 and $group_2$ defines those studies with an underlying OR of 1.5.

Egger's fixed effects regression on standard error

$$\frac{y_i}{se_i} = \beta + \frac{\alpha}{se_i} + \gamma \cdot group_i + \varepsilon_i \quad \text{(Model 1E)}$$

Egger's fixed effects regression on inverse of sample size

$$y_i \cdot size_i = \alpha + \beta \cdot size_i + \gamma \cdot group_i + \varepsilon_i \quad \text{(Model 2E)}$$

Linear fixed effects regression on standard error

$$y_i = \alpha + \beta.se_i + \gamma.group_i + \varepsilon_i, \text{ weighted by } \frac{1}{se_i^2} \quad \text{(Model 3E)}$$

Linear fixed effects regression on sample size

$$y_i = \alpha + \beta.size_i + \gamma.group_i + \varepsilon_i, \text{ weighted by } \frac{1}{se_i^2} \quad \text{(Model 4aE)}$$

Linear fixed effects regression on sample size – FPV model (Macaskill et al., 2001)

$$y_i = \alpha + \beta.size_i + \gamma.group_i + \varepsilon_i, \text{ weighted by } \left(\frac{1}{a_i + b_i} + \frac{1}{c_i + d_i} \right)^{-1} \quad \text{(Model 4bE)}$$

Linear fixed effects regression on inverse of sample size

$$y_i = \alpha + \frac{\beta}{size_i} + \gamma.group_i + \varepsilon_i, \text{ weighted by } \left(\frac{1}{a_i + b_i} + \frac{1}{c_i + d_i} \right)^{-1} \quad \text{(Model 4cE)}$$

Linear random effects regression on standard error

$$y_i = \alpha + \beta.se_i + \gamma.group_i + \mu_i + \varepsilon_i, \text{ weighted by } \frac{1}{se_i^2} \quad \text{(Model 5E)}$$

Linear random effects regression on sample size

$$y_i = \alpha + \beta.size_i + \gamma.group_i + \mu_i + \varepsilon_i, \text{ weighted by } \frac{1}{se_i^2} \quad \text{(Model 6E)}$$

8.4 Results

8.4.1 Between-study heterogeneity

When the covariate group is ignored, moment-based estimates of between-study heterogeneity (obtained from the META command in Stata (StataCorp, 2004)) in the simulated meta-analyses clearly demonstrate that there is more heterogeneity in meta-analyses from combinations 2 (OR=1, 5) and 3 (OR=1.5, 5), compared to combination 1 (OR=1, 1.5) (Table 8.4). This is as expected since the underlying ORs in combinations 2 and 3 are more diverse than those in combination 1.

Between-study heterogeneity is around 15-20 times greater in combination 2 than in combination 1, and 9-11 times greater in combination 3 than in combination 1.

Table 8.4 Summary estimates of between-study heterogeneity in the simulated meta-analyses

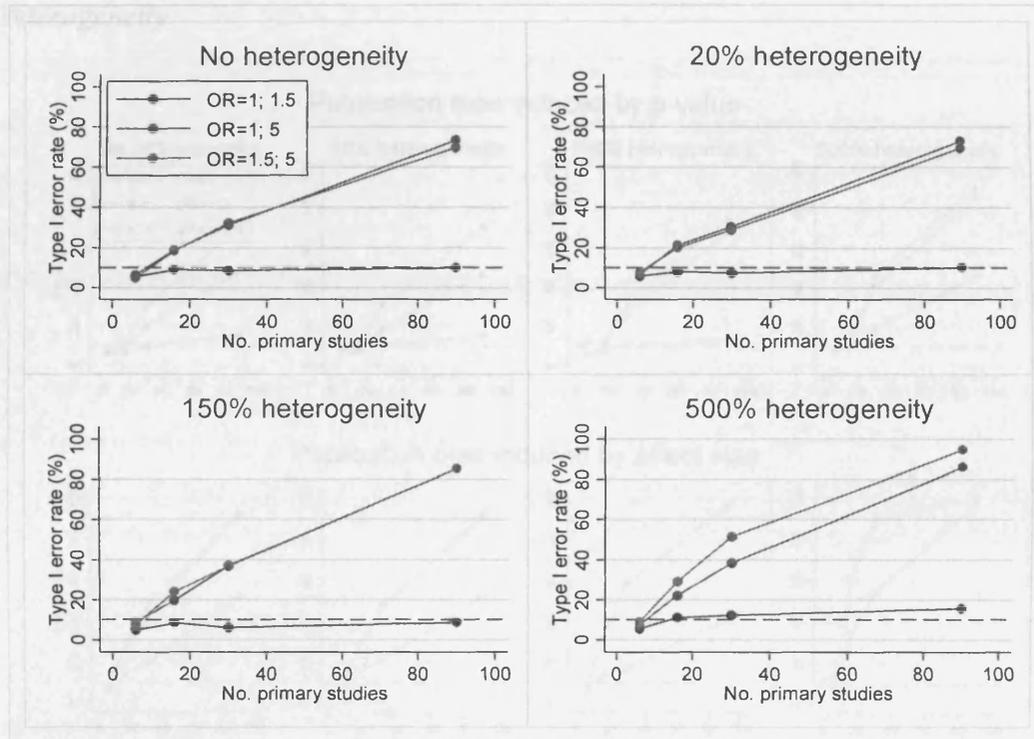
Median estimate of between-study heterogeneity		Ratio of between-study heterogeneity	Number of primary studies in meta-analysis
Combination 1	Combination 2	[Combination 2 ÷ Combination 1]	
0.0375	0.7584	20.22	6
0.0416	0.6735	16.19	16
0.0407	0.6482	15.93	30
0.0409	0.6356	15.54	90

Median estimate of between-study heterogeneity		Ratio of between-study heterogeneity	Number of primary studies in meta-analysis
Combination 1	Combination 3	[Combination 3 ÷ Combination 1]	
0.0375	0.4101	10.94	6
0.0416	0.375	9.01	16
0.0407	0.3668	9.01	30
0.0409	0.3551	8.68	90

8.4.2 Naïve assessment of publication bias - the rank correlation test

When the two subgroups come from relatively similar underlying distributions (i.e. combination 1, ORs = 1 and 1.5), the funnel plots from these two groups overlap (see Figures 8.7 and 8.8) and this is reflected in the 10% type I error rates from the rank correlation test (Figure 8.9). However, when the meta-analyses are simulated under combinations 2 and 3 where there is greater disparity between the underlying ORs from which the studies are drawn, the type I error rates are much larger (Figure 8.9).

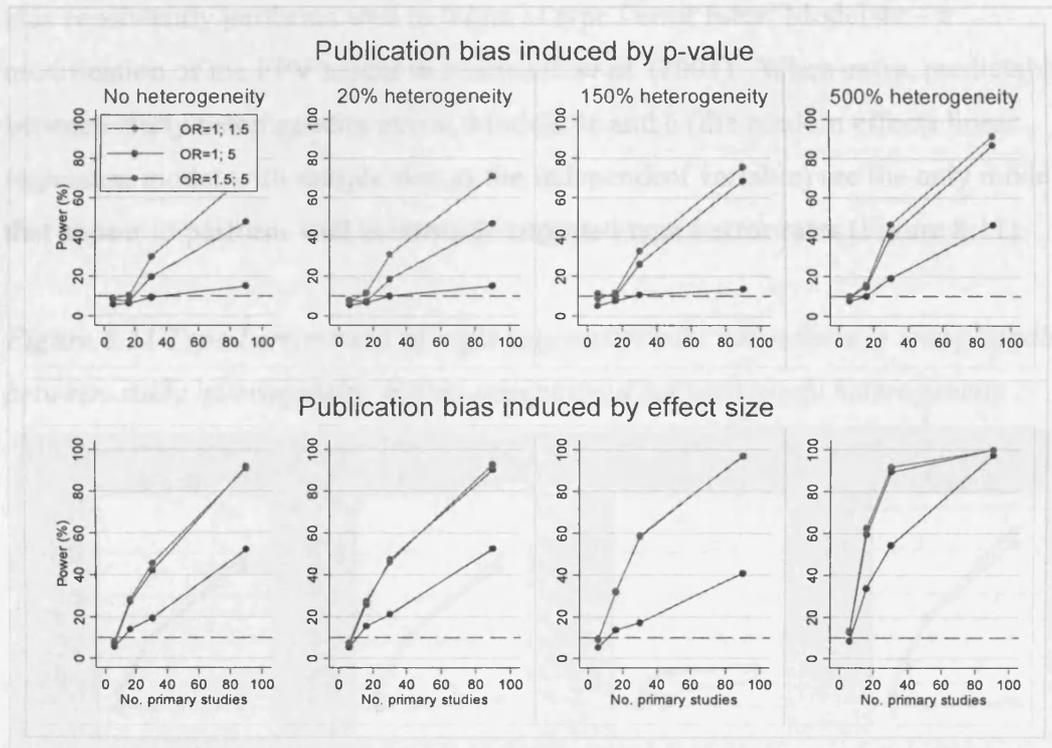
Figure 8.9 Type I error rates for the rank correlation test when there is explainable between-study heterogeneity and differing levels of unexplained between-study heterogeneity



The type I error rates are greater for meta-analyses simulated from the underlying OR combinations 2 and 3, than they are if all studies come from an underlying OR of 5 (Figure 7.7; Section 7.6.1). This is explained by the greater amount of heterogeneity in these meta-analyses caused by the large differences in the two underlying ORs. The type I error rates increase as the amount of unexplained between-study heterogeneity increases (Figure 8.9).

When 'severe' publication bias is induced by effect size the rank correlation test appears powerful to detect it (Figure 8.10), but it is difficult to distinguish this power from the high type I error rates (Figure 8.9). However, since the type I error rates for combination 1 are as expected, the rank correlation test appears to have moderate power to detect 'severe' publication bias when it is induced by effect size (Figure 8.10).

Figure 8.10 Power of the rank correlation test to detect 'severe' publication bias induced by p-value (top row) and effect size (bottom row) when there is explainable between-study heterogeneity and differing levels of unexplainable between-study heterogeneity



As seen in Figure 8.10, levels of power for the rank correlation test are higher when 'severe' publication bias is induced by effect size compared to when 'severe' publication bias is induced by p-value. Results for the rank correlation test and 'moderate' publication bias follow the same trend, but levels of power are much lower (see Figure J.1 of Appendix J).

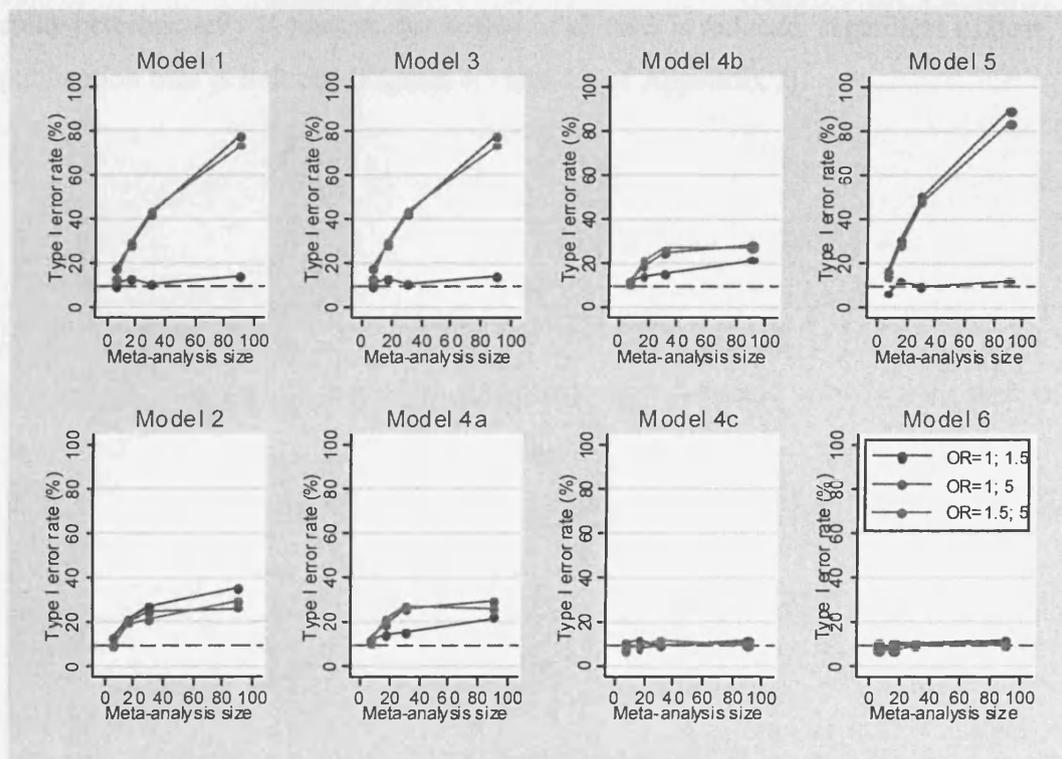
Results suggest that the rank correlation test does not perform well (in terms of type I error rates) when the primary studies in a meta-analysis are drawn from two quite different underlying ORs. These type I error rates increase as the amount of unexplained between-study heterogeneity increase.

8.4.3 Naïve assessment of publication bias - the regression tests

P-values obtained from the usual t-test

In Chapter 7, it was found that only one model used for the detection of publication bias consistently performs well in terms of type I error rates: Model 4c – a modification of the FPV model in Macaskill *et al.* (2001). When extra, predictable between-study heterogeneity exists, Models 4c and 6 (the random effects linear regression model with sample size as the independent variable) are the only models that appear to perform well in terms of expected type I error rates (Figure 8.11).

Figure 8.11 Type I error rates of eight regression tests when there is unexplainable between-study heterogeneity, but no unexplained between-study heterogeneity



When there is a great deal of unexplained between-study heterogeneity (500% of within study heterogeneity) Models 4c and 6 have the expected type I error rates regardless of the underlying ORs and the number of primary studies in the meta-analysis (Figure J.2 of Appendix J), all other models exceed the 10% type I error rate.

The power of some of these tests (Models 1, 3 and 5) looks to be high when publication bias is induced by p-value (Figure 8.12), but because of the corresponding high type I error rates, it is difficult to see how powerful these tests are for the detection of 'severe' publication bias. Model 4c, which had the appropriate type I error rates, has little power to detect publication bias, regardless of the characteristics of the simulated meta-analyses and whether publication bias is induced by p-value (Figure 8.12) or effect size (Figure 8.13).

Model 6 appears to have power to detect 'severe' publication bias when it is induced by effect size and the simulated meta-analyses are drawn under combination 1, $OR_1 = 1$ and $OR_2 = 1.5$ (see Figure 8.13). In all other scenarios power to detect publication bias does not exceed 20%. When unexplained between-study heterogeneity is present, the power of all tests is reduced, regardless of how publication bias is induced (Figures J.3 and J.4 of Appendix J).

Figure 8.12 Power of regression tests to detect 'severe' publication bias (induced by p -value) when there is explainable between-study heterogeneity but no unexplainable between-study heterogeneity

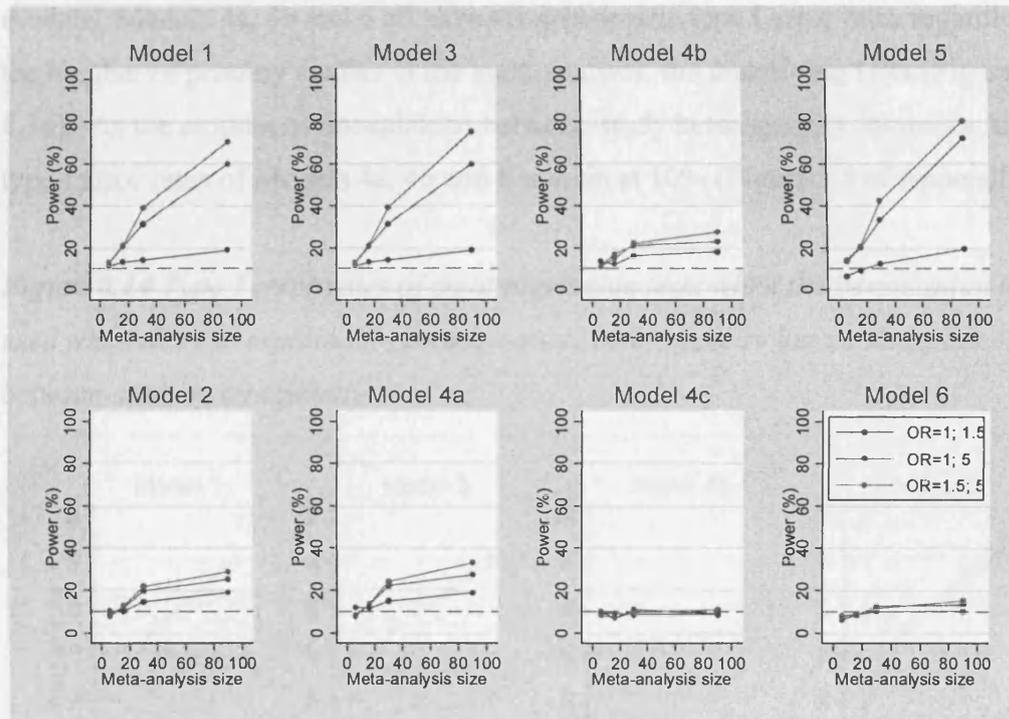
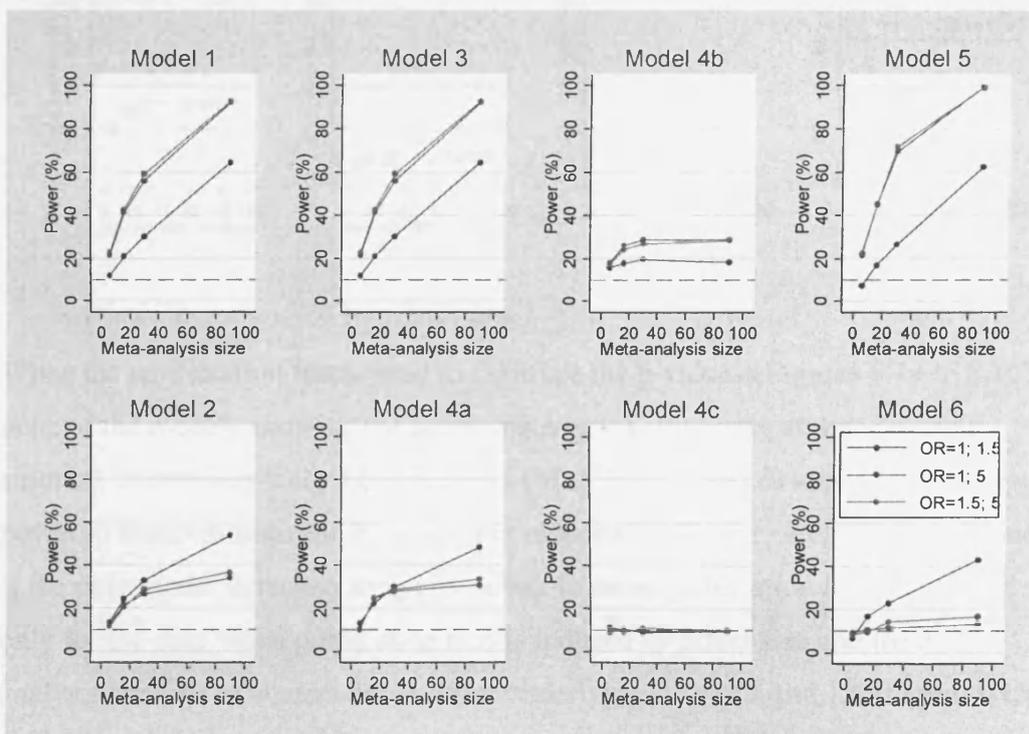


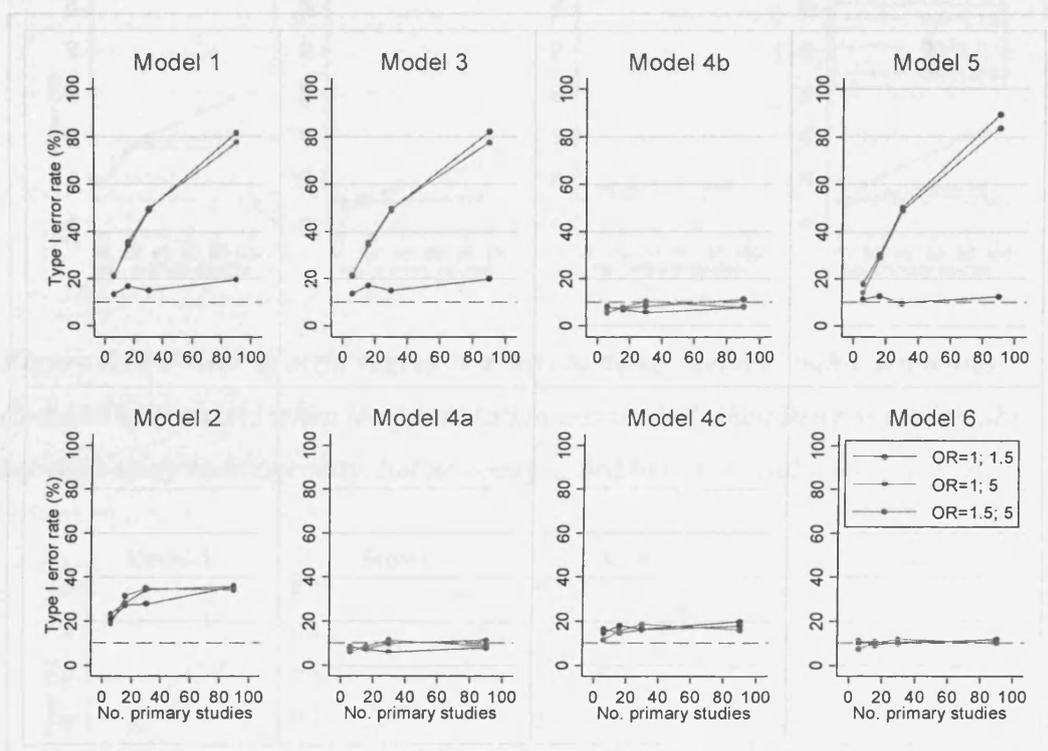
Figure 8.13 Power of regression tests to detect 'severe' publication bias (induced by effect size) when there is explainable between-study heterogeneity but no unexplained between-study heterogeneity



P-values obtained from the permutation test

The following results describe the performance of Models 1 – 6 when the permutation test is used to calculate the p-values of the coefficients from these models. Models 4a, 4b and 6 all have the appropriate type I error rates regardless of the number of primary studies in the meta-analysis, the underlying ORs (Figure 8.14). As the amount of unexplained between-study heterogeneity increases, the type I error rates of Models 4a, 4b and 6 remain at 10% (Figure J.5 of Appendix J).

Figure 8.14 Type I error rates of eight regression tests when the permutation test is used when there is explainable between-study heterogeneity but no unexplained between-study heterogeneity



When the permutation test is used to calculate the p-values (Figures 8.14 to 8.16), none of the models showing the appropriate type I error rates across differing amounts of between-study heterogeneity (Models 4a, b and 6) appear to have good power to detect publication bias whether induced by p-value or effect size. Model 6 is the only model demonstrating any power to detect publication bias and this is only for the case when publication bias is induced by effect size and the meta-analysis consists of studies drawn from underlying ORs of 1 and 1.5 (Figure 8.15).

Figure 8.15 Power of eight regression tests to detect 'severe' publication bias (induced by effect size) when the permutation test is used when there is explainable between-study heterogeneity but no unexplained between-study heterogeneity

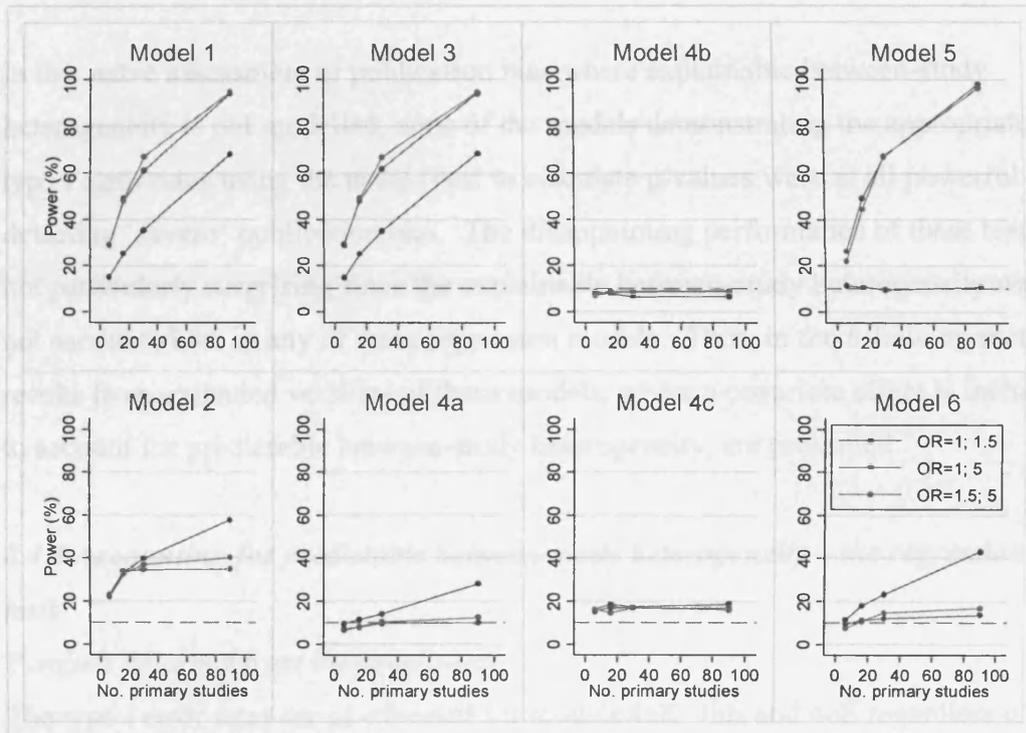
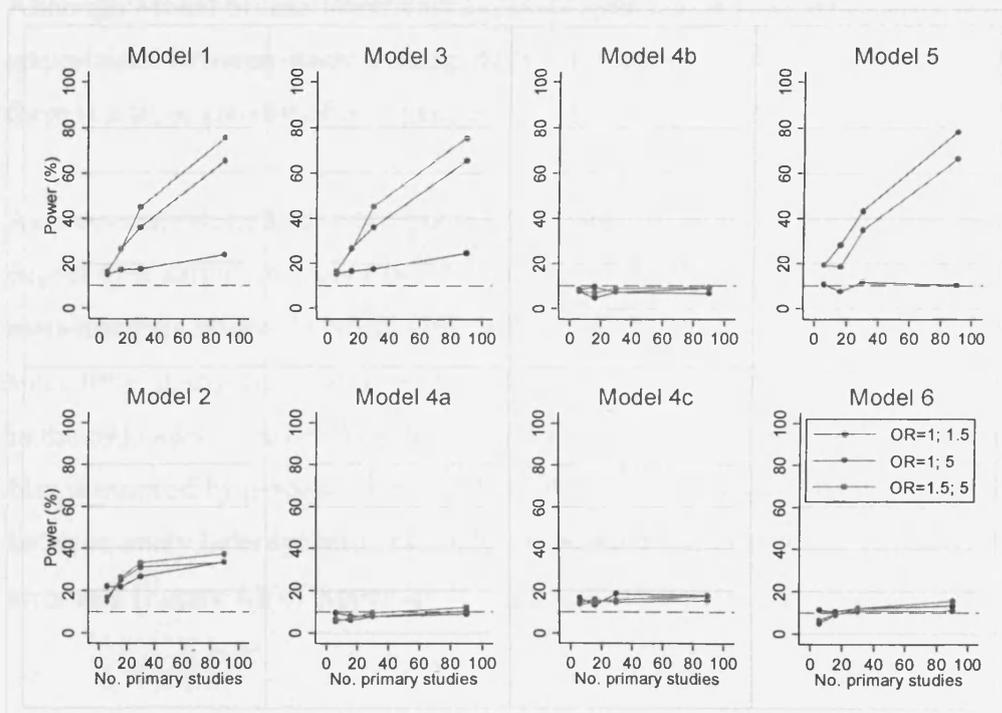


Figure 8.16 Power of eight regression tests to detect 'severe' publication bias (induced by p-value) when the permutation test is used when there is explainable between-study heterogeneity, but no unexplained between-study heterogeneity



Levels of power from Models 4a, 4b and 6 remain at 10% regardless of the amount of unexplained between-study heterogeneity that is present (Figures J.6 and J.7 of Appendix J).

In this naïve assessment of publication bias where explainable between-study heterogeneity is not modelled, none of the models demonstrating the appropriate type I error rates using the usual t-test to calculate p-values were at all powerful in detecting ‘severe’ publication bias. The disappointing performance of these tests is not particularly surprising since the explainable between-study heterogeneity was not accounted for in any of these regression models. Thus, in the following section, results from extended versions of these models, where a covariate effect is included to account for predictable between-study heterogeneity, are presented.

8.4.4 Accounting for predictable between-study heterogeneity – the regression tests

P-values obtained from the usual t-test

The type I error rates are as expected for models 4aE, 4bE and 4cE regardless of the number of primary studies in the meta-analysis and the underlying ORs (Figure 8.17). When a great deal of unexplained between-study heterogeneity exists, only Model 4cE continues to have the appropriate type I error rates (Figure 8.18).

Although Model 6E has lower than expected type I error rates when there is no unexplained between-study heterogeneity these error rates are as expected when there is a large amount of unexplained between-study heterogeneity.

As previously described, when publication bias is induced by p-value one would expect little difference between the estimates of power and type I error rates for meta-analyses where the underlying OR is far from the null (combinations 2 and 3), since little, if any, publication bias is actually induced. This is seen to some extent in the estimates of power from the extended regression models when publication bias is induced by p-value (Figure 8.19). When there is a great deal of unexplained between-study heterogeneity, estimates of power are similar to the estimated type I error rate (Figure J.8 of Appendix J).

Figure 8.17 Type I error rates for eight extended regression tests when there is explainable between-study heterogeneity but no unexplained between-study heterogeneity

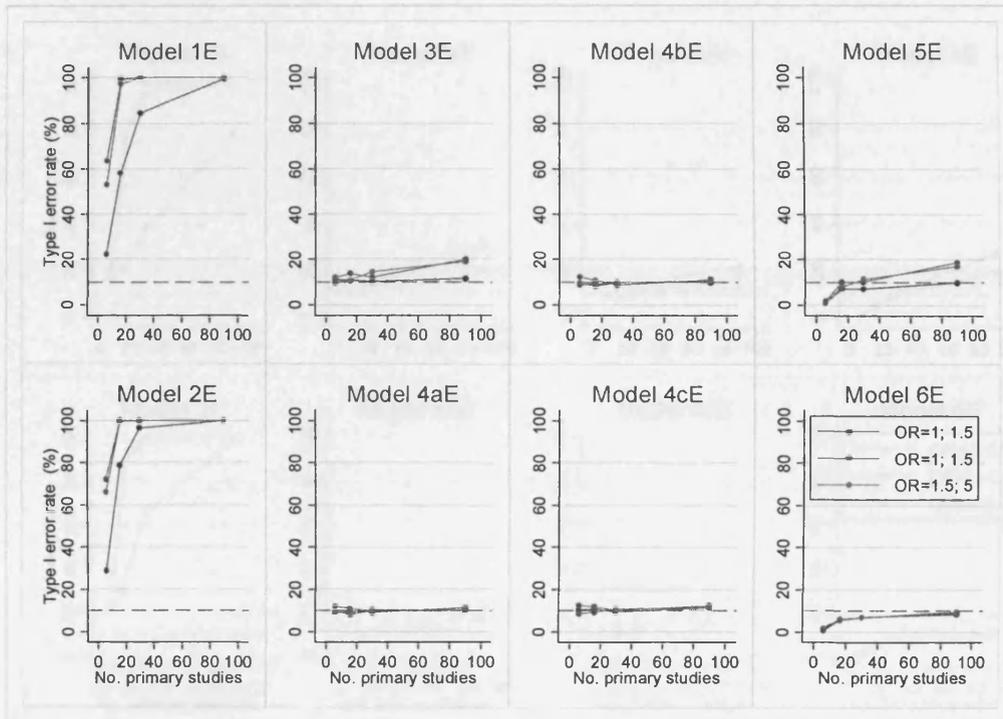


Figure 8.18 Type I error rates for eight extended regression tests when there is explainable between-study heterogeneity and a great deal of unexplained between-study heterogeneity (500%)

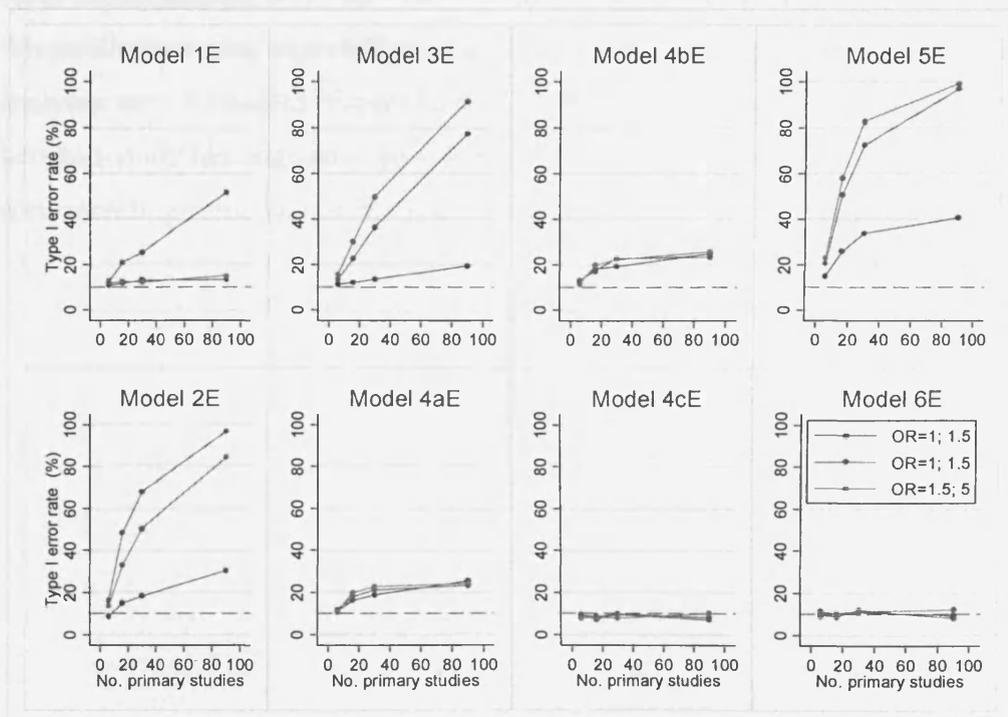
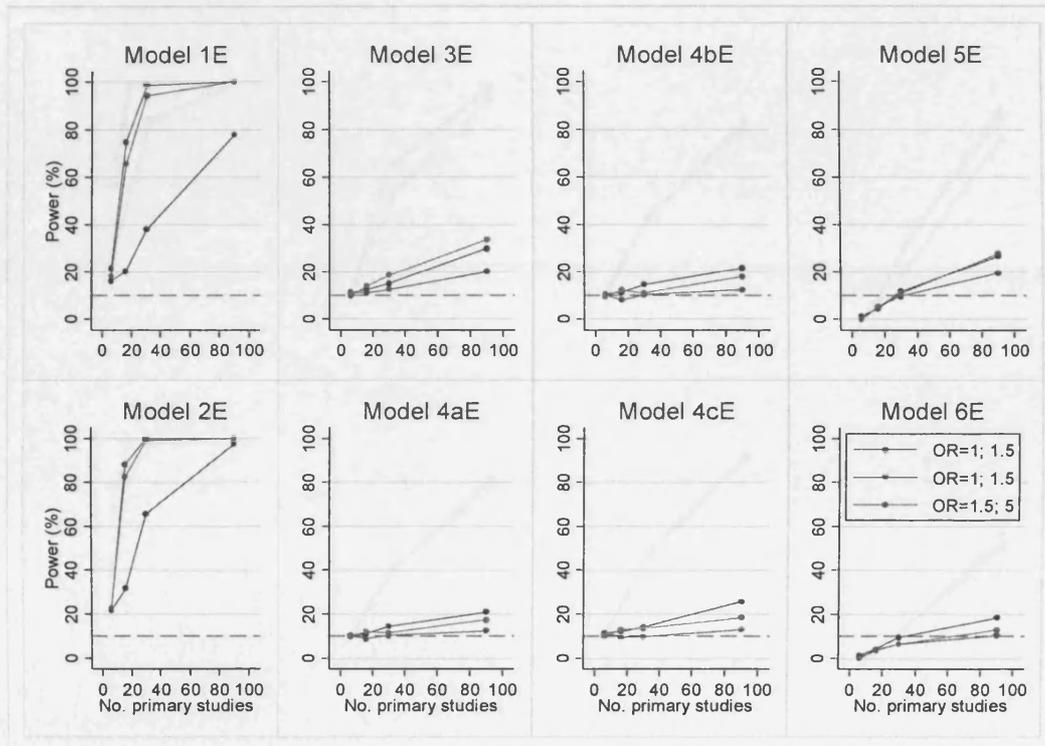
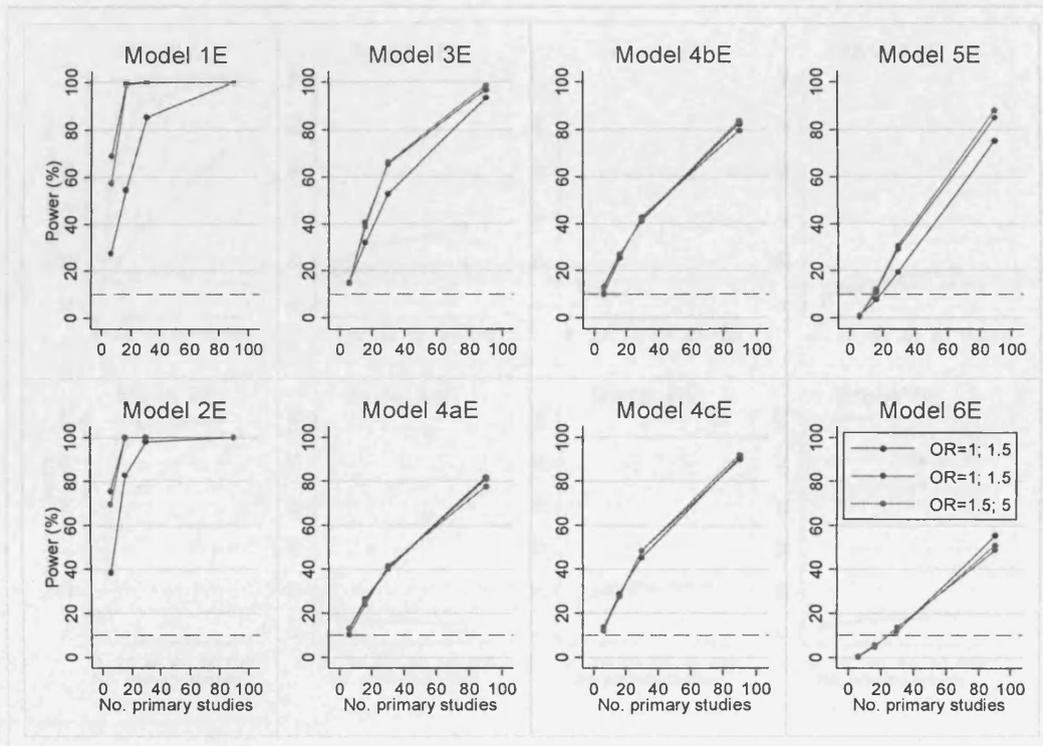


Figure 8.19 Power of eight extended regression tests to detect 'severe' publication bias (induced by p -value) when there is explainable between-study heterogeneity and no unexplained between-study heterogeneity



When publication bias is induced by effect size and there is no unexplained between study heterogeneity, all of the extended models appear to have good power to detect this publication bias, especially for large meta-analyses, regardless of how the meta-analyses were simulated (Figure 8.20). When there is a great deal of unexplained between-study heterogeneity, power is difficult to distinguish from the corresponding estimates of the type I error rates (Figure J.9 of Appendix J).

Figure 8.20 Power of eight extended regression tests to detect 'severe' publication bias (induced by effect size) when there is explainable between-study heterogeneity but no unexplained between-study heterogeneity



P-values obtained from the permutation test

When the permutation test is used to calculate the p-values given by Models 1E - 6E, none of the models appear to attain the appropriate type I error rates in all scenarios (Figures 8.21 and 8.22). Although Models 4aE, 4bE and 6E have the expected 10% type I error rate when there is a great deal of unexplained between-study heterogeneity (Figure 8.22).

When publication bias is induced by effect size, Models 4aE and 6E are the only models having reasonable type I error rates demonstrating good levels of power (Figure 8.23). This power is diminished when a great deal of unexplained between-study heterogeneity is present (Figure J.10 of Appendix J) and non-existent when publication bias is induced by p-value (Figure J.11 in Appendix J).

Figure 8.21 Type I error rates (when the permutation test is used to calculate the *p*-values) for eight extended regression tests when there is explainable between-study heterogeneity but no unexplained between-study heterogeneity

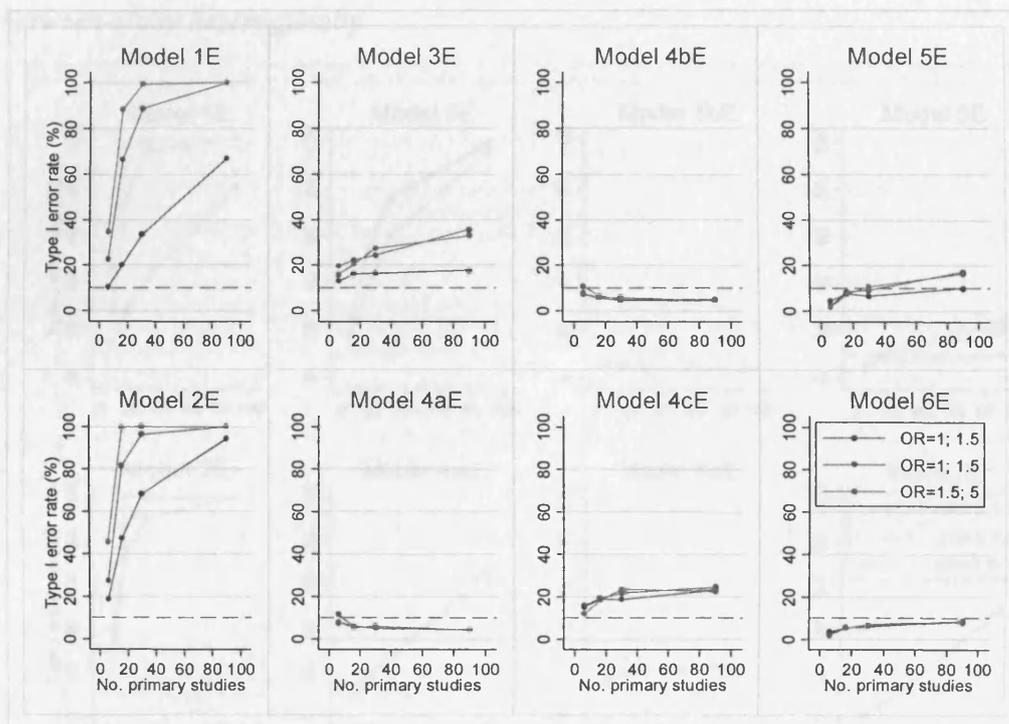


Figure 8.22 Type I error rates (when the permutation test is used to calculate the *p*-values) for eight regression tests when there is explainable between-study heterogeneity and a great deal of unexplained between-study heterogeneity (500%)

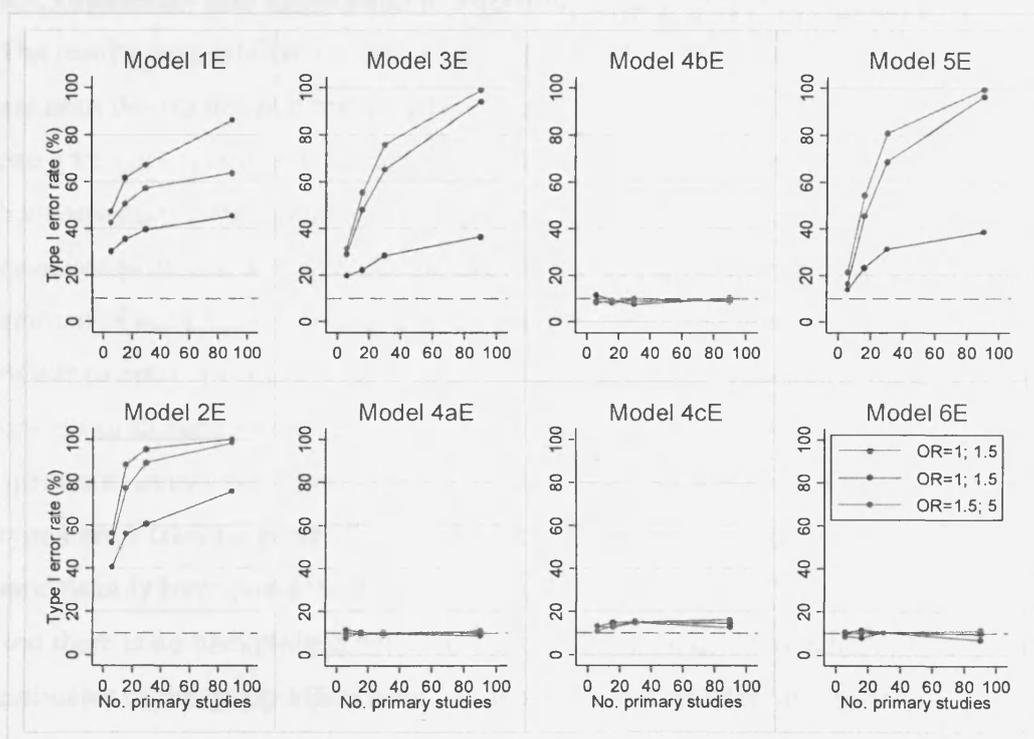
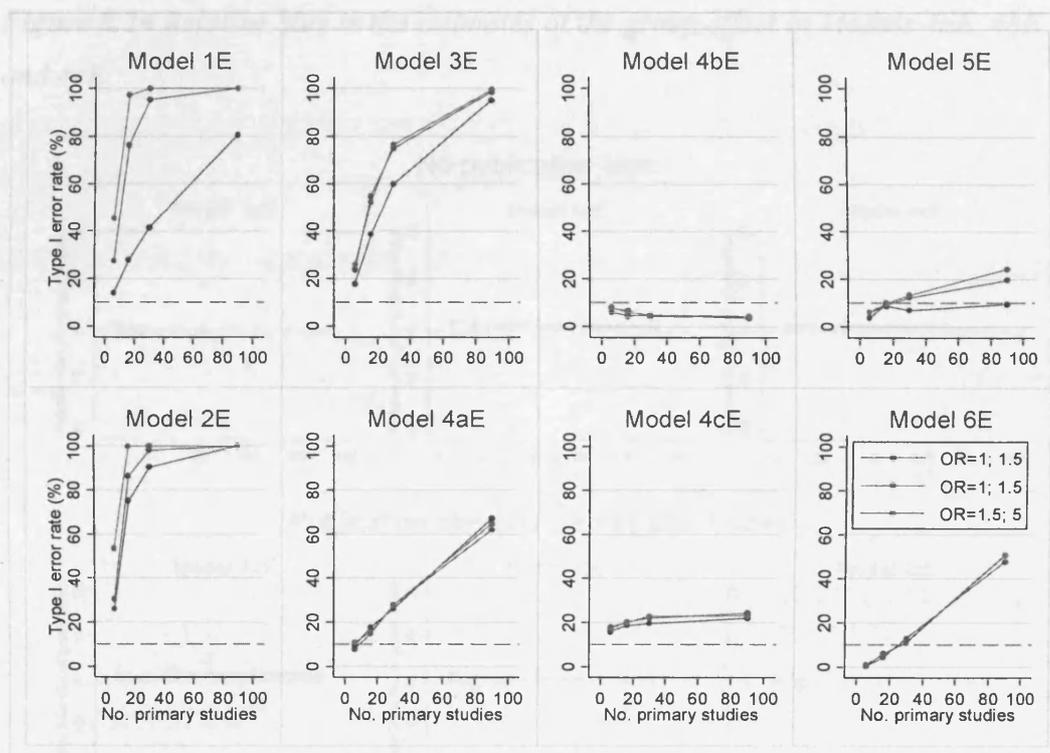


Figure 8.23 Power (when the permutation test is used to calculate the p -values) of eight extended regression tests to detect 'severe' publication bias (induced by effect size) when there is explainable between-study heterogeneity but no unexplained between-study heterogeneity

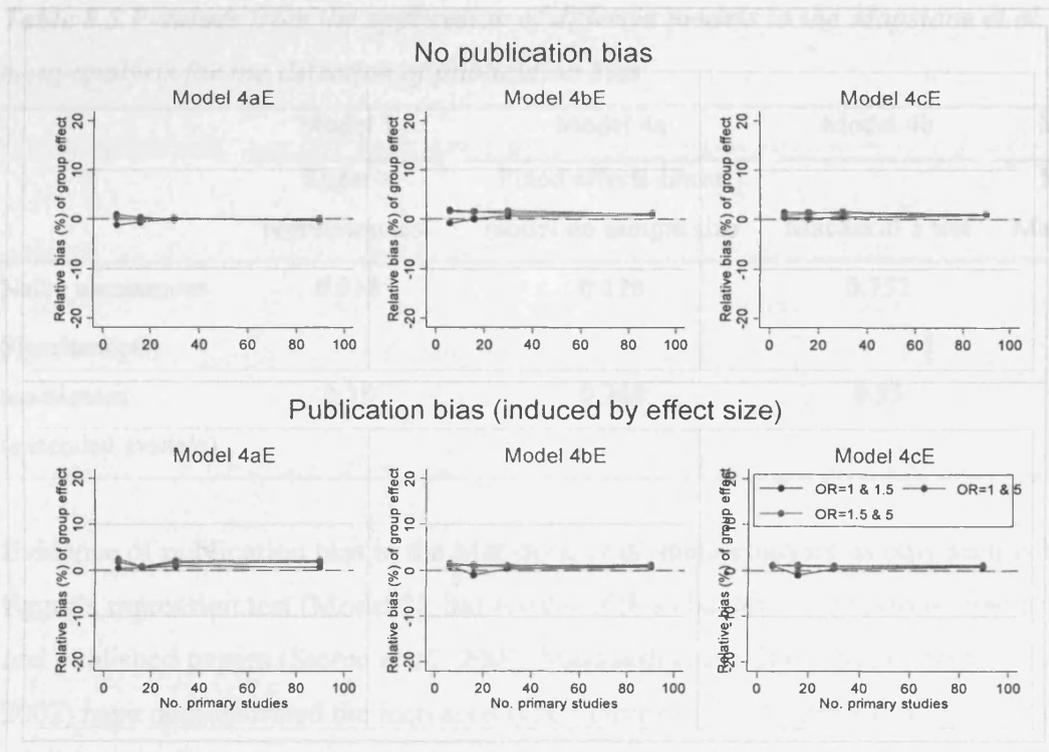


8.4.5 Summary and application to Mapstone et al. (2003) meta-analysis

The results suggest that a naïve assessment of publication bias, that incorrectly assumes the studies in a meta-analysis are all from the same underlying distribution, can give misleading results. This is particularly so when a great deal of unexplained between-study heterogeneity is also present. When using the permutation test to calculate p -values, Model 6 has the expected type I error rates regardless of the amount of unexplained between-study heterogeneity, and seems to have some power to detect publication bias in a particular scenario. Extending these regression models to include a term to account for explainable between-study heterogeneity appears to improve the performance of Models 4a, 4b and 4c in terms of appropriate type I error rates (compare Figures 8.11 and 8.17). However, these extended models only have power to detect publication bias when it is induced by effect size and there is no unexplained between-study heterogeneity. The relative bias in the estimates of the group effect from Models 4aE, 4bE and 4cE are given in

Figure 8.24 for the case where these models appear to perform well, i.e. in the absence and presence of publication bias (induced by effect size), when there is no unexplained between-study heterogeneity.

Figure 8.24 Relative bias in the estimates of the group effect in Models 4aE, 4bE and 4cE



It seems reasonable to attempt to simultaneously model publication bias and explainable between-study heterogeneity using these extended models, and as Figure 8.24 shows in the presence of publication bias, the predictable between-study heterogeneity (here the group effect) can be estimated without bias in models that performed well in terms of type I error rates and power for the detection of publication bias (Models 4aE, 4bE and 4cE). Thus, simultaneously modelling publication bias and explainable between-study heterogeneity may be of potential use, however in this chapter such models have only been found to perform well in a limited scenario (publication bias induced by effect size and no unexplained between-study heterogeneity) and it is questionable as to how realistic this scenario is in practice.

For comparison Model 1 (Egger's regression test), Model 4a (the fixed effects linear regression model on sample), Model 4b (Macaskill's FPV model) and Model 4c (the modified Macaskill test that performed well in Chapter 7) and their extended versions (Models 1E, 4aE, 4bE and 4cE) are applied to the Mapstone *et al.* meta-analysis (Table 8.5).

Table 8.5 *P-values from the application of different models to the Mapstone et al meta-analysis for the detection of publication bias*

	Model 1	Model 4a	Model 4b	Model 4c
	Egger's regression test	Fixed effects linear model on sample size	Macaskill's test	Modified Macaskill test
Naïve assessment	0.018	0.126	0.752	0.647
Simultaneous assessment (extended models)	0.35	0.268	0.55	0.695

Evidence of publication bias in the Mapstone *et al.* meta-analysis, is only seen with Egger's regression test (Model 1), but results of this chapter, the previous chapter and published papers (Sterne *et al.*, 2000; Macaskill *et al.*, 2001; Schwarzer *et al.*, 2002) have demonstrated the increased type I error rates of Egger's test. Further interpretation of these p-values from the naïve and simultaneous publication bias assessments should be made cautiously as Models 4aE, 4bE and 4cE only demonstrate favourable properties in limited settings. Moreover, in the Mapstone *et al.* (2003) meta-analysis there are five subgroups (Sprague-Dawley rats, Hebrew University rats, Wistar rats, pigs and sheep) within the meta-analysis that explain some of the between-study heterogeneity, whereas in the simulations described in this chapter only two subgroups have been considered. Thus, although there is some potential as to the use of the extended Models 4aE, 4bE and 4cE, further work, including assessing the performance of these models when there are more than two subgroups, is needed.

8.5 Summary

The results of the simulation analyses described in this chapter suggest that regression models for publication bias do not perform well when between-study heterogeneity, explainable by a measured study-level covariate, is present. This is regardless of whether the usual t-test or the permutation test is used to calculate p-values from the regression models. There is, however, some evidence that models extended to simultaneously account for publication bias and explainable between-study heterogeneity may have the potential to assist in detecting publication bias when explainable between-study heterogeneity exists (i.e. Models 4aE, 4bE and 4cE). In fact preliminary findings suggest that these models have a great deal of power to distinguish between publication bias and explainable between-study heterogeneity when the meta-analysis is very large (1000 studies), suggesting that the observed poor performance is due to lack of power.

However, as pointed out, these results are only relevant for the simulation analyses defined in this chapter. In particular, the way in which publication bias was induced here may not be realistic. By inducing publication bias within each group in a meta-analysis, it is assumed that publication bias depends on the knowledge that a particular study comes from a particular underlying distribution (i.e. from a particular subgroup). Since publication bias depends on beliefs on what is interesting, important and/or favourable according to authors, editors and reviewers, these subgroups may not be recognised and so the way in which publication bias has been induced in this chapter may not fully reflect reality.

Furthermore, as noted in Chapter 7, publication bias probably depends upon both the p-value *and* the effect size, in addition to a number of other mechanisms that can or cannot be measured. Using evidence on reasons why some articles get published and others do not (e.g. Kerr *et al.*, 1977; Kupfersmid *et al.*, 1991; Frank, 1994; Weber *et al.*, 1998), simulations based on these assumptions could be attempted. Nevertheless, there are many possibilities and uncertainties in the publication process, and so, as with the analyses carried out in this chapter and in Chapter 7, the extent to which the simulations would reflect reality is unknown.

However, the fact that findings in this chapter are consistent regardless of whether publication bias is induced by p-value or effect size, gives some confidence to the robustness of these results.

The simulation analyses described in this and the previous chapter offer the best approach to assessing and comparing the performance of the tests and methods for dealing with publication bias since analytical solutions cannot be obtained.

However, one of the limiting factors of simulation analyses is that they do not provide the reasons for why one test may perform better than another in a particular scenario. Often more detailed investigation is required to understand why, for instance, Model 4c performs better than Models 4a and 4b. Thus, such investigations would form part of further work.

Although the models assessed in this chapter have generally demonstrated poor performance when assessing publication bias, the scenarios outlined in this chapter are realistic scenarios: between-study heterogeneity (explained to some extent by a measured covariate, e.g. different species or strains used in an experiment) and publication bias can be simultaneous features of a meta-analysis. Therefore ignoring one or both in the meta-analysis can give rise to misleading results. In human health risk assessments explainable between-study heterogeneity is likely to exist, thus as this chapter has shown, it is important that an assessment of publication bias takes into account the likely between-study heterogeneity. This was demonstrated in Section 8.2 with the Mapstone *et al.* (2003) meta-analysis, however, the question still remains as to how best to assess between-study heterogeneity or publication bias in the presence of the other. There has been little work in this area, although one technique that has been proposed is the covariate-adjusted funnel plot. Petticrew *et al.* (1999) present an adjusted funnel plot for the assessment of publication bias. Using regression, Petticrew *et al.* adjusted the funnel plot by study quality, the covariate believed to explain the observed between-study heterogeneity. The covariate-adjusted funnel plot suggested little evidence of publication bias.

Although not a formal test, this approach is analogous to the use of the extended regression models (Models 1E – 6E). In both cases, the data in the meta-analysis

are adjusted for the covariate effect and then assessed for publication bias. In Models 1E – 6E publication bias is tested, in the covariate-adjusted funnel plots it is subjectively assessed. However, since few of the extended models performed well in the different scenarios, the performance of the covariate-adjusted funnel plot will need to be fully investigated.

Given that the extended regression models did not perform well (when the size of the meta-analysis reflects that likely in reality), rather than attempting to account simultaneously for between-study heterogeneity and publication bias, it may be more informative to stratify the data by the covariate believed to explain the excess between-study heterogeneity and assess publication bias within each subgroup. This was carried out in Section 8.2.2 for the Mapstone *et al.* (2003) meta-analysis. However, the simulations in this chapter have only considered categorical study-level covariates to explain between-study heterogeneity. In some situations, these study-level covariates may be continuous (e.g. distance from the equator in a meta-analysis of BCG vaccine (cited in Sutton *et al.*, 2000)) and the performance of these extended models should be assessed in these scenarios.

Such an approach may be more relevant as one is unlikely to be interested in the overall pooled effect from a meta-analysis, if a covariate can explain heterogeneity between studies. In this case, one is more likely to be interested in the findings within each subgroup and so an assessment of publication bias within these may be more appropriate. However, this approach will result in a significant loss of power since subgroups, such as species and strain, will contain fewer studies and so tests will be less likely to be able to detect publication bias if it occurs (Begg and Mazumdar, 1994; Sterne *et al.*, 2000; Macaskill *et al.*, 2001; Peters *et al.*, 2006 (Appendix H)). The application of systematic review and meta-analysis methods in human health risk assessments provides a quantitative framework for the assessment of between-study heterogeneity and publication bias. Although the characteristics of such meta-analyses (e.g. large between-study heterogeneity due to species differences) may mean that assessment of publication bias is not at all straightforward, as this chapter and Chapter 7 have shown, investigation into publication bias will nonetheless be greatly facilitated by the systematic review and meta-analysis framework.

Discussion

9.1 Thesis summary

Human health risk assessments for environmental exposures involve the evaluation of a great deal of often quite diverse evidence. Current methods used in risk assessment have their limitations; wider use of systematic review and meta-analysis methods could help to overcome them. The use of these methods in the risk assessment of exposure to environmental chemicals is different from many of the other areas where systematic reviews and meta-analyses methods have been applied (e.g. clinical interventions, epidemiological associations), because evidence from animal experiments is often used in addition to human evidence. This has led to particular consideration of the extent, quality and issues involved in systematic reviews and meta-analyses of animal experiments in this thesis (Chapter 4).

Findings from a systematic review of these articles indicate an increase in the past few years in the publication of systematic reviews and meta-analyses of animal evidence. However, an assessment of the reporting of these articles has found many to be of poor quality, especially when compared to sets of meta-analyses of human RCTs. Based on QUOROM, guidelines have been proposed in this thesis to help improve the reporting and conduct of systematic reviews and meta-analyses of animal experiments. However, it is imperative that the quality and reporting of the original animal experiments should also improve since the quality of a meta-analysis depends on the quality of the primary studies.

The next step in assessing the potential of systematic reviews and meta-analyses to help in risk assessments of environmental exposures was to apply these, and more sophisticated cross design synthesis, methods to two risk assessments of the human health effects from exposure to environmental chemicals (Chapters 5 and 6). In one example these methods were successfully applied throughout the risk assessment process, from identifying the relevant evidence through to the derivation of a 'safe'

exposure limit (Chapter 5: total THMs and low birth weight). In the second example however, the data were much more diverse and sparse, and so the application of meta-analysis methods was of limited value in the evaluation of the evidence (Chapter 6: Mn and neurobehavioural effects). Even so, the approach taken in Chapter 6 demonstrated the advantages of a systematic review, even though a meta-analysis across species was not feasible. If these methods are adopted in the human health risk assessment context, issues of between-study heterogeneity and publication bias will need to be addressed. In this thesis an improved method for the detection of publication bias was proposed (Chapter 7). Approaches for detecting publication bias when differences between studies may be explained by study-level covariates (a situation potentially common in the synthesis of evidence for a human health risk assessment, e.g. species, exposure route) have also been assessed and findings suggest one should proceed with caution in such situations (Chapter 8).

In the following three sections, the main issues arising in this thesis – i) the application of systematic review and meta-analysis methods to animal evidence, ii) the cross-design synthesis of human and animal evidence in environmental exposure risk assessment contexts, and iii) methods for the detection of publication bias – are discussed in the context of the thesis aims, considering possible areas for further work.

9.2 Systematic reviews and meta-analyses of animal experiments

As many have pointed out, systematic reviews have the potential to help summarise, review and evaluate data from multiple animal experiments (Pound *et al.*, 2004; Roberts *et al.*, 2002a; Roberts and Sandercock, 2002; Horn *et al.*, 2001). Their use may also lead to a decrease in the number of animal experiments, and human studies, carried out (Horn *et al.*, 2001), strengthening the UK 3Rs aim of reducing the number of animal experiments. However, no assessment of the quality of the reporting of these methods for evaluating animal experiments had previously been carried out to assess their use in practice. In this thesis a systematic review of

systematic reviews and meta-analyses of animal experiments (Chapter 4) found that much improvement is needed in the conduct and reporting of these methods to evaluate animal evidence. In particular, the meta-analyses of animal experiments concerning environmental exposures tended to be of poor quality, with none of them being preceded by a systematic review. Although the sources of evidence were described in some of the environmental exposure meta-analyses (e.g. multi-laboratory experiments, databases of toxicology results), many others did not identify the origin of the data. The fact that there has been less use of systematic review and meta-analysis methods to combine human environmental epidemiological studies than human RCTs for medical interventions (Blair *et al.*, 1995) may have impacted upon their uptake in the review of animal evidence for environmental exposures. For instance, the systematic reviews and meta-analyses of animal experiments assessing medical interventions reviewed in Chapter 4 tend to use approaches similar to those used in systematic reviews and meta-analyses of human RCTs. This suggests that systematic review and meta-analysis methods used to synthesise animal evidence in medical interventions have translated from those used in the synthesis of human RCTs.

The systematic review described in Chapter 4 was based on the reports of systematic reviews and meta-analyses of animal experiments, so a distinction between the conduct and the reporting of the articles could not be made. This point is clearly made by the number of articles initially thought to be relevant, but subsequently excluded from the review as they did not meet the pre-specified inclusion criteria. For instance, as pointed out in Chapter 4, many articles referred to themselves as systematic reviews of animal experiments, but provided very few, or no, details on the systematic review process. Hence, a clear distinction needs to be made between articles excluded because they are not reports of systematic reviews or meta-analyses, and articles excluded because the systematic review process was poorly reported.

In addition to there being no guidelines for reporting of systematic reviews and meta-analyses of animal experiments, a possible factor in the poor reporting of these articles is that, unlike human RCTs, published guidelines for the reporting of *primary* animal experiments are not available. There is evidence to suggest that

such guidelines are associated with improved reporting in human primary studies (Moher *et al.*, 2001) and a number of publications and websites do offer guidance for the conduct and reporting of primary animal experiments. For instance, a whole issue of the ILAR journal (Institute for Laboratory Animal Research; Vol 43, No 3, 2002) is dedicated to giving guidance on both the design and the analysis of animal toxicology studies and an approach for evaluating the quality of toxicological studies has been proposed (Klimisch *et al.*, 1997). In the UK, the National Centre for the Replacement, Refinement and Reduction of Animals in Research is working towards improvements in experimental design (<http://www.nc3rs.org.uk/>) and guidelines for animal testing are also available from a number of US and international organisations, such as the Center for Alternatives to Animal Testing (CAAT) based at the John Hopkins University (<http://caat.jhsph.edu/>) and the Organisation for Economic Co-operation and Development (OECD) (http://www.oecd.org/departement/0,2688,en_2649_34377_1_1_1_1_1,00.html). Initiatives to improve the quality of reporting of the primary studies should go hand-in-hand with both the conduct and reporting of systematic reviews and/or meta-analyses of animal experiments. It is hoped that, initially, the guidelines proposed in Figure 4.7 should help to improve the quality of both the conduct and reporting of systematic reviews and meta-analyses of animal experiments. Further work would involve much discussion and revision of these guidelines, with researchers carrying out systematic reviews and meta-analyses of animal experiments on a regular basis to increase their relevance to animal experiments, leading to improvements in the quality of reporting. This could, and should, lead to a more efficient use of the animal evidence to inform human health, while contributing to the 3Rs programme if researchers, particularly those involved in environmental exposure outcomes, are willing to adopt these methods.

9.3 Synthesis of human and animal evidence

When it comes to the review and evaluation of both human and animal evidence, there are likely to be few objections to the use of systematic review methods for this purpose. As has been pointed out, systematic reviews which have been conducted thoroughly ensure all relevant data are included and help identify areas where more

evidence is required. Thus, advantages of a systematic review approach are greater than the more narrative, qualitative approaches to human health risk assessment described in Chapter 2. The application of meta-analysis for the synthesis of human and animal evidence is another issue and one that has, so far, received little attention. To synthesise relatively diverse evidence an understanding is required on the assumptions made to 1) obtain comparable, meaningful estimates of effect from each study, and 2) to combine these estimates in a sensible and constructive way, minimising the potential for bias in the overall estimate of effect. For some, these assumptions may initially seem to be too strong. However, the assumptions are parallel to those currently made in the risk assessment process, and through the transparent nature of systematic reviews and meta-analyses these assumptions are forced to be made explicit. DuMouchel and colleagues (DuMouchel and Harris, 1983; DuMouchel and Groer, 1989) showed how the use of a Bayesian model provides a flexible and realistically complex framework for the synthesis of human and animal evidence to inform human health risks from environmental exposures which allow explicit modelling of the relevance of one species to another in the prior distributions.

In this thesis, alternative Bayesian models have been used to illustrate the potential of systematic review and meta-analysis methods to synthesise human and animal evidence for a human health risk assessment (Chapters 5 and 6). These Bayesian models have been previously applied to different sources of human evidence (Prevost *et al.*, 2000; Sutton and Abrams, 2001). In this thesis, their use to combine animal and human evidence to obtain a 'better' estimate of effect in humans, rather than an overall species effect, has been illustrated. Extensive sensitivity analyses, as in Chapter 5, have demonstrated the necessity of investigating the impact of the many different assumptions made in such a synthesis, and this has been greatly facilitated in the quantitative framework provided by the use of meta-analysis methods. However, in practice, the use of meta-analysis may be restricted by the diversity of available relevant evidence as illustrated in Chapter 6 with the Mn example. In the Mn example, meta-analysis methods could only be used to a certain extent because the available evidence concerned such a diverse set of studies and experiments. Nevertheless, use of these systematic review and meta-analysis methods facilitated the evaluation of the evidence in the Mn example by allowing

synthesis of similar sorts of evidence within the whole evidence base, thus reducing the amount of evidence to be considered. The presentation of evidence was also facilitated by a systematic review and meta-analysis approach such that data were displayed in a clear and concise manner, allowing inconsistencies to be identified (see Table 6.7). However, estimation towards a 'safe' exposure limit for Mn was limited by the inconsistent reporting of exposure data in the primary studies and concerns related to the ecological fallacy (see Section 6.5.2), thus highlighting advantages in obtaining individual (human and animal) subject data.

Within a meta-analysis framework, of the kind presented in Chapters 5 and 6, the number and type of further studies needed to inform the human health risk assessment could also be considered. For instance, the Expected Value of different types of Information (EVI) in the risk assessment could be calculated and used to inform the direction of future research (Claxton, 1999), e.g. should a human epidemiological study be carried out or an animal experiment to provide the most information. This approach already has applications in the decision-making process for new medical technologies (Claxton *et al.*, 2002; Tappenden *et al.*, 2004) and it is not difficult to see how the application of EVI to the risk assessment examples in Chapters 5 and 6 could form further work in this area. Such an application would have potentially important implications for funding bodies and for the design of future human studies *and* animal experiments.

Since there has been very little research into the use of cross-design synthesis methods for human and animal evidence in the risk assessment context, there is a great deal of scope for further work. More immediate work could involve returning to the THMs example (Chapter 5) and updating the synthesis models to account for the more recent human and animal evidence now available (Christian *et al.*, 2002; Toledano *et al.*, 2005; Wright *et al.*, 2005). Further modelling of the THMs example could involve i) describing and estimating the relationship between the THM exposures (i.e. how total THMs, chloroform, BDCM, CDBM and bromoform are related), ii) broadening the health outcomes to account for, and model, related effects (such as premature birth and stillbirth in addition to low birth weight), iii) applying model averaging techniques (Kass and Raftery, 1995; Sutton and Abrams, 2001) to the dose-response models (as mentioned in Section 5.9) and iv)

comparing the fit of the synthesis models, not just the dose-response models as in Section 5.7.1. Further work has already begun in the use of the Bayesian models proposed by DuMouchel (DuMouchel and Harris, 1983) for the THMs example and this will be followed by critical comparison of different Bayesian hierarchical models to synthesise the relevant evidence (Peters *et al.*, 2003).

Further work in the Mn example requires careful consideration of the exposure levels in the relevant studies and experiments. As noted above, and in Chapter 6, estimation of a 'safe' exposure limit could not be carried out using the current evidence, but use of individual subject data would greatly facilitate this assessment. However, obtaining the individual subject data is unlikely to be a trivial task and its costs and benefits could also be assessed from an EVI approach alongside consideration of whether a human study or animal experiment is likely to provide the most valuable, evidence given the likely costs of obtaining that evidence and the funds available. Such an assessment can only be based on the existing evidence and so further work in the Mn example would involve more informed synthesis of evidence across species and health effects, accounting for multiple outcomes in the synthesis, as in Riley *et al.* (2006), and differences in effects between species. This would also provide a larger framework for investigation of possible differential publication bias in the generalised synthesis of this evidence.

9.4 Publication bias in the presence of between-study heterogeneity

Regardless of whether human or animal evidence are being synthesised in a meta-analysis, between-study heterogeneity and publication bias are both common and important features of meta-analyses (Engels *et al.*, 2000; Sutton *et al.*, 2000; Egger *et al.*, 2001; Villar *et al.*, 2001). Ignoring either can have implications for the findings and conclusions made from a meta-analysis no matter what the application is. If exposure limits are set on the basis of biased evidence the implications could have a large impact for the general population. When they are simultaneously present, assessment and interpretation of one of these may be affected by the other. However, investigating the possible presence of between-study heterogeneity and/or

publication bias must be carried out and implications of the findings of the meta-analysis must be explored. In Chapter 7 an improved test for publication bias, based on the inverse of sample size and an alternative specification for study weighting (Model 4c), was proposed and found to have more appealing properties (i.e. appropriate type I error rates and reasonable power) compared to all of the regression tests assessed, including the commonly used Egger's regression test (Peters *et al.*, 2006). Based on the inverse of the sample size, this alternative test avoids the correlation induced between an estimate of lnOR and its standard error and also avoids possible problems of regression dilution bias, since sample size is a known quantity whereas standard error is estimated. Although initial results suggest that the modified Macaskill regression test is superior to Egger's regression test when the relative risk is the summary estimate, further investigation into the performance of Model 4c is needed under different conditions (e.g. when the event is rare, when the number of subjects in a study is small). In addition to this, a comparison of Model 4c with models recently put forward by Harbord *et al.* (2006) and Copas *et al.* (2005) is required with recommendations on choice of model in different circumstances.

The scenarios in Chapter 7 were extended in Chapter 8 to include investigation of publication bias when some, or all, of the between-study heterogeneity can be explained by a measured study-level covariate, e.g. species and strain of animal. Results indicate that when this *explainable* between-study heterogeneity is not modelled, none of the eight regression tests or the rank correlation test (Begg and Mazumdar, 1994) assessed here performed well. Extended versions of fixed effects linear regression models based on sample size, and a modification of Model 4c showed some potential for identifying publication bias and correctly estimating species effects, but only in limited scenarios (with preliminary findings suggesting that for very large meta-analyses these models could be quite powerful). Simultaneous modelling of publication bias and explainable between-study heterogeneity has important advantages of power over individual assessment of publication bias within homogenous subgroups. Such models are, therefore, worth pursuing in addition to more subjective methods such as the covariate-adjusted funnel plots mentioned in Section 8.5.

It must be noted, however, that all of the results from Chapters 7 and 8 are confined to the meta-analysis scenarios simulated in those chapters. For instance, because a number of factors are thought to influence publication bias, actually inducing publication bias realistically is difficult. Many authors induce publication bias based on p-value (Macaskill *et al.* 2001, Hedges and Vevea, 1996; Begg and Mazumdar, 1994) or on effect size (Duval and Tweedie, 2000b). In this thesis both mechanisms have been used, but any modelling of publication bias is unlikely to capture all aspects of the process. In reality, publication bias is likely to be based on both the effect size *and* the p-value from a primary study, in addition to many other possible factors. It is, therefore, difficult to make general conclusions on the power of the tests assessed in this thesis, but similar performance across different mechanisms of publication bias (as seen with Model 4c in Section 7.6.2) gives some confidence to the generalisability of these findings.

Just testing for the presence of publication bias is not enough. In the survey of meta-analyses published in JAMA and BMJ (described in Section 7.3), a number of approaches were taken in these articles when publication bias was suspected. These ranged from excluding studies from funnel plots and claiming no evidence of publication bias; acknowledging possible publication bias but giving no detail of its extent or impact; carrying out sensitivity analyses on the impact of possible publication bias (e.g. comparing unadjusted estimates to trim and fill adjusted estimates); and discussion of the implications, advising caution in the interpretation of the findings of the articles and calling for larger, well-designed studies to be carried out.

An ideal approach, taken by a number of the meta-analyses reviewed in JAMA and the BMJ, would be to assess the impact suspected publication bias may have on the overall estimate of effect. None of the three approaches assessed in Chapter 7 to obtain an estimate of the underlying effect ((i) the trim and fill method, (ii) using the estimate from the largest study, (iii) using the estimate from the most precise study) performed particularly well in terms of how well the estimate of effect compares to the underlying effect (i.e. the bias in the estimate), especially when unexplainable between-study heterogeneity was present. Although it may seem appealing to use the largest study, or studies, in a meta-analysis and believe them to be less likely to

be affected by publication bias, when a great deal of between-study heterogeneity is induced, the findings of these simulations suggest that publication bias does impact upon the larger studies. These findings are similar to those of Stern and Simes (1997), who still found evidence of publication bias in their meta-analysis of 'large' studies.

These, and other, approaches such as Copas' sensitivity method (Copas and Jackson, 2004) provide a means for sensitivity analyses when publication bias is suspected. However, in many situations decisions need to be made on the evidence available, even if it is heterogeneous and/or likely to be subject to publication bias. Future work needs to consider not just the appropriateness and performance of particular tests for publication bias but also investigation into methods to obtain appropriately adjusted estimates of effect reflecting the likely impact of publication bias. Until this is done, meta-analyses need to be interpreted with care, especially when between-study heterogeneity and/or publication bias are suspected. As such further research in this area would extend the scenarios reported in Chapter 8 to that where the between-study heterogeneity covariates are continuous rather than binary. The performance of the tests in these scenarios can then be examined and compared to the results given in Chapter 8 and to inform recommendations for practice and future investigations into publication bias in the presence of between-study heterogeneity.

9.5 Conclusions

This thesis has illustrated and critically explored the potential for systematic review and meta-analysis methods to assist in a human health risk assessment of environmental exposures and contributed to the development of methods for doing so. In particular, methods for the synthesis of human and animal evidence have been illustrated (Peters *et al.*, 2005) using Bayesian hierarchical models only previously used to combine different sources of human evidence. However, improvements on current practice are needed, especially in the conduct and reporting of systematic reviews and meta-analyses of animal experiments, as is an acknowledgement of the limitations of meta-analyses in particular. Current use of

these methods to review animal evidence suggests that improvements in reporting need to be made to make better use of the evidence to inform human health. It is hoped that guidelines (given in Chapter 4) for good quality reporting of systematic reviews and meta-analyses of animal experiments will help to do this (Peters *et al.*, accepted). As Chapter 6 demonstrated, systematic review and meta-analysis methods have an important role in the search, review and evaluation of the evidence for a risk assessment of environmental exposures. However, because of the diverse, and sometimes sparse, nature of the evidence available for a human health risk assessment full cross-design synthesis may not be appropriate or achievable. Nevertheless, advantages of such a framework far out-weigh the alternatives including narrative reviews, decreased power, and inefficient use of all the relevant evidence. Using systematic review and meta-analysis methods, features such as study quality, between-study heterogeneity and publication bias can be assessed and their impact on results and subsequent decisions investigated. Work carried out in Chapter 7 has contributed to the debate on how best to identify publication bias in different meta-analysis scenarios (Peters *et al.*, 2006) and publication of the results from Chapter 8 will build on this for scenarios where at least some between-study heterogeneity can be explained. As discussed in Section 9.3, use of systematic review and meta-analysis methods may lead on to more informed use of time, energy and money by directing future research using a value of information approach. Human health risks from exposure to environmental chemicals can only be assessed when all the relevant evidence, regardless of its sources, are considered together. Systematic reviews and meta-analyses can help to make this process more structured and transparent, ultimately leading to a more efficient informed assessment of human health risks to environmental chemicals.

Bibliography

Abd El Naby S, Hussanein M (1965) Neuropsychiatric manifestations of chronic manganese poisoning. *Journal of Neurology, Neurosurgery and Psychiatry* 28:282-287.

Abrams KR, Gillies CL, Lambert PC (2005) Meta-analysis of heterogeneously reported trials assessing change from baseline. *Statistics in Medicine* 24:3823-3844.

Adar R, Critchfield G, Eddy DM (1989) A confidence profile analysis of the results of femoropopliteal percutaneous transluminal angioplasty in the treatment of lower extremity ischemia. *Journal of Vascular Surgery* 10:57-67.

Ades AE, Sutton AJ (2006) Multiparameter evidence synthesis in epidemiology and medical decision-making: current approaches. *Journal of the Royal Statistical Society Series A, Statistics in Society* 169:5-36.

ATSDR (2000) *Toxicological Profile for manganese*. Agency for Toxic Substances and Disease Registry. US Department of Health and Human Services, Atlanta GA, USA.

Babapulle MN, Joseph L, Belisle P, Brophy JM, Eisenberg MJ (2004) A hierarchical Bayesian meta-analysis of randomised clinical trials of drug-eluting stents. *Lancet* 364:583-591.

Bae K, Mallick BK (2004) Gene selection using a two-level hierarchical Bayesian model. *Bioinformatics* 20:3423-3430.

Baird SJ, Cohen JT, Graham JD, Shyakhter AI, Evans JS (1996) Noncancer risk assessment: probabilistic characterization of population threshold doses. *Journal of Human and Ecological Risk Assessment* 2:79-102.

- Begg CB, Mazumdar M (1994) Operating characteristics of a rank correlation test for publication bias. *Biometrics* 50:1088-1101.
- Begg CB, Pilote L (1991) A model for incorporating historical controls into a meta-analysis. *Biometrics* 47:899-906.
- Berlin JA, Begg CB, Louis TA (1989) An assessment of publication bias using a sample of published clinical trials. *Journal of the American Statistical Association* 84:381-392.
- Bertani H, Gelmini R, Del Buono MG, De Maria N, Girardis M, Solfrini V, Villa E (2002) Literature overview on artificial liver support in fulminant hepatic failure: a methodological approach. *International Journal of Artificial Organs* 25:903-10.
- Beuter A, Edwards R, deGeoffroy A, Mergler D, Hudnell K (1999) Quantification of neuromotor function for detection of the effects of manganese. *Neurotoxicology* 20:355-366.
- Bickell WH, Bruttig SP, Millnamow GA, O'Benar J, Wade CE (1992) Use of hypertonic saline/dextran versus lactated Ringer's solution as a resuscitation fluid after uncontrolled aortic hemorrhage in anesthetised swine. *Annals of Emergency Medicine* 21:1077-1085.
- Biondi-Zoccai GGL, Abbate A, Parisi Q, Agostoni P, Burzotta F, Sandroni C, Zardini P, Biasucci LM (2003) Is vasopressin superior to adrenaline or placebo in the management of cardiac arrest? A meta-analysis. *Resuscitation* 59:221-224.
- Blair A, Burg J, Foran J, Gibb H, Greenland S, Morris R, Raabe G, Savitz D, Teta J, Wartenberg D, Wong O, Zimmerman R (1995) Guidelines for application of meta-analysis in environmental epidemiology. *Regulatory Toxicology and Pharmacology* 22:189-197.
- Blettner M, Sauerbrei W, Shlehofer B, Scheuchenpflug T, Friedenreich C (1999) Traditional reviews, meta-analyses and pooled analyses in epidemiology. *International Journal of Epidemiology* 28:1-9.
- Boffetta P, Silverman DT (2001) A meta-analysis of bladder cancer and diesel exhaust exposure. *Epidemiology* 12:125-130.

- Bonilla E (1984) Chronic manganese intake induces changes in the motor activity of rats. *Experimental Neurology* 84:696-700.
- Borenstein M (2005) Software for Publication Bias, in *Publication Bias in Meta-Analysis: Prevention, Assessments and Adjustments*, (Rothstein HR, Sutton AJ, Borenstein M eds), pp 193-220. Wiley, Chichester, England.
- Borrelli F, Ernst E (2002) Cimicifuga racemosa: a systematic review of its clinical efficacy. *European Journal of Clinical Pharmacology* 58:235-41.
- Borrelli F, Izzo AA, Ernst E (2003) Pharmacological effects of Cimicifuga racemosa. *Life Sciences* 73:1215-29.
- Bove FJ, Fulcomer MC, Klotz JB, Dufficy EM, Zagraniski RT (1992) Report on Phase IV-A: Public drinking water contamination and birthweight, fetal deaths and birth defects. New Jersey Department of Health, US.
- Boyd HA, Flanders WD, Addiss DG, Waller LA (2005) Residual spatial correlation between geographically referenced observations: a Bayesian hierarchical modeling approach. *Epidemiology* 16:532-541.
- Bracken MB (2001) Commentary: toward systematic reviews in epidemiology. *International Journal of Epidemiology* 30:954-957.
- Brooks SP, Gelman A (1998) General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* 7: 434-455
- Brown DSO, Wills CE, Yousefi V, Nell V (1991) Neurotoxic effects of chronic exposure to manganese dust. *Neuropsychiatry, Neuropsychology and Behavioral Neurology* 4:238-250.
- Brown KG, Strickland JA (2003) Utilizing data from multiple studies (meta-analysis) to determine effective dose-duration levels. Example: rats and mice exposed to hydrogen sulfide. *Regulatory Toxicology and Pharmacology* 37:305-17.
- Budtz-Jørgensen E, Keiding N, Grandjean P (2001) Benchmark dose calculation from epidemiological data. *Biometrics* 57:698-706.

- Calabrese EJ, Baldwin LA, Holland CD (1999) Hormesis: a highly generalizable and reproducible phenomenon with important implications for risk assessment. *Risk Analysis* 19:261-81.
- Camerino D, Cassitto MG, Gilioli R (1993) Prevalence of abnormal neurobehavioural scores in populations exposed to different industrial chemicals. *Environmental Research* 61:251-257.
- Cantor KP, Lynch CF, Hildesheim ME, Dosemeci M, Lubin J, Alavanja M, Craun G (1999) Drinking water source and chlorination byproducts in Iowa. III. Risk of brain cancer. *American Journal of Epidemiology* 150:552-560.
- Carroll RJ, Simpson DG, Zhou H (1994) Stratified ordinal regression: a tool for combining information from disparate toxicological studies. National Institute of Statistical Sciences, Research Triangle Park, US.
- CDC (2005) Third National Report on Human Exposure to Environmental Chemicals. Centres for Disease Control and Prevention, Atlanta, Georgia.
- Chalmers I, Hedges LV, Cooper H (2002) A brief history of research synthesis. *Evaluation and the Health Professionals* 25:12-37.
- Chan AW, Krieza-Jeric K, Schmid I, Altman DG (2004) Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. *Canadian Medical Association Journal* 171:735-740.
- Chan A-W, Altman DG (2005) Identifying outcome reporting bias in randomised trials on PubMed: reviews of publications and survey of authors. *British Medical Journal* 330:753.
- Chandra SV (1972) Histological and histochemical changes in experimental manganese encephalopathy in rabbits. *Archives of Toxicology* 29:29-38.
- Chandra AV, Ali MM, Saxena DK, Murthy RC (1981) Behavioural and neurochemical changes in rats simultaneously exposed to manganese and lead. *Archives of Toxicology* 49:49-56.

- Chinn S (2000) A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in Medicine* 19:3127-3131.
- Christensen E (2001) Quality of reporting of meta-analyses: the QUOROM statement. Will it help? *Journal of Hepatology* 34:342-345.
- Christian MS, York RG, Hoberman AM, Fisher LC, Brown WR (2002) Oral (drinking water) two-generation reproductive toxicity study of bromodichloromethane (BDCM) in rats. *International Journal of Toxicology* 21:115-146.
- Claxton K (1999) Bayesian approaches to the value of information: implications for the regulation of pharmaceuticals. *Health Economics* 8:269-274.
- Claxton K, Schulper M, Drummond M (2002) A rational framework for decision making by the National Institute for Clinical Excellence (NICE). *Lancet* 360:711-715.
- Clewell HJ, Lawrence GA, Calne DB, Crump KS (2003) Determination of an occupational exposure guideline for manganese using the benchmark method. *Risk Analysis* 23:1031-1046.
- Cogliano VJ (2005) Principles of cancer risk assessment: the risk assessment paradigm, in *Recent advances in quantitative methods in cancer and human health risk assessment*, (Edler L, Kitsos CP eds), pp 5-21. Wiley, Chichester, England.
- Cochrane Collaboration (2005 [updated 24 August 2005]) The Cochrane Manual Issue 4.
- Collins JJ, Ness R, Tyl RW, Krivanek N, Esmen NA, Hall TA (2001) A review of adverse pregnancy outcomes and formaldehyde exposure in human and animal studies. *Regulatory Toxicology and Pharmacology* 34:17-34.
- Commission of the European Communities (2001) White Paper: Strategy for a future chemicals policy.
- Cook DJ, Sackett DL, Spitzer WO (1995) Methodologic guidelines for systematic reviews of randomized control trials in health care from the Potsdam consultation on meta-analysis. *Journal of Clinical Epidemiology* 48:167-171.

- Copas J (1999) What works? Selectivity models and meta-analysis. *Journal of the Royal Statistical Society Series A, Statistics in Society* 162:95-109.
- Copas JB, Shi JQ (2001) A sensitivity analysis for publication bias in systematic reviews. *Statistical Methods in Medical Research* 10:251-265.
- Copas J, Jackson D (2004) A bound for publication bias based on the fraction of unpublished studies. *Biometrics* 60:146-153.
- Copas J, Lozada-Can C (2005) Modified Egger's test. Presented at the Methods in Meta-analysis meeting, held at the Royal Statistical Society, London. December 2005.
- Corpet DE, Tache S (2002) Ranking chemopreventive agents on rat colon carcinogenesis. *IARC Scientific Publications* 156:381-384.
- Covello VT, Merkhofer MW (1994) *Risk Assessment Methods: Approaches for Assessing Health and Environmental Risks*. Plenum Press, New York & London.
- Cox LH, Piegorsch WW (1994) Combining environmental information: environmetric research in ecological monitoring, epidemiology, toxicology, and environmental data reporting. National Institute of Statistical Science Report 12.
- Craig JC, Wheeler DM, Irwig L, Howman-Giles RB (2000) How accurate is dimercaptosuccinic acid scintigraphy for the diagnosis of acute pyelonephritis? A meta-analysis of experimental studies. *Journal of Nuclear Medicine* 41:986-93.
- Critchfield GC, Eddy DM (1987) A confidence profile analysis of the effectiveness of disulfiram in the treatment of chronic alcoholism. *Medical Care* 25:S66-S75.
- Crump KS (1984) A new method for determining allowable daily intakes. *Fundamental and Applied Toxicology* 4:854-871.
- Crump KS, Krewski D, Van Landingham C (1999) Estimates of the proportions of carcinogens and anticarcinogens in bioassays conducted by the U.S. National Toxicology Program. Application of a new meta-analytic approach. *Annals of the New York Academy of Sciences* 895:232-244.

- Deeks J, Glanville J, Sheldon T (2001) *Undertaking systematic reviews of research on effectiveness: CRD guidelines for those carrying out or commissioning reviews*. NHS Centre for Reviews and Dissemination, York, UK.
- de Nijs RNJ, Jacobs JWG, Algra A, Lems WF, Bijlsma JW (2004) Prevention and treatment of glucocorticoid-induced osteoporosis with active vitamin D₃ analogues: a review with meta-analysis of randomized controlled trials including organ transplantation studies. *Osteoporosis International* 15:589-602.
- Deschamps FJ, Guillaumot M, Raux S (2001) Neurological effects in workers exposed to manganese. *Journal of Occupational and Environmental Medicine* 43:127-132.
- Dixon E, Hameed M, Sutherland F, Cook DJ, Doig C (2005) Evaluating meta-analyses in the general surgical literature: a critical appraisal. *Annals of Surgery* 241:450-459.
- Do M, Birkett NJ, Johnson KC, Krewski D, Villeneuve P (2005) Chlorination disinfection by-products and pancreatic cancer risk. *Environmental Health Perspectives* 113:418-424.
- Dodds L, King W, Woolcott C, Pole J (1999) Trihalomethanes in public water supplies and adverse birth outcomes. *Epidemiology* 10:233-237.
- DoE (1993) *Risk Assessment of Existing Substances*. Department of the Environment, London, UK.
- Dominici F, Parmigiani G, Reckhow KH, Wolpert RL (1997) Combining information from related regressions. Technical Report. Duke University.
- Dominici F, Samet J, Zeger SL (2000) Combining evidence on air pollution and daily mortality from the largest 20 U.S. cities: a hierarchical modeling strategy (with discussion). *Journal of the Royal Statistical Society Series A, Statistics in Society* 163:263-302.
- Dominici F, Zanobetti A, Zeger SL, Schwartz J, Samet JM (2004) Hierarchical bivariate time series models: a combined analysis of the effects of particulate matter on morbidity and mortality. *Biostatistics* 5:341-360.

- Dorman DC, Struve MF, Vitarella D, Byerly FL, Goetz J, Miller R (2000) Neurotoxicity of manganese chloride in neonatal and adult CD rats following subchronic (21-day) high-dose exposure. *Journal of Applied Toxicology* 20:179-187.
- Dourson ML, Stara JF (1983) Regulatory history and experimental support of uncertainty (safety) factors. *Regulatory Toxicology and Pharmacology* 3:224-238.
- Dourson ML, Felter SP, Robinson D (1996) Evolution of science-based uncertainty factors in noncancer risk assessment. *Regulatory Toxicology and Pharmacology* 24:108-120.
- DuMont GJH, de Visser SJ, Cohen AF, van Gerven JMA (2005) Biomarkers for the effects of selective serotonin reuptake inhibitors (SSRIs) in healthy subjects. *British Journal of Clinical Pharmacology* 59:495-510.
- DuMouchel WH, Harris JE (1983) Bayes methods for combining the results of cancer studies in humans and other species. *Journal of the American Statistical Association* 78:293-308.
- DuMouchel W, Groër PG (1989) A Bayesian methodology for scaling radiation studies from animals to man. *Health Physics* 57:411-418.
- DuMouchel W (1994) Hierarchical Bayes linear models for meta-analysis. National Institute of Statistical Science 27.
- Duval S, Tweedie R (2000a) Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics* 56:455-463.
- Duval S, Tweedie RL (2000b) A nonparametric "Trim and Fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Society* 95(449):89-98.
- Duval S (2005) The trim and fill method, in *Publication bias in meta-analysis*, (Rothstein H, Sutton A, Borenstein M eds), pp127-144. Wiley, Chichester, England.
- EA, DEFRA (2002) *The contaminated land exposure assessment model (CLEA): technical basis and algorithms*. Environment Agency, Bristol, UK.

- Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR (1991) Publication bias in clinical research. *Lancet* 337:867-872.
- EC (1997) *Study of toxic effects in the central and peripheral nervous system of workers in the ferro-alloy industry*. Health and Safety at Work, European Commission, Brussels, Belgium.
- ECETOC (2001) *Exposure factors sourcebook for European populations (with focus on UK data)*. European Centre for Ecotoxicology and Toxicology of Chemicals, Brussels, Belgium.
- Eddy DM, Hasselblad V, McGivney W, Hendee W (1988) The value of mammography screening in women under age 50. *Journal of the American Medical Association* 259:1512-1519.
- Eddy DM (1989) The confidence profile methods: a Bayesian method for assessing health technologies. *Operations Research* 37:210-228.
- Edler L, Kopp-Schneider A, Heinzl H (2005) Dose-response modelling, in *Recent advances in quantitative methods in cancer and human health risk assessment*, (Edler L, Kitsos CP eds), pp 211-238. Wiley, Chichester, England.
- EEC (1967) The approximation of laws, regulations and administrative provisions relating to the classification, packaging and labelling of dangerous substances. Council Directive 67/548/EEC. European Economic Community.
- Egger M, Davey Smith G (1997) Meta-analysis: potentials and promise. *British Medical Journal* 315:1371-1374.
- Egger M, Davey Smith G, Schneider M, Minder C (1997) Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal* 315:629-634.
- Egger M, Davey Smith G, Altman DG (2001) *Systematic reviews in health care: meta-analysis in context*. BMJ, London, UK.

- Egger M, Juni P, Bartlett C, Hoenstein F, Sterne J (2003) How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. *Health Technology Assessment* 7(1).
- Eichacker PQ, Parent C, Kalil A, Esposito C, Cui X, Banks SM, Gerstenberger EP, Fitz Y, Danner RL, Natanson C (2002) Risk and the efficacy of antiinflammatory agents: retrospective and confirmatory studies of sepsis. *American Journal of Respiratory and Critical Care Medicine* 166:1197-205.
- Engels EA, Schmid CH, Terrin N, Olkin I, Lau J (2000) Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. *Statistics in Medicine* 19:1707-1728.
- EPA (1998) National primary drinking water regulations: disinfectants and disinfection byproducts. Report 63(241). US Environmental Protection Agency.
- EPA (2003) *Benchmark Dose Software*. Available at: www.epa.gov/nceawww1/bmds.htm [accessed 2nd Nov 2005]. US Environmental Protection Agency.
- Eriksson H, Maegiste K, Plantin L-O, Fonnum F, Hedstroem K-G, Theodorsson-Norheim E, Kristensson K, Staalberg E, Heilbronn E (1987) Effects of manganese oxide on monkeys as revealed by a combined neurochemical, histological and neurophysiological evaluation. *Archives of Toxicology* 61:46-52.
- Ernstgard L, Shibata E, Johanson G (2005) Uptake and disposition of inhaled methanol vapor in humans. *Toxicological Sciences* 88:30-38.
- EU (1998) On the quality of water intended for human consumption. European Union directive 98/83/EC (EU 98/83/EC).
- Evans JT, Green JD, Carlin PE, Barrett LO (1995) Meta-analysis of antibiotics in tube thoracostomy. *American Surgeon* 61:215-219.
- Fawell J, Robinson D, Bull RJ, Birnbaum L, Boorman GA, Butterworth B, Daniel P, Galal-Gorchev H, Hauchman F, Julkunen P, Klaassen C, Krasner S, Orme-Zavalet J, Reif J, Tardiff RG (1997) Disinfection by-products in drinking water: critical issues in health effects research. *Environmental Health Perspectives* 105:108-109.

- Feihl F, Waeber B, Liaudet L (2001) Is nitric oxide overproduction the target of choice for the management of septic shock? *Pharmacology and Therapeutics* 91:179-213.
- Fern-Pollak L, Whone AL, Brooks DJ, Mehta MA (2004) Cognitive and motor effects of dopaminergic medication withdrawal in Parkinson's disease. *Neuropsychologia* 42:1917-1926
- Fleischauer AT, Arab L (2001) Recent advances on the nutritional effects associated with the use of garlic as a supplement. *Journal of Nutrition* 131:1032S-1040S.
- Fleiss JL, Gross AJ (1991) Meta-analysis in epidemiology, with special reference to studies of the association between exposure to environmental tobacco smoke and lung cancer: a critique. *Journal of Clinical Epidemiology* 44:127-139.
- Flinn RH, Neal PA, Fulton WB (1941) Industrial manganese poisoning. *Journal of Industrial Hygiene and Toxicology* 23:374-387.
- Foster PM, Auton TR (1995) Application of benchmark dose risk assessment methodology to developmental toxicity: an industrial view. *Toxicology Letters* 82-83:555-559.
- Fox NI, Wikle CK (2005) A Bayesian quantitative precipitation nowcast scheme. *Weather and Forecasting* 20:264-275.
- Frank E (1994) Authors' criteria for selecting journals. *Journal of the American Medical Association* 272:163-164.
- Gallagher MD, Nuckols JR, Stallones L, Savitz DA (1998) Exposure to trihalomethanes and adverse pregnancy outcomes. *Epidemiology* 9:484-489.
- Gallizo JL, Jimenez F, Salvador M (2002) Adjusting financial ratios: a Bayesian analysis of the Spanish manufacturing sector. *Omega* 30:185-195.
- GAO (1992) Cross design synthesis: a new strategy for medical effectiveness research. General Accounting Office, Washington, DC.

- Gavin NI, Hasselblad V, Eddy DM (1987) The role of adjuvant tamoxifen in the treatment of postmenopausal women with operable, node-positive breast cancer. Center for Health Policy Research and Education, Duke University.
- Gaylor D, Ryan L, Krewski D, Zhu Y (1998) Procedures for calculating benchmark doses for health risk assessment. *Regulatory Toxicology and Pharmacology* 28:150-164.
- Gemperli A, Vounatsou P, Kleinschmidt I, Bagayoko M, Lengeler C, Smith T (2004) Spatial patterns of infant mortality in Mali: the effect of malaria endemicity. *American Journal of Epidemiology* 159:64-72.
- Gibbs JP, Crump KS, Houck DP, Warren PA, Mosley WS (1999) Focused medical surveillance: a search for subclinical movement disorders in a cohort of US workers exposed to low levels of manganese dust. *Neurotoxicology* 20:299-314.
- Gilks WR, Richardson S, Spiegelhalter DJ (1996) *Markov chain Monte Carlo in practice*. Chapman and Hall, London, England.
- Glatt SJ, Bolanos CA, Trksak GH, Jackson D (2000) Effects of prenatal cocaine exposure on dopamine system development: a meta-analysis. *Neurotoxicology and Teratology* 22:617-29.
- Gray LEJ, Laskey JW (1980) Multivariate analysis of the effects of manganese on the reproductive physiology and behavior of the male house mouse. *Journal of Toxicology and Environmental Health* 6:861-867.
- Greenland S, Longnecker MP (1992) Methods for trend estimation from summarized dose-response data, with applications to meta-analysis. *American Journal of Epidemiology* 135:1301-1309.
- Guth DJ, Carroll RJ, Simpson DG, Zhou H (1997) Categorical regression analysis of acute exposure to tetrachloroethylene. *Risk Analysis* 17:321-32.
- Hahn S, Williamson PR, Hutton JL (2002) Investigation of within-study selective reporting in clinical research: follow-up of applications submitted to a local research ethics committee. *Journal of Evaluation in Clinical Practice* 8:353-359.

- Harbord RM, Egger M, Sterne JAC (2006) A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. *Statistics in Medicine* (early access online).
- Hartling L, McAlister FA, Rowe BH, Ezekowitz J, Friesen C, Klassen TP (2005) Challenges in systematic reviews of therapeutic devices and procedures. *Annals of Internal Medicine* 142:1100-1111.
- Hasselblad V, Critchfield GC (1987) An analysis of neonatal screening for maple syrup urine disease. Center for Health Policy Research and Education, Duke University.
- Hasselblad V (1995) Meta-analysis of environmental health data. *The Science of the Total Environment* 160/161:545-558.
- Hasselblad V (1994) Meta-analysis in environmental statistics. *Handbook of Statistics* 12:691-716.
- Healy MJR (1972) Animal litters as experimental units. *Journal of the Royal Statistical Society Series C, Applied Statistics* 21:155-159.
- Hedges LV, Olkin I (1985) *Statistical methods for meta-analysis*. Academic Press, Inc, Orlando, Florida.
- Hedges LV, Vevea JL (2005) Selection method approaches, in *Publication bias in meta-analysis: Prevention, assessment and adjustments*, (Rothstein H, Sutton AJ, Borenstein M eds), pp 145-174. Wiley, Chichester, England.
- Hedges LV, Vevea JL (1996) Estimating effect size under publication bias: small sample properties and robustness of a random effects selection model. *Journal of Educational and Behavioral Statistics* 21:299-332.
- Hemels MEH, Vincente C, Sadri H, Masson MJ, Einarson TR (2004) Quality assessment of meta-analyses of RCTs of pharmacotherapy in major depressive disorder. *Current Medical Research and Opinion* 20:477-484.
- Higgins JPT, Green S (2005) *Cochrane Handbook for Systematic Reviews of Interventions* 4.2.5 [updated May 2005]. The Cochrane Library, Issue 3.

- Higgins JPT, Thompson SG (2002) Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* 21:1539-1558.
- Higgins JPT, Thompson SG (2004) Controlling the risk of spurious findings from meta-regression. *Statistics in Medicine* 23:1663-1682.
- Hirst WM, Murray RD, Ward WR, French NP (2002) Generalised additive models and hierarchical logistic regression of lameness in dairy cows. *Preventive Veterinary Medicine* 55:37-46.
- Hoar TJ, Milliff RF, Nychka D, Wikle CK, Berliner LM (2003) Winds from a Bayesian hierarchical model: computation for atmosphere-ocean research. *Journal of Computational and Graphical Statistics* 12:781-807.
- Horn J, de Haan RJ, Vermeulen M, Luiten PG, Limburg M (2001) Nimodipine in animal model experiments of focal cerebral ischemia: a systematic review. *Stroke* 32:2433-8.
- Hsu CH, Jeng WL, Chang RM, Chien LC, Han BC (2001) Estimation of potential lifetime cancer risks for trihalomethanes from consuming chlorinated drinking water in Taiwan. *Environmental Research* 85:77-82.
- Hua M-S, Huang C-C (1991) Chronic occupational exposure to manganese and neurobehavioural function. *Journal of Clinical and Experimental Neuropsychology* 13:495-507.
- Huang C-C, Lu C-S, Chu N-S, Hochberg F, Lilienfeld D, Olanow W, Calne DB (1993) Progression after chronic manganese exposure. *Neurology* 43:1479-1483.
- Huang Q, Liu W, Pan C (1990) The neurobehavioral changes of ferromanganese smelting workers. *Occupational Epidemiology* 329-332.
- Hurwitz EL, Aker PD, Adams AH, Meeker WC, Shekelle PG (1996) Manipulation and mobilisation of the cervical spine. A systematic review of the literature. *Spine* 21:1746-1759.

- IARC (2000) *Some industrial chemicals; IARC Monographs in the evaluation of carcinogenic risk to humans*. vol 77. International Agency for Research on Cancer, World Health Organization, Lyon, France.
- Ibrahim JG, Chen M-H (2000) Power prior distributions for regression models. *Statistical Science* 15:46-60.
- IEH, IOM (2004) Occupational exposure limits: criteria document for manganese and inorganic manganese compounds. Institute for Environment and Health; Institute of Occupational Medicine, Leicester, UK.
- Ioannidis JPA (1998) Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. *Journal of the American Medical Association* 279:281-286.
- Ioannidis JPA (2005) Differentiating bias from genuine heterogeneity: distinguishing artifactual from substantive effects, in *Publication bias in meta-analysis: Prevention, assessment and adjustments*, (Rothstein H, Sutton AJ, Borenstein M eds). Wiley, Chichester, UK.
- Iregren A (1999) Manganese neurotoxicity in industrial exposures: proof of effects, critical exposure level and sensitive tests. *Neurotoxicology* 20:315-324.
- Irwig L, Macaskill P, Berry G, Glasziou P (1998) Graphical test is itself biased. *British Medical Journal* 316:470.
- Irwig L, Tosteson ANA, Gatsonis C, Lau J, Colditz G, Chalmers TC, Mosteller F (1994) Guidelines for meta-analyses evaluating diagnostic tests. *Annals of Internal Medicine* 120:667-676.
- Iyengar S, Greenhouse J (1988) Selection models and the file drawer problem (with discussion). *Statistical Science* 3:109-135.
- Jadad AR, Cook DJ, Jones A, Klassen TP, Tugwell P, Moher M, Moher D (1998) Methodology and reports of systematic reviews and meta-analyses. *Journal of the American Medical Association* 280:278-280.

- Jiao LR, Seifalian AM, Mathie RT, Habib N, Davidson BR (2000) Portal flow augmentation for liver cirrhosis. *British Journal of Surgery* 87:984-91.
- Joffres MR, Sampalli T, Fox RA (2005) Physiologic and symptomatic responses to low-level substances in individuals with and without chemical sensitivities: a randomized controlled blinded pilot boot study. *Environmental Health Perspectives* 113:1178-1183.
- Jonderko G, Kujawsa A, Langauer-Lewowicka H (1971) Problems of chronic manganese poisoning on the basis of investigations of workers at a manganese alloy foundry. *Internationales Archiv fuer Arbeitsmedizin* 28:250-264.
- Juni P, Witschi A, Bloch R, Egger M (1999) The hazards of scoring the quality of clinical trials for meta-analysis. *Journal of the American Medical Association* 282:1054-1060.
- Kass RE, Raftery AE (1995) Bayes factors. *Journal of the American Statistical Association* 90:773-795.
- Kawamura R, Ikuta H, Fukuzumi S, Yamada R, Tsubaki S, Kodama T, Kurata S (1941) Intoxication by manganese in well water. *Kisasato Archives of Experimental Medicine* 18:145-169.
- Kelley G (1996) Mechanical overload and skeletal muscle fiber hyperplasia: a meta-analysis. *Journal of Applied Physiology* 81:1584-8.
- Kerr S, Tolliver J, Petree D (1977) Manuscript characteristics which influence acceptance for management and social science journals. *Academy of Management Journal* 20:132-141.
- Khan KS, Mignini L (2005) Surveying the literature from animal experiments: avoidance of bias is objective of systematic reviews, not meta-analysis. *British Medical Journal* 331:110-111.
- Kilburn CJ (1987) Manganese, malformations and motor disorders: findings in a manganese-exposed population. *Neurotoxicology* 8:421-430.

- Kilburn KH (1998) Neurobehavioral impairment and symptoms associated with aluminium remelting. *Archives of Environmental Health* 53:329-335.
- Kilburn KH (1999) Neurobehavioral and respiratory findings in jet engine repair workers: a comparison of exposed and unexposed volunteers. *Environmental Research* 80:244-252.
- Klimisch HJ, Andreae M, Tillmann U (1997) A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. *Regulatory Toxicology and Pharmacology* 25:1-5.
- Klotzbeucher CM, Ross PD, Landsman PB, Abbott TA, Berger M (2000) Patients with prior fractures have an increased risk of future fractures: a summary of the literature and statistical synthesis. *Journal of Bone and Mineral Research* 15:721-739.
- Koivusalo M, Vartianinen T (1997) Drinking water chlorination by-products and cancer. *Reviews on Environmental Health* 12:81-90.
- Komura J, Sakamoto M (1991) Short-term oral administration of several manganese compounds in mice: physiological and behavioral alterations caused by different forms of manganese. *Bulletin of Environmental Contamination and Toxicology* 46:921-928.
- Kondakis XG, Makris N, Leotsinidis M, Prinou M, Papapetropoulos T (1989) Possible health effects of high manganese concentration in drinking water. *Archives of Environmental Health* 44:175-178.
- Kramer MD, Lynch CF, Isacson P, Hanson JW (1992) The association of waterborne chloroform with intrauterine growth retardation. *Epidemiology* 3:407-413.
- Kroeze WK, Kristiansen K, Roth BL (2002) Molecular biology of serotonin receptors structure and function at the molecular level. *Current Topics in Medicinal Chemistry* 2:507-28.
- Kroll MW, Anderson KM, Supino CG, Adams TP (1993) Decline in defibrillation thresholds. *Pacing and Clinical Electrophysiology* 16:213-7.

- Kuhnert PM, Do K-A (2003) Fitting genetic models to twin data with binary and ordered categorical responses: a comparison of structural equation modelling and Bayesian hierarchical models. *Behavior Genetics* 33:441-454.
- Kupfersmid J, Fiala M (1991) A survey of attitudes and behaviors of authors who publish in psychology and education journals. *The American Psychologist* 46:249-250.
- 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* 24:2401-2428.
- Laskey JW, Edens FW (1985) Effects of chronic high-level manganese exposure on male behavior in the Japanese quail (*Coturnix coturnix japonica*). *Poultry Science* 64:579-584.
- Lee SC, Guo H, Lam SMJ, Lau SLA (2004) Multipathway risk assessment on disinfection by-products of drinking water in Hong Kong. *Environmental Research* 94:47-56.
- Lefevre F, Aronson N (2000) Ketogenic diet for the treatment of refractory epilepsy in children: a systematic review of efficacy. *Pediatrics* 105:E46.
- Light RJ, Pillemer DB (1984) *Summing up: the science of reviewing research*. Harvard University Press, Cambridge, MA.
- Loder E (2003) Safety of sumatriptan in pregnancy: a review of the data so far. *CNS Drugs* 17:1-7.
- Lopez-Vizcaino ME, Vidal-Rodeiro CL, Santiago-Perez MI, Vazquez-Fernandez E, Hervada-Vidal X (2002) An evaluation of spatio-temporal models for the estimation of the mortality relative risk from breast cancer. *Journal of Cancer Epidemiology and Prevention* 7:181-193.
- Lovell DP, Thomas G (1997) Quantitative Risk Assessment, in *Food Chemical Risk Analysis*, (Tennant DR ed), pp 57-86. Blackie Academic and Professional, London, UK.

- Lown BA, Morganti JB, Agostino RB, Stineman CH, Massaro EJ (1984) Effects of the postnatal development of the mouse of preconception, postconception and/or suckling exposure to manganese via maternal inhalation exposure to MnO₂ dust. *Neurotoxicology* 5:119-129.
- Lucchini R, Apostoli P, Perrone C, Placidi D, Albin E, Migliorati P, Mergler D, Sassine M-P, Palmi S (1999) Long term exposure to "low levels" of manganese oxides and neurofunctional changes in ferroalloy workers. *Neurotoxicology* 20:287-298.
- Macaskill P, Walter SD, Irwig L (2001) A comparison of methods to detect publication bias in meta-analysis. *Statistics in Medicine* 20:641-654.
- Macleod MR, Ebrahim S, Roberts I (2005) Surveying the literature from animal experiments: systematic reviews and meta-analyses are important contributions. *British Medical Journal* 331:110.
- Mapstone J, Roberts I, Evans P (2003) Fluid resuscitation strategies: a systematic review of animal trials. *Journal of Trauma-Injury Infection and Critical Care* 55:571-89.
- Martin DO, Austin H (2000) An exact method for meta-analysis of case-control and follow-up studies. *Epidemiology* 11:255-260.
- McCullagh P, Nelder JA (1996) *Generalized linear models*. vol 37. Chapman and Hall, Cambridge, England.
- McGeehin MA, Reif JS, Becher JC, Mangione EJ (1993) Case-control study of bladder-cancer and water disinfection methods in Colorado. *American Journal of Epidemiology* 138:492-501.
- McKnight B (1992) Considerations in the conduct of meta-analysis using data from animal carcinogenicity experiments. IARC Scientific Publications 116:557-69.
- Mena I, Marin O, Fuenzalida S, Cotzias GC (1967) Chronic manganese poisoning: clinical picture and manganese turnover. *Neurology* 17:128-136.

- Mendes JJA, Pluygers E (2005) Risk assessment and chemical and radiation hormesis: a short commentary and bibliographic review, in *Recent advances in quantitative methods in cancer and human health risk assessment*, (Edler L, Kitsos CP eds), pp 97-108. Wiley, Chichester, England.
- Mergler D, Huel G, Bowler R, Iregren A, Bélanger S, Baldwin M, Tardif R, Smargiassi A, Martin L (1994) Nervous system dysfunction among workers with long-term exposure to manganese. *Environmental Research* 64:151-180.
- 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* 354:1896.
- Moher D, Jones A, Lepage L (2001) Use of the CONSORT statement and quality of reports of randomized trials - a comparative before-and-after evaluation. *Journal of the American Medical Association* 285:1992-1995.
- Moolenaar RJ (1994) Default assumptions in carcinogenic risk assessment used by regulatory agencies. *Regulatory Toxicology and Pharmacology* 20:S135-41.
- Morris RD, Audet A, Angelillo IF, Chalmers TC, Mosteller F (1992) Chlorination, chlorination by-products and cancer: a meta-analysis. *American Journal of Public Health* 82:955-963.
- Mueller I, Vounatsou P, Allen BJ, Smith T (2001) Spatial patterns of child growth in Papua New Guinea and their relation to environment, diet, socio-economic status and subsistence abuse. *Annals of Human Biology* 28:263-280.
- Muller P, Parmigiani G, Schildkraut J, Tardella L (1999) A Bayesian hierarchical approach for combining case-control and prospective studies. *Biometrics* 55:858-866.
- Munafò MR, Flint J (2004) Meta-analysis of genetic association studies. *Trends in Genetics* 20:439-444.
- Murrell JA, Portier CJ, Morris RW (1998) Characterizing dose-response I: critical assessment of the benchmark dose concept. *Risk Analysis* 18:13-26.

- Murthy RC, Lal S, Saxena DK, Shukla GS, Ali MM, Chandra SV (1981) Effect of manganese and copper interaction on behavior and biogenic amines in rats fed a 10% casein diet. *Chemico-biological Interactions* 37:299-308.
- Myers JE, teWaterNaude JM, AbieZogoe HB, Fourie M, Naik I, Theodorou P, Tassell H, Daya A, Thompson M (2002) Two Phase Longitudinal or Prospective Study of the Nervous System Effects of Occupational Environmental Exposures on Mineworkers or Processing Plant Workers at Two Manganese Mines. Safety in Mines Research Advisory Committee (SIMRAC), Cape Town, South Africa.
- Nachtman JP, Tubbem RE, Commissaris RL (1986) Behavioral effects of chronic manganese administration in rats: locomotor activity studies. *Neurobehavioral Toxicology and Teratology* 8:711-715.
- Narotsky MG, Pegram RA, Kavlock RJ (1997) Effect of dosing vehicle on the developmental toxicity of bromodichloromethane and carbon tetrachloride in rats. *Fundamental and Applied Toxicology* 40:30-36.
- Nava-Ocampo AA, Reyes-Perez H, Bello-Ramirez AM, Mansilla-Olivares A, Ponce-Monter H (2000) For ischemic brain damage, is preclinical evidence of neuroprotection by presynaptic blockade of glutamate release enough? *Medical Hypotheses* 54:77-79.
- Nieuwenhuijsen MJ, Toledano MB, Eaton NE, Fawell J, Elliott P (2000) Chlorination disinfection by-products in water and their association with adverse reproductive outcomes: a review. *Occupational and Environmental Epidemiology* 57:73-85.
- Oberwalder M, Connor J, Wexner SD (2003) Meta-analysis to determine the incidence of obstetric anal sphincter damage. *British Journal of Surgery* 90:1333-1337.
- ONS (2000) Health Statistics Quarterly, 7. Office of National Statistics.
- Pappas BA, Zhang D, Davidson CM, Crowder T, Park GAS, Fortin T (1997) Perinatal manganese exposure: behavioral, neurochemical and histopathological effects in the rat. *Neurotoxicology and Teratology* 19:17-25.

- Pasquali SK, Hasselblad V, Li JS, Kong DF, Sabders SP (2002) Coronary artery pattern and outcome of arterial switch operation for transposition of the great arteries: a meta-analysis. *Circulation* 106:2575-2580.
- Peng RD, Dominici F, Pastor-Barriuso R, Zeger SL, Samet JM (2005) Seasonal analyses of air pollution and mortality in 100 US cities. *American Journal of Epidemiology* 161:585-594.
- Perna A, Remuzzi G (1996) Abnormal permeability to proteins and glomerular lesions: a meta-analysis of experimental and human studies. *American Journal of Kidney Diseases* 27:34-41.
- Peters JL, Rushton L, Sutton AJ, Jones DR, Abrams K, Mugglestone MA (2005) Bayesian methods for the cross-design synthesis of epidemiological and toxicological evidence. *Journal of the Royal Statistical Society Series C, Applied Statistics* 54:159-172.
- Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L (2006) Comparison of two methods to detect publication bias in meta-analysis. *Journal of the American Medical Association* 295:676-680.
- Peters JL, Sutton AJ, Jones DR, Rushton L, Abrams KR A systematic review of systematic reviews and meta-analyses of animal experiments with guidelines for reporting. *Journal of Environmental Science and Health Part B, Pesticides, Food Contaminants, and Agricultural Waste* (accepted).
- Peters JL, Rushton L, Jones DR, Sutton AJ, Abrams KR (2003) Interspecies extrapolation in environmental exposure standard setting: exploration of a Bayesian synthesis approach. Appendix IV of IEH Client Report for the Department of Health 'Contribution of systematic review and meta-analysis methods to the setting of environmental exposure standards'.
- Petticrew M, Gilbody S, Sheldon TA (1999) relation between hostility and coronary heart disease: evidence does not support a link. *British Medical Journal* 319:917.
- Pound P (2001) Scientific debate on animal model in research is needed. *British Medical Journal* 323:1252.

- Pound P, Ebrahim S (2002) Supportive evidence is lacking for report on animal studies. *British Medical Journal* 325:1038.
- Pound P, Ebrahim S, Sandercock P, Bracken MB, Roberts I (2004) Where is the evidence that animal research benefits humans? *British Medical Journal* 328:514-517.
- Preda A, Turetschek K, Daldrup H, Floyd E, Novikov V, Shames DM, Roberts TPL, Carter WO, Brasch RC (2005) The choice of region of interest measures in contrast-enhanced magnetic resonance image characterization of experimental breast tumors. *Investigative Radiology* 40:349-354.
- Preston C, Ashby D, Smyth R (2004) Adjusting for publication bias: modelling the selection process. *Journal of Evaluation in Clinical Practice* 10:313-322.
- 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* 19:3359-3376.
- Pries AR, Secomb TW, Sperandio M, Gaehtgens P (1998) Blood flow resistance during hemodilution: effect of plasma composition. *Cardiovascular Research* 37:225-35.
- Rang HP, Dale MM, Ritter JM (1999) *Pharmacology*. Churchill Livingstone, Edinburgh, UK.
- Ranta J, Tuominen P, Majjala R (2005) Estimation of true salmonella prevalence jointly in cattle herd and animal populations using Bayesian hierarchical modeling. *Risk Analysis* 25:23-37.
- RATSC (1999a) *Physiologically-Based Pharmacokinetic Modelling: A Potential Tool for Use in Risk Assessment*. Risk Assessment and Toxicology Steering Committee, MRC Institute for Environment and Health, Leicester, UK.
- RATSC (1999b) *Risk Assessment Procedures used by UK Government for Evaluating Human Health Effects of Chemicals*. Risk Assessment and Toxicology Steering Committee, MRC Institute for Environment and Health, Leicester, UK.

- Reif JS, Hatch MC, Bracken M, Holmes LB, Schwetz BA, Singer PC (1996) Reproductive and developmental effects of disinfection by-products in drinking water. *Environmental Health Perspectives* 104:1056-1061.
- Renwick AG (1993) Data-derived safety factors for the evaluation of food additives and environmental contaminants. *Food Additives and Contaminants* 10:275-305.
- Renwick AG, Lazarus NR (1998) Human variability and noncancer risk assessment - an analysis of the default uncertainty factor. *Regulatory Toxicology and Pharmacology* 27:3-20.
- Richardson S, Monfort C, Green M, Draper G, Muirhead C (1995) Spatial variation of natural radiation and childhood leukaemia incidence in Great Britain. *Statistics in Medicine* 14:2487-2501.
- Riley RD, Abrams K, Sutton AJ, Lambert PC, Jones DR, Heney D, Burchill SA (2003) Reporting of prognostic markers: current problems and development of guidelines for evidence-based practice in the future. *British Journal of Cancer* 88:1191-1198.
- Riley RD, Abrams KR, Lambert PC, Sutton AJ, Thompson JR (2006) An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. *Statistics in Medicine* (early access online)
- Roberts I, Kwan I, Evans P, Haig S (2002a) Does animal experimentation inform human healthcare? Observations from a systematic review of international animal experiments on fluid resuscitation. *British Medical Journal* 324:474-6.
- Roberts KA, Dixon-Woods M, Fitzpatrick R, Abrams KR, Jones DR (2002b) Factors affecting uptake of childhood immunisation: a Bayesian synthesis of qualitative and quantitative evidence. *Lancet* 360:1596-1599.
- Rodier J (1955) Manganese poisoning in Moroccan mines. *British Journal of Industrial Medicine* 12:21-35.
- Roels HA, Ghyselen P, Buchet J-P, Ceulemans E, Lauwerys RR (1992) Assessment of the permissible exposure level to manganese in workers exposed to manganese dioxide dust. *British Journal of Industrial Medicine* 49:25-34.

- Rosenthal R (1979) The file drawer problem and tolerance for null results. *Psychological Bulletin* 86:638-641.
- Rowlett JK, Woolverton WL (1996) Assessment of benzodiazepine receptor heterogeneity in vivo: apparent pA2 and pKB analyses from behavioral studies. *Psychopharmacology* 128:1-16.
- Rubery ED, Barlow SM, Steadman JH (1990) Criteria for setting quantitative estimates of acceptable intakes of chemicals in food in the UK. *Food Additives and Contaminants* 7:287-302.
- Ruddick JA, Villeneuve DC, Chu I (1983) A teratological assessment of four trihalomethanes in the rat. *Journal of Environmental Science and Health B* 18:333-349.
- Russel WMS, Burch RL (1959) *Principles of Humane Experimental Technique*. Methuen, London.
- Sackett DL, Rosenberg WMC, Muir Gray JA, Haynes RB, Richardson WS (1996) Evidence based medicine: what it is and what it isn't. *British Medical Journal* 312:71-72.
- Sakles JC, Sena MJ, Knight DA, Davis JM (1997) Effect of immediate fluid resuscitation on the rate, volume and duration of pulmonary vascular hemorrhage in a sheep model of penetrating thoracic trauma. *Annals of Emergency Medicine* 29:392-399.
- Sandercock P, Roberts I (2002) Systematic reviews of animal experiments. *Lancet* 360:586.
- Saric M, Markicevic A, Hrustic O (1977) Occupational exposure to manganese. *British Journal of Industrial Medicine* 34:114-118.
- Savitz DA, Andrews KW, Pastore LM (1995) Drinking water and pregnancy outcome in Central North Carolina: source, amount and trihalomethane levels. *Environmental Health Perspectives* 103:592-596.
- Scher RW, Dickersin K, Langenberg P (1994) Full publication of results initially presented in abstracts: a meta-analysis. *Journal of the American Medical Association* 272:158-162.

- Schuler P, Oyanguren H, Maturana V, Valenzuela A, Cruz E, Plaza V, Schmidt E, Haddad R (1957) Manganese poisoning: environmental and medical study at a Chilean mine. *Industrial Medicine and Surgery* 26:167-173.
- Schwarz G (1978) Estimating the dimension of a model. *Annals of Statistics* 6:461-464.
- Schwarzer G, Antes G, Schumacher M (2002) Inflation of type I error rate in two statistical tests for the detection of publication bias in meta-analyses with binary outcomes. *Statistics in Medicine* 21:2465-2477.
- Shea B, Moher D, Graham I, Pham B, Tugwell P (2002) A comparison of the quality of Cochrane reviews and systematic reviews published in paper-based journals. *Evaluation and the Health Professionals* 25:116-129.
- Shekelle PG, Adams AH, Chassin MR, Hurwitz EL, Brook RH (1992) Spinal manipulation for low-back pain. *Annals of Internal Medicine* 117:590-598.
- Shephard RJ, Futcher R (1997) Physical activity and cancer: how may protection be maximized? *Critical Reviews in Oncogenesis* 8:219-272.
- Shinotoh H, Snow BJ, Hewitt KA, Pate BD, Doudet D, Nugent R, Perl DP, Olanow W, Calne DB (1995) MRI and PET studies of manganese-intoxicated monkeys. *Neurology* 45:1199-1204.
- Sillanpaa MJ, Kilpikari R, Ripatti S, Onkamo P, Uimari P (2001) Bayesian association mapping for quantitative traits in a mixture of two populations. *Genetic Epidemiology* 21:S692-S699.
- Simmonds SJ, Piegorsch WW, Nitcheva D, Zeiger E (2003) Combining environmental information via hierarchical modeling: an example using mutagenic potencies. *Environmetrics* 14:159-168.
- Singer AJ, Thode HC (2004) A review of the literature on octylcyanoacrylate tissue adhesive. *American Journal of Surgery* 187:238-248.
- Sjögren B, Gustavsson P, Hogstedt C (1990) Neuropsychiatric symptoms among welders exposed to neurotoxic metals. *British Journal of Industrial Medicine* 47:704-707.

- Slavin RE (1986) Best-evidence synthesis: an alternative to meta-analytic and traditional reviews. *Educational Researcher* 15:5-11.
- Slavin RE (1995) Best evidence synthesis: an intelligent alternative to meta-analysis. *Journal of Clinical Epidemiology* 48(1):9-18.
- Slob W, Pieters MN (1998) A probabilistic approach for deriving acceptable human intake limits and human health risks from toxicological studies: general framework. *Risk Analysis* 18:787-798.
- Smith AFM, Roberts GO (1993) Bayesian computation via the Gibbs sampler and related Markov Chain Monte Carlo methods. *Journal of the Royal Statistical Society Series B, Methodology* 55:3-23.
- Smith R (2001) Animal research: the need for a middle ground. *British Medical Journal* 322:248-249.
- Smorenburg SM, Van Noorden CJ (2001) The complex effects of heparins on cancer progression and metastasis in experimental studies. *Pharmacological Reviews* 53:93-105.
- Song F, Eastwood AJ, Gilbody S, Duley L, Sutton AJ (2000) Publication and related biases. *Health Technology Assessment* 4(10).
- Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR (2000a) Bayesian method in health technology assessment. *Health Technology Assessment* 4.
- Spiegelhalter DJ, Thomas A, Best NG (2000b) *WinBugs Version 1.3. User manual*. MRC Biostatistics Unit, Cambridge, UK.
- Spiegelhalter DJ, Best NG (2003) Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effectiveness modelling. *Statistics in Medicine* 22:3687-3709.
- Spiegelhalter DJ, Abrams KR, Myles JP (2004) *Bayesian approaches to clinical trials and health-care evaluation*. Wiley, Chichester, England.

Splus (1999) Splus 2000. MathSoft, Seattle, WA.

StataCorp (2004) Stata Statistical Software. Release 8.2. College Station, TX: Stata Corporation.

Stern JM, Simes RJ (1997) Publication bias: evidence of delayed publication in a cohort study of clinical research projects. *British Medical Journal* 315:640-645.

Sterne JAC, Egger M (2001) Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. *Journal of Clinical Epidemiology* 54:1046-1055.

Sterne JAC, Egger M, Sutton AJ (2001) Meta-analysis software, in *Systematic reviews in health care: meta-analysis in context*, vol Chapter 17, pp 336-346. BMJ, London.

Sterne JAC, Gavaghan D, Egger M (2000) Publication and related bias in meta-analysis: Power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology* 53:1119-1129.

Stevenson MA, Morris RS, Lawson AB, Wilesmith JW, Ryan JBM, Jackson R (2005) Area-level risks for BSE in British cattle before and after the July 1988 meat and bone meal feed ban. *Preventive Veterinary Medicine* 69:129-144.

Stroup DF, Berlin JA, Morton SC, Olkin I, Wialliamson GD, Rennie D, Moher D, Becker BJ, Sipe TA, Thacker SB (2000) Meta-analysis of observational studies in epidemiology: a proposal for reporting. *Journal of the American Medical Association* 283:2008-2012.

Stýblová V, Bencko V, Drobný M, Chumchal O, Rimská V, Žlab L (1979) Clinical and epidemiological study in workers exposed to manganese. *Activ Nerv* 21:290-291.

Subhash MN, Padmashree TS (1991) Effect of manganese on biogenic amine metabolism in regions of the rat brain. *Food and Chemical Toxicology* 29:579-582.

Sutton AJ (2005) Evidence concerning the consequences of publication and related biases, in *Publication Bias in Meta-analysis: Prevention, Assessment and Adjustments*, (Rothstein HR, Sutton AJ, Borenstein M eds), pp 175-192. Wiley, Chichester, England.

- Sutton AJ, Abrams KR, Jones DR (2002) Generalized synthesis of evidence and the threat of dissemination bias: the example of electronic fetal heart rate monitoring (EFM). *Journal of Clinical Epidemiology* 55:1013-1024
- Sutton AJ, Abrams KR (2001) Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research* 10:277-303.
- Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F (2000) *Methods for Meta-Analysis in Medical Research*. Wiley, Chichester.
- Suzuki Y, Mouri T, Suzuki Y, Nishiyama K, Fujii N, Yano H (1975) Study of subacute toxicity of manganese dioxide in monkeys. *The Tohoku Journal of Experimental Medicine* 22:5-10.
- Swartout JC, Price PS, Dourson ML, Carlson-Lynch HL, Keenan R (1998) A probabilistic framework for the reference dose (probabilistic RfD). *Risk Analysis* 18:271-282.
- Tachibana T (1989) Behavioral teratogenic effect of methylmercury and D- amphetamine - meta-analysis and power analysis of data from the collaborative behavioral teratology study of National Center for Toxicological Research. *Teratology* 40:93-100.
- Tachibana T, Terada Y, Fukunishi K, Tanimura T (1996) Estimated magnitude of behavioral effects of phenytoin in rats and its reproducibility: a collaborative behavioral teratology study in Japan. *Physiology and Behavior* 60:941-52.
- Talih M, Hengartner N (2005) Structural learning with time-varying components: tracking the cross-section of financial time series. *Journal of the Royal Statistical Society Series B, Methodology* 67:321-341.
- Tappenden P, Chilcott JB, Eggington S, Oakley J, McCabe C (2004) Methods for expected value of information analysis in complex health economic models: developments on the health economics of interferon-beta and glatiramer acetate for multiple sclerosis. *Health Technology Assessment* 8(27).
- Terrin N, Schmid CH, Lau J, Olkin I (2003) Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine* 22:2113-2126.

- Thompson SG, Higgins JPT (2002) How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine* 21:1159-1573.
- Thompson DJ, Warner SD, Robinson VB (1974) Teratology studies on orally administered chloroform in the rat and rabbit. *Toxicology and Applied Pharmacology* 29:348-357.
- Toledano MB, Nieuwenhuijsen MJ, Best N, Whitaker H, Hambly P, de Hoogh C, Fawell J, Jarup L, Elliott P (2005) Relation of trihalomethane concentrations in public water supplied to stillbirth and birthweight in three water regions in England. *Environmental Health Perspectives* 113:225-232.
- 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* 14:545-569.
- Vieregge P, Heinzow B, Korf G, Teichert H-M, Schleifenbaum P, Mosinger H-U (1995) Long term exposure to manganese in rural well water has no neurological effects. *Canadian Journal of Neurological Sciences* 22:286-289.
- Villanueva CM, Cantor KP, Cordier S, Jaakkola JJK, King WD, Lynch CF, Porru S, Kogevinas M (2004) Disinfection byproducts and bladder cancer - a pooled analysis. *Epidemiology* 15:357-367.
- Villanueva CM, Fernandez F, Malats N, Grimalt JO, Kogevinas M (2003) Meta-analysis of studies on individual consumption of chlorinated drinking water and bladder cancer. *Journal of Epidemiology and Community Health* 57:166-173.
- Villar J, Mackey ME, Carroli G, Donner A (2001) Meta-analyses in systematic reviews of randomized controlled trials in perinatal medicine: comparison of fixed and random effects models. *Statistics in Medicine* 20:3635-3647.
- Waldmann P, Garcia-Gil MR, Sillanpaa MJ (2005) Comparing Bayesian estimates of genetic differentiation of molecular markers and quantitative traits: an application to *Pinus sylvestris*. *Heredity* 94:623-629.

- Wang J-D, Huang C-C, Hwang Y-H, Chiang J-R, Lin J-M, Chen J-S (1989) Manganese induced parkinsonism: an outbreak due to unrepaired ventilation control system in a ferromanganese smelter. *British Journal of Industrial Medicine* 46:856-859.
- Watanabe KH (2005) Modeling exposure and target organ concentrations, in *Recent advances in quantitative methods in cancer and human health risk assessment*, (Edler L, Kitsos CP eds), pp 115-124. Wiley, Chichester, England.
- Weber EJ, Callahan ML, Wears RL, Barton C, Young G (1998) Unpublished research from a medical speciality meeting. Why investigators fail to publish. *Journal of the American Medical Association* 280:257-259.
- Wegener K (1979) Systematic review of thorotrast data and facts: animal experiments. *Virchows Archiv. A, Pathological Anatomy and Histology* 381:245-68.
- Wennberg A, Iregren A, Struwe G, Cizinsky G, Hagman M, Johansson L (1991) Manganese exposure in steel smelters: a health hazard to the nervous system. *Scandinavian Journal of Work, Environment and Health* 17:255-262.
- Whitaker H, Nieuwenhuijsen MJ, Best N, Fawell J, Gowers A, Elliott P (2003) Description of trihalomethane levels in three UK water supplies. *Journal of Exposure Analysis and Environmental Epidemiology* 13:17-23.
- Whitehead A, Bailey AJ, Elbourne D (1999) Combining summaries of binary outcomes with those of continuous outcomes in a meta-analysis. *Journal of Biopharmaceutical Statistics* 9:1-16.
- Whitehead A (2002) *Meta-analysis of controlled clinical trials*. Wiley, Chichester, England.
- WHO (1981) *Manganese*. World Health Organization, Geneva, Switzerland.
- WHO (1994) *Assessing Human Health Risks of Chemicals: Derivation of Guidance Values for Health-Based Exposure Limits*. World Health Organization, Geneva, Switzerland.
- WHO (1996) *Guidelines for drinking water quality: health criteria and other supporting information*. World Health Organization, Geneva, Switzerland.

- WHO (2004) Low birthweight: country, regional and global estimates. WHO and UNICEF, New York.
- Williams DA (1975) The analysis of binary responses from toxicological experiments using reproduction and teratogenicity. *Biometrics* 31:949-952.
- Williamson PR, Gamble C (2005) Identification and impact of outcome selection bias in meta-analysis. *Statistics in Medicine* 24:1547-1561.
- Williamson PR, Gamble C, Altman DG, Hutton JL (2005) Outcome selection bias in meta-analysis. *Statistical Methods in Medical Research* 14:515-524.
- Wolpert RL, Warren-Hicks WJ (1992) Bayesian hierarchical logistic models for combining field and laboratory survival data. *Bayesian Statistics* 4:525-546.
- Wolpert RL, Mengersen KL (2004) Adjusted likelihood for synthesizing empirical evidence from studies that differ in quality and design: effects of environmental tobacco smoke. *Statistical Science* 19:450-471.
- Woodruff LD, Bounkeo JM, Brannon WM, Dawes KS, Barham CD, Waddell DL, Enwemeka CS (2004) The efficacy of laser therapy in wound repair: a meta-analysis of the literature. *Photomedicine and Laser Surgery* 22:241-247.
- Woolley A (2003) *A guide to practical toxicology: evaluation, prediction and risk*. Taylor and Francis, London.
- Wright JM, Schwartz J, Dockery DW (2004) The effect of disinfection by-products and mutagenic activity on birth weight and gestational duration. *Environmental Health Perspectives* 112:920-925.
- Zhong J, Dujovny M, Park HK, Perez E, Perlin AR, Diaz FG (2003) Advances in ICP monitoring techniques. *Neurological Research* 25:339-50.
- Zock PL, Katan MB (1998) Linoleic acid intake and cancer risk: a review and meta-analysis. *American Journal of Clinical Nutrition* 68:142-53.

Appendix A: Summary of the derivation of occupational exposure limits for manganese (Mn)

WHO Study Group (1980) Occupational Exposure Limit
0.3 mg/m³ respirable dust (TWA)

Based on a review of animal and human evidence, the WHO Study Group suggest that the most sensitive and important health outcomes from exposure to manganese are those affecting the central nervous system and lungs. A narrative review of the evidence suggests central nervous system effects occur at levels below 5 mg/m³ and adverse effects in the lungs are observed at levels of 0.3–0.5 mg/m³. From this review, the study group suggest 0.3 mg/m³ respirable dust as an occupational exposure limit.

DFG (1999) Maximum Workplace Concentration
0.5 mg/m³ total dust

From a review of animal and human evidence, the authors state that the most important health outcome is effects on the central nervous system. Using data from four human occupational epidemiology studies assessing central nervous system effects (Iregren, 1990; Roels *et al.*, 1992; Wennberg *et al.*, 1992; Mergler *et al.*, 1994), the range of exposure levels where effects are observed is reported to be 0.25-1 mg/m³. Data from animal experiments are also reviewed to assess whether other health outcomes are of significance, but conclude that the central nervous system is the main outcome of interest. Based on the fact that there are differences in the sampling equipment used in Germany and in Sweden (where the study with the lowest observed effect - 0.25 mg/m³ - was carried out), and that concentrations measured in Germany are twice those measured in the US and Sweden, the

maximum workplace concentration is reported to be 0.5 mg/m^3 (i.e. twice that observed in the Swedish study).

ACGIH (2002) Threshold Limit Value for Occupational Exposure
0.03 mg/m³ respirable dust (TWA)

Using evidence from a number of human studies, the ACGIH declare that effects on the central nervous system and the lungs are of primary importance in the derivation of an occupational exposure limit. Taking LOAELs from three occupational epidemiology studies (Roels *et al.*, 1992: LOAEL 0.15 mg/m^3 ; Mergler *et al.*, 1992: LOAEL 0.035 mg/m^3 ; Lucchini *et al.*, 1999: LOAEL 0.037 mg/m^3), a threshold limit value for occupational exposure of 0.03 is obtained. Few details are available on exactly how this level is obtained from the three LOAELs, but the authors state that no UFs were applied because of the agreement between LOAELs from Mergler *et al.* and Lucchini *et al.* Results from six additional human studies and other possibly important health outcomes are discussed in light of the recommended exposure limit.

Clewell *et al.* (2003) Occupational Exposure Guideline
0.1-0.3 mg/m³ respirable dust (TWA)

Based on a review of the human evidence, Clewell *et al.* argue that neurological symptoms are the most sensitive to manganese exposure. From a review of these human studies, four were identified as being relevant for calculation of the benchmark dose (Roels *et al.*, 1992; Gibbs *et al.*, 1999; Iregren, 1990; Mergler *et al.*, 1994). However data from only two of these studies was used since individual patient data were available (Roels *et al.*, 1992; Gibbs *et al.*, 1999). Results from a number of tests reported in each study were analysed using the benchmark dose approach: Three dose-response models were defined; a Weibull model for quantal data (Roels *et al.*, 1992; Gibbs *et al.*, 1999) and a Weibull model for continuous

data (Gibbs *et al.*, 1999) and a K-power model for continuous data (Gibbs *et al.*, 1999).

In total nineteen benchmark dose levels are calculated and presented in Clewell *et al.* (2003), ranging from 0.09 to 0.27 mg/m³, with no UFs being applied at any point. The authors do not report the exposure limit the smallest benchmark dose level for a number of reasons, including the fact the smaller levels resulted from continuous data being redefined as quantal data and that the health outcomes used in each study do not coincide with what is considered to be ‘material impairment’ by the US Occupational Safety and Health Administration, therefore taking the lowest benchmark dose is likely to result is a conservative exposure limit. Thus, the following range of limits is recommended 0.1-0.3 mg/m³ respirable dust.

IEM and IOM (2004)

Occupational Exposure Limit

0.1 mg/m³ respirable dust

Initial large review of human and animal studies leads authors to the following assumptions on which the Occupational Exposure Limit is based:

- Respirable route is most relevant (i.e. the dust likely to enter the respiratory tract) since the evidence of there is insignificant gastrointestinal absorption
- Neurological endpoints are the first indications of an adverse effect
- Even with prolonged exposure to Mn there may be a threshold effect, although this is likely to vary between individuals

The authors conclude that there are sufficient relevant human data for the basis of the derivation of the occupational exposure limit and cite, and briefly describe, the three human epidemiology studies used (Roels *et al.*, 1992; Gibbs *et al.*, 1999; Myers *et al.*, 2002).

Few details are given on how the limit of 0.1 mg/m³ respirable dust is obtained from these studies: “Based on these three studies, it is concluded that limiting exposure to 0.1 mg/m³ respirable manganese will prevent most workers from developing the

subtlest detectable effect, that is, a small non-clinical decrement in motor neurobehavioural function". The authors only state that no UFs were applied, since the population in each of the three studies is representative of the intended population subject to the occupational exposure limit. A supplementary limit is reported just in case the gastrointestinal route of exposure is important (0.5 mg/m^3 total dust).

Dietary intake, Clewell's calculation of a benchmark dose (Clewell *et al.*, 2003) and data from one animal experiment (St-Pierre *et al.*, 2001) are all considered as supporting evidence for an occupational exposure limit of 0.1 mg/m^3 respirable dust.

Appendix B: Search strategies for systematic reviews and meta-analyses of animal experiments

Table B.1 Searching the electronic databases

Database		Search strategy
Medline (1966-July 2005)	1	review.ab
(adaptation of CRD	2	review.pt
Strategy 2.1 ^a)	3	meta-analysis.ab
	4	meta-analysis.pt
	5	meta-analysis.ti
	6	or/1-5
	7	letter.pt
	8	comment.pt
	9	editorial.pt
	10	or/7-9
	11	animals.sh
	12	experiment.tw
	13	11 or 12
	14	6 not 10
	15	13 and 14
	16	systematic.tw
	17	15 and 16
Total number of articles		1464
Potentially relevant (after screening titles and abstracts)		141
<i>Actually relevant (after screening abstracts and full text articles)</i>		<i>40</i>

ScienceDirect (all years)	1	systematic review (title, abstract or keyword)
	2	animal (title, abstract or keyword)
	3	1 and 2
<hr/>		
Total number of articles		27
Potentially relevant (after screening titles and abstracts)		17
<i>Actually relevant (after screening abstracts and full text articles)</i>		<i>11</i>
<hr/>		
ScienceDirect (all years)	1	meta-analysis (title, abstract or keyword)
	2	animal (title, abstract or keyword)
	3	1 and 2
<hr/>		
Total number of articles		58
Potentially relevant (after screening titles and abstracts)		20
<i>Actually relevant (after screening abstracts and full text articles)</i>		<i>11</i>
<hr/>		
EMBASE (1980-May 2004)	1	Animal Experiment/
	2	animal stud\$.mp
	3	Animal Model/
	4	animal model\$.mp
	5	environmental exposure.mp or exp Environmental Exposure/ toxicology.mp or exp
	6	Toxicology/ or exp Genetic Toxicology/ or exp Comparative Toxicology
	7	mutagen\$.mp or exp Mutagenic

	Agent/
8	carcinogen\$.mp or exp Carcinogen/
9	to.fs
10	ae.fs
11	si.fs
12	Meta Analysis/
13	(systematic\$ and review\$).mp.
14	(systematic\$ and overview\$).mp.
15	or/1-4
16	or/5-11
17	Nonhuman/
18	review.pt.
19	13 or 14 or 18
20	systematic.mp.
21	19 and 20
22	12 or 21
23	15 and 16 and 22
24	17 and 23
<hr/>	
Total number of articles	295
Potentially relevant (after screening titles and abstracts)	63
<i>Actually relevant (after screening abstracts and full text articles)</i>	<i>17</i>

Medline (1966-May 2004)	1	meta-analysis.sh
	2	meta-analysis.pt
	3	meta analy\$.tw
	4	met analy\$.tw
	5	metanaly\$.tw
	6	metaanaly\$.tw
	7	animal.sh
	8	or/1-6
	9	7 & 8
Total number of articles		812
Potentially relevant (after screening titles and abstracts)		151
<i>Actually relevant (after screening abstracts and full text articles)</i>		47
TOXNET (all years)	1	meta analysis
	2	animal
	3	1 & 2
Total number of articles		191
Potentially relevant (after screening titles and abstracts)		14
<i>Actually relevant (after screening abstracts and full text articles)</i>		2

^a available at <http://www.york.ac.uk/inst/crd/search.htm>

ab, abstract; pt, publication type; ti, title; sh, subject heading; tw, text word; mp, free text; exp, term exploded; fs, floating subheading; to, toxicity; ae, adverse drug reaction; si, side effects

Table B.2 Searching the grey literature

Database	Keywords	Hits	Potentially relevant	Actually relevant
Cancerlit http://www.cancer.gov/search/	(meta analys OR systematic review) AND (animal or experiment)	58	0	
UK DoH http://www.info.doh.gov.uk/doh/poin t.nsf/Publications?ReadForm	meta-analysis systematic review & animal systematic review & experiment	2 1 0	0 0 0	
British Library http://catalogue.bl.uk	meta analys* systematic review & animal systematic review & experiment	121 0 0	0	
Graylit http://graylit.osti.gov/	meta analys* systematic review & animal systematic review & experiment	256+ 250+ 250+	1 0 0	0
Agricola http://www.nal.usda.gov/ag98/ag98.h tml	meta-analysis systematic & review meta-analyses	7 21 0	1 0	0

‘*’ indicates truncated search (e.g. analys* will identify analysis, analyses and analyse); ‘+’ indicates that the maximum number of articles were shown, but there may have been more relevant articles.

Appendix C: List of references reviewed in Chapter 4

57 systematic reviews of animal experiments

- Aghajafari F, Murphy K, Matthews S, Ohlsson A, Amankwah K, Hannah M (2002) Repeated doses of antenatal corticosteroids in animals: a systematic review. *American Journal of Obstetrics and Gynecology* 186:843-849.
- Albertazzi P (2002) Purified phytoestrogens in postmenopausal bone health: is there a role for genistein? *Climacteric* 5:190-196.
- Albertazzi P, Coupland K (2002) Polyunsaturated fatty acids. Is there a role in postmenopausal osteoporosis prevention? *Maturitas* 42:13-22.
- Apostoli P, Lucchini R, Alessio L (2000) Are current biomarkers suitable for the assessment of manganese exposure in individual workers? *American Journal of Industrial Medicine* 37:283-290.
- Astorg P (2004) Dietary n-6 and n-3 polyunsaturated fatty acids and prostate cancer risk: a review of epidemiological and experimental evidence. *Cancer Causes and Control* 15:367-386.
- Bailey B (2003) Glucagon in beta-blocker and calcium channel blocker overdoses: a systematic review. *Journal of Toxicology - Clinical Toxicology* 41:595-602.
- Bartels C, Gerdes A, Babin-Ebell J, Beyersdorf F, Boeken U, Doenst T, Feindt P, Heiermann M, Schlensak C, Sievers HH (2002) Cardiopulmonary bypass: evidence or experience based? *Journal of Thoracic and Cardiovascular Surgery* 124:20-27.
- Bidani A, Tzouanakis AE, Cardenas VJJ, Zwischenberger JB (1994) Permissive hypercapnia in acute respiratory failure. *Journal of the American Medical Association* 272:957-962.
- Blackman K, Brown SG, Wilkes GJ (2001) Plasma alkalization for tricyclic antidepressant toxicity: a systematic review. *Emergency Medicine Australasia* 13:204-210.
- Borrelli F, Ernst E (2002) Cimicifuga racemosa: a systematic review of its clinical efficacy. *European Journal of Clinical Pharmacology* 58:235-241.
- Borrelli F, Izzo AA, Ernst E (2003) Pharmacological effects of Cimicifuga racemosa. *Life Sciences* 73:1215-1229.
- Breil M, Chariot P (1999) Muscle disorders associated with cyclosporine treatment. *Muscle and Nerve* 22:1631-1636.

- Cairney S, Maruff P, Clough AR (2002) The neurobehavioural effects of kava. *Australian and New Zealand Journal of Psychiatry* 36:657-662.
- Chasan-Taber L, Stampfer MJ (1998) Epidemiology of oral contraceptives and cardiovascular disease. *Annals of Internal Medicine* 128:467-477.
- Chrubasik S, Pittler MH, Roufogalis BD (2005) Zingiberis rhizoma: a comprehensive review on the ginger effect and efficacy profiles. *Phytomedicine* 12(9):684-701
- Clarkson BH, Rafter ME (2001) Emerging methods used in the prevention and repair of carious tissues. *Journal of Dental Education* 65:1114-1120.
- Collins JJ, Ness R, Tyl RW, Krivanek N, Esmen NA, Hall TA (2001) A review of adverse pregnancy outcomes and formaldehyde exposure in human and animal studies. *Regulatory Toxicology and Pharmacology* 34:17-34.
- Corpet DE, Pierre F (2003) Point: From animal models to prevention of colon cancer. Systematic review of chemoprevention in min mice and choice of the model system. *Cancer Epidemiology, Biomarkers and Prevention* 12:391-400.
- Corpet DE, Tache S (2002) Most effective colon cancer chemopreventive agents in rats: a systematic review of aberrant crypt foci and tumor data, ranked by potency. *Nutrition and Cancer* 43:1-21.
- Curran MP, Perry CM (2004) Cabergoline: a review of its use in the treatment of Parkinson's disease. *Drugs* 64:2125-2141.
- Cvetkovic RS, Scott LJ (2005) Dexrazoxane: a review of its use for cardioprotection during anthracycline chemotherapy. *Drugs* 66:1005-1024.
- Davids E, Gastpar M (2004) Buprenorphine in the treatment of opioid dependence. *European Neuropsychopharmacology* 14:209-216.
- de Lemos ML (2001) Effects of soy phytoestrogens genistein and daidzein on breast cancer growth. *Annals of Pharmacotherapy* 35:1118-1121.
- de Nijs RNJ, Jacobs JWG, Algra A, Lems WF, Bijlsma JW (2004) Prevention and treatment of glucocorticoid-induced osteoporosis with active vitamin D₃ analogues: a review with meta-analysis of randomized controlled trials including organ transplantation studies. *Osteoporosis International* 15:589-602.
- Demos LL, Kazda H, Cicuttini FM, Sinclair MI, Fairley CK (2001) Water fluoridation, osteoporosis, fractures--recent developments. *Australian Dental Journal* 46:80-87.
- Fehlings MG, Perrin RG (2005) The role and timing of early compression for cervical spinal cord injury: update with a review of recent clinical evidence. *Injury, Internal Journal of the Care of the Injured* 36:B13-26.
- Fehlings MG, Tator CH (1999) An evidence-based review of decompressive surgery in acute spinal cord injury: rationale, indications, and timing based on experimental and clinical studies. *Journal of Neurosurgery* 91:1-11.
- Fox CH, Eberl M (2002) Phytic acid (IP6), novel broad spectrum anti-neoplastic agent: a systematic review. *Complementary Therapies in Medicine* 10:229-234.

- Girard P, Stern JB, Parent F (2002) Medical literature and vena cava filters: so far so weak. *Chest* 122:963-967.
- Gutt CN, Oniu T, Schemmer P, Mehrabi A, Büchler MW (2004) Fewer adhesions induced by laparoscopic surgery? *Surgical Endoscopy* 18:898-906.
- Hill PD, Chatterton RTJ, Aldag JC (2003) Neuroendocrine responses to stressors in lactating and nonlactating mammals: a literature review. *Biological Research for Nursing* 5:79-86.
- Hirschberg E (1963) Patterns of response of animal tumors to anticancer agents. A systematic analysis of the literature in experimental cancer chemotherapy--1945-1958. *Cancer Research* 23:521-980.
- Klinge B, Gustafsson A, Berglundh T (2002) A systematic review of the effect of anti-infective therapy in the treatment of peri-implantitis. *Journal of Clinical Periodontology* 29:213-225.
- Li-Ling, Irving M (2001) The effectiveness of growth hormone, glutamine and a low-fat diet containing high-carbohydrate on the enhancement of the function of remnant intestine among patients with short bowel syndrome: A review of published trials. *Clinical Nutrition* 20:199-204.
- McCann UD, Seiden LS, Rubin LJ, Ricaurte GA (1997) Brain serotonin neurotoxicity and primary pulmonary hypertension from fenfluramine and dexfenfluramine. A systematic review of the evidence. *Journal of the American Medical Association* 278:666-672.
- Meune C, Spaulding C, Mahe I, Lebon P, Bergmann J-F (2003) Risks versus Benefits of NSAIDs Including Aspirin in Myocarditis: A Review of the Evidence from Animal Studies. *Drug Safety* 26:975-981.
- Pearl ML, Rayburn WF (2004) Choosing abdominal incision and closure techniques. *Journal of Reproductive Medicine* 49:662-670.
- Quadrilatero J, Hoffman-Goetz L (2003) Physical activity and colon cancer: A systematic review of potential mechanisms. *Journal of Sports Medicine and Physical Fitness* 43:121-138.
- Rantala H, Tarkka R, Uhari M (2001) Systematic review of the role of prostaglandins and their synthetase inhibitors with respect to febrile seizures. *Epilepsy Research* 46:251-257.
- Salvi GE, Lang NP (2004) Diagnostic parameters for monitoring peri-implant conditions. *International Journal of Oral and Maxillofacial Implants* 19:116-127.
- Sawynok J, Reid A (2003) Chronic intrathecal cannulas inhibit some and potentiate other behaviors elicited by formalin injection. *Pain* 103:7-9.
- Schwingl PJ, Guess HA (2000) Safety and effectiveness of vasectomy. *Fertility and Sterility* 73:923-936.
- Shephard RJ, Futcher R (1997) Physical activity and cancer: how may protection be maximized? *Critical Reviews in Oncogenesis* 8:219-272.

- Short DJ, El Masry WS, Jones PW (2000) High dose methylprednisolone in the management of acute spinal cord injury - a systematic review from a clinical perspective. *Spinal Cord* 38:273-286.
- Shotan A, Widerhorn J, Hurst A, Elkayam U (1994) Risks of angiotensin-converting enzyme inhibition during pregnancy: experimental and clinical evidence, potential mechanisms, and recommendations for use. *American Journal of Medicine* 96:451-456.
- Shrier I, Matheson GO, Kohl HW (1996) Achilles tendonitis: are corticosteroid injections useful or harmful? *Clinical Journal of Sport Medicine* 6:245-250.
- Singer AJ, Thode HC (2004) A review of the literature on octylcyanoacrylate tissue adhesive. *American Journal of Surgery* 187:238-248.
- Stocchi L, Nelson H (2000) Wound recurrences following laparoscopic-assisted colectomy for cancer. *Archives of Surgery* 135:948-958.
- Szeto AL, Rollwagen F, Jonas WB (2004) Rapid induction of protective tolerance to potential terrorist agents: a systematic review of low- and ultra-low dose research. *Homeopathy* 93:179-178.
- Tickner JA, Schettler T, Guidotti T, McCally M, Rossi M (2001) Health risks posed by use of Di-2-ethylhexyl phthalate (DEHP) in PVC medical devices: a critical review. *American Journal of Industrial Medicine* 39:100-111.
- Tsatsaris V, Cabrol D, Carbonne B (2004) Pharmacokinetics of tocolytic agents. *Clinical Pharmacokinetics* 43:833-844.
- Van Maele-Fabry G, Laurent C, Willems JL (2000) Dichlorvos and carcinogenicity: a systematic approach to a regulatory decision. *Regulatory Toxicology and Pharmacology* 31:13-21.
- Walfisch A, Hallak M, Mazor M (2001) Multiple courses of antenatal steroids: risks and benefits. *Obstetrics and Gynecology* 98:491-497.
- Weigl M, Tenze G, Steinlechner B, Skhirtladze K, Reining G, Bernardo M, Pedicelli E, Dworschak M (2005) A systematic review of currently available pharmacological neuroprotective agents as a sole intervention before anticipated or induced cardiac arrest. *Resuscitation* 65:21-39.
- Whysner J, Reddy MV, Ross PM, Mohan M, Lax EA (2004) Genotoxicity of benzene and its metabolites. *Mutation Research* 566:99-130.
- Wunsch MJ, Stanard V, Schnoll SH (2003) Treatment of pain in pregnancy. *Clinical Journal of Pain* 19:148-155.
- Zock PL, Katan MB (1998) Linoleic acid intake and cancer risk: a review and meta-analysis. *American Journal of Clinical Nutrition* 68:142-153.

29 Systematic reviews and meta-analyses of animal experiments

- Baron M, Haas R, Dortbudak O, Watzek G (2000) Experimentally induced peri-implantitis: a review of different treatment methods described in the literature. *International Journal of Oral and Maxillofacial Implants* 15:533-544.
- Bertani H, Gelmini R, Del Buono MG, De Maria N, Girardis M, Solfrini V, Villa E (2002) Literature overview on artificial liver support in fulminant hepatic failure: a methodological approach. *International Journal of Artificial Organs* 25:903-910.
- Biondi-Zoccai GGL, Abbate A, Parisi Q, Agostoni P, Burzotta F, Sandroni C, Zardini P, Biasucci LM (2003) Is vasopressin superior to adrenaline or placebo in the management of cardiac arrest? A meta-analysis. *Resuscitation* 59:221-224.
- Craig JC, Wheeler DM, Irwig L, Howman-Giles RB (2000) How accurate is dimercaptosuccinic acid scintigraphy for the diagnosis of acute pyelonephritis? A meta-analysis of experimental studies. *Journal of Nuclear Medicine* 41:986-993.
- Dirx MJ, Zeegers MP, Dagnelie PC, van den Bogaard T, van den Brandt PA (2003) Energy restriction and the risk of spontaneous mammary tumors in mice: a meta-analysis. *International Journal of Cancer* 106:766-770.
- Dumas P, Tremblay J, Hamet P (1994) Stress modulation by electrolytes in salt-sensitive spontaneously hypertensive rats. *American Journal of the Medical Sciences* 307:S130-137.
- Fay MP, Freedman LS, Clifford CK, Midthune DN (1997) Effect of different types and amounts of fat on the development of mammary tumors in rodents: a review. *Cancer Research* 57:3979-3988.
- Freedman LS (1994) Meta-analysis of animal experiments on dietary fat intake and mammary tumours. *Statistics in Medicine* 13:709-718.
- Glatt SJ, Bolanos CA, Trksak GH, Jackson D (2000) Effects of prenatal cocaine exposure on dopamine system development: a meta-analysis. *Neurotoxicology and Teratology* 22:617-629.
- Horn J, de Haan RJ, Vermeulen M, Luiten PG, Limburg M (2001) Nimodipine in animal model experiments of focal cerebral ischemia: a systematic review. *Stroke* 32:2433-2438.
- Jiao LR, Seifalian AM, Mathie RT, Habib N, Davidson BR (2000) Portal flow augmentation for liver cirrhosis. *British Journal of Surgery* 87:984-991.
- Kelley G (1996) Mechanical overload and skeletal muscle fiber hyperplasia: a meta-analysis. *Journal of Applied Physiology* 81:1584-1588.
- Lee DS, Nguyen QT, Lapointe N, Austin PC, Ohlsson A, Tu JV, Stewart DJ, Rouleau JL (2003) Meta-analysis of the effects of endothelin receptor blockade on survival in experimental heart failure. *Journal of Cardiac Failure* 9:368-374.
- Linde K, Jonas WB, Melchart D, Worku F, Wagner H, Eitel F (1994) Critical review and meta-analysis of serial agitated dilutions in experimental toxicology. *Human and Experimental Toxicology* 13:481-492.

- Lucas C, Criens-Poublon LJ, Cockrell CT, de Haan RJ (2002) Wound healing in cell studies and animal model experiments by Low Level Laser Therapy; were clinical studies justified? a systematic review. *Lasers in Medical Science* 17:110-134.
- Macleod MR, O'Collins T, Horky LL, Howells DW, Donnan GA (2005a) Systematic review and meta-analysis of the efficacy of melatonin in experimental stroke. *Journal of Pineal Research* 38:35-41.
- Macleod MR, O'Collins T, Horky LL, Howless DW, Donnan GA (2005b) Systematic review and metaanalysis of the efficacy of FK506 in experimental stroke. *Journal of Cerebral Blood Flow and Metabolism* 25:713-721.
- Macleod MR, O'Collins T, Howelss DW, Donnan GA (2004) Pooling of animal experimental data reveals influence of study design and publication bias. *Stroke* 35:1203-1208.
- Mapstone J, Roberts I, Evans P (2003) Fluid resuscitation strategies: a systematic review of animal trials. *Journal of Trauma-Injury Infection and Critical Care* 55:571-589.
- Nava-Ocampo AA, Reyes-Perez H, Bello-Ramirez AM, Mansilla-Olivares A, Ponce-Monter H (2000) For ischemic brain damage, is preclinical evidence of neuroprotection by presynaptic blockade of glutamate release enough? *Medical Hypotheses* 54:77-79.
- Perna A, Remuzzi G (1996) Abnormal permeability to proteins and glomerular lesions: a meta-analysis of experimental and human studies. *American Journal of Kidney Diseases* 27:34-41.
- Pries AR, Secomb TW, Sperandio M, Gaehtgens P (1998) Blood flow resistance during hemodilution: effect of plasma composition. *Cardiovascular Research* 37:225-235.
- Roberts I, Kwan I, Evans P, Haig S (2002) Does animal experimentation inform human healthcare? Observations from a systematic review of international animal experiments on fluid resuscitation. *British Medical Journal* 324:474-476.
- Rowlett JK, Woolverton WL (1996) Assessment of benzodiazepine receptor heterogeneity in vivo: apparent pA2 and pKB analyses from behavioral studies. *Psychopharmacology* 128:1-16.
- Sumner BEH, Cruise LA, Slattery DA, Hill DR, Shahid M, Henry B (2004) Testing the validity of c-fos expression profiling to aid the therapeutic classification of psychoactive drugs. *Psychopharmacology* 171:306-321.
- Syeda B, Schukro C, Heinze G, Modaresi K, Glogar D, Maurer G, Mohl W (2004) The salvage potential of coronary sinus interventions: meta-analysis and pathophysiologic consequences. *Journal of Thoracic and Cardiovascular Surgery* 127:1709-1712.
- Willmot M, Gibson C, Gray L, Murphy S, Bath P (2005a) Nitric oxide synthase inhibitors in experimental ischemic stroke and their effects on infarct size and cerebral blood flow: a systematic review. *Free Radical Biology and Medicine* 39:412-425.
- Willmot M, Grat L, Gibson C, Murphy S, Bath PMW (2005b) A systematic review of nitric oxide donors and L-arginine in experimental stroke; effects on infarct size and cerebral blood flow. *Nitric Oxide* 12:141-149.

Woodruff LD, Bounkeo JM, Brannon WM, Dawes KS, Barham CD, Waddell DL, Enwemeka CS (2004) The efficacy of laser therapy in wound repair: a meta-analysis of the literature. *Photomedicine and Laser Surgery* 22:241-247.

17 meta-analyses of animal experiments

- Benignus VA (1994) Behavioral effects of carbon monoxide: meta analyses and extrapolations. *Journal of Applied Physiology* 76:1310-1316.
- Benignus VA, Boyes WK, Bushnell PJ (1998) A dosimetric analysis of behavioral effects of acute toluene exposure in rats and humans. *Toxicological Sciences* 43:186-195.
- Bouzom F, Laveille C, Merdjan H, Jochemsen R (2000) Use of nonlinear mixed effect modeling for the meta-analysis of preclinical pharmacokinetic data: application to S 20342 in the rat. *Journal of Pharmaceutical Sciences* 89:603-613.
- Brown KG, Strickland JA (2003) Utilizing data from multiple studies (meta-analysis) to determine effective dose-duration levels. Example: rats and mice exposed to hydrogen sulfide. *Regulatory Toxicology and Pharmacology* 37:305-317.
- Carroll RJ, Simpson DG, Zhou H (1994) Stratified ordinal regression: a tool for combining information from disparate toxicological studies. National Institute of Statistical Sciences, Research Triangle Park, US.
- Corpet DE, Tache S (2002) Ranking chemopreventive agents on rat colon carcinogenesis. *IARC Scientific Publications* 156:381-384.
- Crump KS, Krewski D, Van Landingham C (1999) Estimates of the proportions of carcinogens and anticarcinogens in bioassays conducted by the U.S. National Toxicology Program. Application of a new meta-analytic approach. *Annals of the New York Academy of Sciences* 895:232-244.
- Eichacker PQ, Parent C, Kalil A, Esposito C, Cui X, Banks SM, Gerstenberger EP, Fitz Y, Danner RL, Natanson C (2002) Risk and the efficacy of antiinflammatory agents: retrospective and confirmatory studies of sepsis. *American Journal of Respiratory and Critical Care Medicine* 166:1197-1205.
- Guth DJ, Carroll RJ, Simpson DG, Zhou H (1997) Categorical regression analysis of acute exposure to tetrachloroethylene. *Risk Analysis* 17:321-332.
- Hendry JH, Roberts SA (1990) Analysis of dose-incidence relationships for marrow failure in different species, in terms of radiosensitivity of tissue-rescuing units. *Radiation Research* 122:155-160.
- Kroll MW, Anderson KM, Supino CG, Adams TP (1993) Decline in defibrillation thresholds. *Pacing and Clinical Electrophysiology* 16:213-217.
- Marino AA (1990) Meta-analysis of multi-generational studies of mice exposed to power-frequency electric fields. *Journal of Bioelectricity* 9:213-231.
- Pakzaban P, Isacson O (1994) Neural xenotransplantation: reconstruction of neuronal circuitry across species barriers. *Neuroscience* 62:989-1001.
- Preda A, Turetschek K, Daldrup H, Floyd E, Novikov V, Shames DM, Roberts TPL, Carter WO, Brasch RC (2005) The choice of region of interest measures in contrast-enhanced magnetic resonance image characterization of experimental breast tumors. *Investigative Radiology* 40:349-354.

- Tachibana T (1989) Behavioral teratogenic effect of methylmercury and D- amphetamine - meta-analysis and power analysis of data from the collaborative behavioral teratology study of National Center for Toxicological Research. *Teratology* 40:93-100.
- Tachibana T, Terada Y, Fukunishi K, Tanimura T (1996) Estimated magnitude of behavioral effects of phenytoin in rats and its reproducibility: a collaborative behavioral teratology study in Japan. *Physiology and Behavior* 60:941-952.
- Valberg PA, Crouch EA (1999) Meta-analysis of rat lung tumors from lifetime inhalation of diesel exhaust. *Environmental Health Perspectives* 107:693-699.

Appendix D: Summaries of the meta-analyses of animal experiments

Table D.1 Details of the 46 meta-analyses of animal experiments

Study	Setting	Origin of data	Species	Number studies	Study quality	Methodology
<i>Unknown</i>						
Bertani <i>et al.</i> (2002)	Use of artificial liver support in fulminant hepatic liver failure	Systematic review	No details of species or strain	12 animal experiments in meta-analysis	Not reported	<p>Effect estimates: Odds ratio for survival</p> <p>Heterogeneity: Assessed but none suspected (p-value of 0.113), give no details on how assessed heterogeneity</p> <p>Synthesis method: Not given, but reference Cook <i>et al</i> 1995 <i>Journal of Clinical Epidemiology</i></p> <p>Subgroup analyses: No</p> <p>Assess publication bias: Do not mention</p>
Italy						
<i>International Journal of Artificial Organs</i>						
<i>Combining p-values</i>						
Jiao <i>et al.</i> (2000)	Augmented portal perfusion for the treatment of portal hypertension in cirrhosis	Systematic review	Rat, pig, rabbit and human	5	Not reported	<p>Effect estimates: The mean intrahepatic portal vascular resistance (IHPR) was calculated. The mean IHPR was compared before and after portal pumping using t-tests</p> <p>Heterogeneity: Mention but do not assess</p> <p>Synthesis method: P-values weighted by sample size and square rooted</p> <p>Subgroup analyses: No</p> <p>Assess publication bias: Do not mention, although point out that all but one study are from the same group of investigators</p>
UK						
<i>British Journal of Surgery</i>						

Study	Setting	Origin of data	Species	Number studies	Study quality	Methodology
<i>Some average</i>						
Corpet and Taché (2002)	To find the most potent chemopreventive agents for colorectal cancer	Not systematic	Rodents	122 (= 171 agents)	Not reported	<i>Effect estimates:</i> Potency “estimated by the ratio of value in control rats divided by the value in treated rats” <i>Heterogeneity:</i> No details <i>Synthesis method:</i> Obtain median potency across all agents <i>Subgroup analyses:</i> No <i>Assess publication bias:</i> Do not mention
France						
<i>IARC Scientific Publications</i>						
Dumas <i>et al.</i> (1994)	Effect of nutritional calcium on blood pressure	Systematic review	Rats	Not explicit	Not reported	<i>Effect estimates:</i> Average difference in blood pressure between experimental and control groups <i>Heterogeneity:</i> No details <i>Synthesis method:</i> “A pooled mean effect was calculated when the data were numerous and similar enough” <i>Subgroup analyses:</i> By type of rat <i>Assess publication bias:</i> Do not mention
Canada						
<i>American Journal of the Medical Sciences</i>						
Kelley (1996)	Effects of mechanical overload on skeletal muscle fibre number	Systematic review	Quail, chickens, rats, cats and mice	17	Not reported	<i>Effect estimates:</i> Percentage change in muscle fibre number <i>Heterogeneity:</i> No details <i>Synthesis method:</i> Non-parametric methods are reported for the meta-analyses of these data: Mann-Whitney and Kruskal-Wallis tests. <i>Subgroup analyses:</i> Yes <i>Assess publication bias:</i> Mention but no assessment was made because of the lack of validity of the statistical procedures. Kelley concludes by warning readers of the use of meta-analysis methods particularly as in this review eleven of the seventeen relevant studies are by the same research group
USA						
<i>Journal of Applied Physiology</i>						

Study	Setting	Origin of data	Species	Number studies	Study quality	Methodology
Linde <i>et al.</i> (1994) Germany <i>Human and Experimental Toxicology</i>	Protective effect of serially agitated dilutions (SADs) of toxin preparations.	Systematic review	Rats	26	Yes – a quality index was developed	<i>Effect estimates:</i> 'Protective indexes' <i>Heterogeneity:</i> No details <i>Synthesis method:</i> "Protective Indexes [were] calculated independently from the combined raw data". No further details on the methodology are reported <i>Subgroup analyses:</i> No <i>Assess publication bias:</i> state that because of the great amount of effort in identifying the research "...and because of the number of negative studies obtained, we feel that our survey is reasonably comprehensive and unlikely to have missed important research in this area
Marino (1990) USA <i>Journal of Bioelectricity</i>	Health effects of exposure to power-frequency electric fields	No details	Mice	5 (= 8 experiments)	Not reported	<i>Effect estimates:</i> Means <i>Heterogeneity:</i> No details <i>Synthesis method:</i> Mean average – no further details <i>Subgroup analyses:</i> Yes <i>Assess publication bias:</i> Do not mention
Nava-Ocampo <i>et al.</i> (2000) Mexico <i>Medical Hypotheses</i>	Neuroprotective effects of drugs inhibiting glutamate release	Systematic review	Not reported	4 (= 30 comparisons)	Yes – as part of the inclusion criteria	<i>Effect estimates:</i> Mean parameters for brain damage <i>Heterogeneity:</i> Assessed using Chi-square test. Excluded 17 comparisons because they were heterogeneous <i>Synthesis method:</i> Weighted average – no details given on the weights <i>Subgroup analyses:</i> No <i>Assess publication bias:</i> State that there is little evidence here as more non-significant than significant studies identified from this review
Pries <i>et al.</i> (1998) Germany <i>Cardiovascular Research</i>	Effects of hemodilution on flow resistance of vascular beds	Systematic review	Dog, rabbit, cat and rat	28	Not reported	<i>Effect estimate:</i> A number of measures of effect <i>Heterogeneity:</i> State there is a great deal of variation, do not say how they assess it <i>Synthesis method:</i> Analysis of variance (using Bonferroni corrections) to investigate associations between outcome and design characteristics <i>Subgroup analyses:</i> No <i>Assess publication bias:</i> Do not mention

Study	Setting	Origin of data	Species	Number studies	Study quality	Methodology
Rowlett and Woolverton (1996)	Benzodiazepine and behavioural effects	Systematic review	Rat, baboon, monkey and pigeons	16	Not reported	<i>Effect estimate:</i> Behavioural measure of effect <i>Heterogeneity:</i> Assess interactions by species/ treatment <i>Synthesis method:</i> ANOVA with Bonferonni corrections and Tukey-Kramer post hoc test <i>Subgroup analyses:</i> No <i>Assess publication bias:</i> Do not mention
USA						
<i>Psychopharmacology</i>						
Sumner <i>et al.</i> (2004)	Effect in brain of a number of psychoactive drugs	Systematic review	No details	No details	Not reported	<i>Effect estimates:</i> Percentage change <i>Heterogeneity:</i> Mention, but do not assess <i>Synthesis method:</i> Mean average percentage change (no details on any weighting) <i>Subgroup analyses:</i> No details <i>Assess publication bias:</i> Do not mention
UK						
<i>Psychopharmacology</i>						
Tachibana (1989)	Possible behavioural effects of exposure to methylmercury and d-amphetamine	Experiments across six laboratories. Collaborative Behavioural Teratology Study	Not specified	24	Not reported	<i>Effect estimates:</i> Standardised mean differences <i>Heterogeneity:</i> Assess (none suspected) but do not report details <i>Synthesis method:</i> A weighted average effect size. This paper references Hedges and Olkin (1980) for the methods. <i>Subgroup analyses:</i> No <i>Assess publication bias:</i> Do not mention
Japan						
<i>Teratology</i>						
Tachibana <i>et al.</i> (1996)	Adverse outcomes associated with exposure to phenytoin	Experiments from laboratories. Collaborative Behavioural Teratology Study	Rat	30	Not reported	<i>Effect estimates:</i> Standardised mean differences <i>Heterogeneity:</i> Assess (heterogeneity suspected - "if an effect size from a laboratory was judged not to share the common effect size, the effect size was excluded from the pooling") but do not report details <i>Synthesis method:</i> A weighted average – no details given on the weights <i>Subgroup analyses:</i> By animal breeder <i>Assess publication bias:</i> Do not mention
Japan						
<i>Physiology and Behavior</i>						

Study	Setting	Origin of data	Species	Number studies	Study quality	Methodology
Woodruff <i>et al.</i> (2004) USA <i>Photomedicine and Laser Surgery</i>	Efficacy of laser therapy in wound repair	Systematic review	Human and animal	24 studies (31 effect sizes)	Not reported	<i>Effect estimates:</i> Cohen's d <i>Heterogeneity:</i> No assessment <i>Synthesis method:</i> Mean d <i>Subgroup analyses:</i> Yes <i>Assess publication bias:</i> Use the Failsafe N method
<i>Pooled precision-weighted fixed effects model</i>						
Glatt <i>et al.</i> (2000) USA <i>Neurotoxicology and Teratology</i>	Prenatal cocaine exposure on dopamine system development	Systematic review	Rats and rabbits	16 studies = 87 experiments	Not reported	<i>Effect estimates:</i> Pearson's product-moment correlation coefficients <i>Heterogeneity:</i> Chi-square test (heterogeneity not suspected) <i>Synthesis method:</i> The median, unweighted, weighted (by total number of subjects in each experiment) are calculated and reported <i>Subgroup analyses:</i> No details <i>Assess publication bias:</i> Do not mention
Pakzaban and Isacson (1994) USA <i>Neuroscience</i>	Transplantation of fetal neuroblasts from and transplanted into the adult brain	Published data (but give no information on how studies searched and identified)	Rats (as hosts)	25	Yes	<i>Effect estimates:</i> Survival <i>Heterogeneity:</i> Mention, but no assessment <i>Synthesis method:</i> Combined using the Mantel-Haenzel test <i>Subgroup analyses:</i> No details <i>Assess publication bias:</i> Do not mention

Study	Setting	Origin of data	Species	Number studies	Study quality	Methodology
Perna and Remuzzi (1996) Italy <i>American Journal of Kidney Disease</i>	Urinary protein excretion and glomerular injury in proteinuric nephropathies	Systematic review	Rats and humans	41 for proteinuria; 16 for albuminuria (+ 17 human)	Not reported	<i>Effect estimates:</i> Correlation coefficient between two variables under study. <i>Heterogeneity:</i> Mention, but do not assess <i>Synthesis method:</i> Effect sizes weighted by sample size <i>Subgroup analyses:</i> No details <i>Assess publication bias:</i> No, but mention possibility of reporting bias within studies
Roberts <i>et al.</i> (2002) UK <i>BMJ</i>	Fluid resuscitation	Systematic review	Rats, pigs and sheep	38	Yes	<i>Effect estimates:</i> Odds ratios for mortality <i>Heterogeneity:</i> Assess (heterogeneity suspected – use stratified analyses), but give no details <i>Synthesis method:</i> Fixed effects meta-analysis <i>Subgroup analyses:</i> Yes <i>Assess publication bias:</i> Mention it, but do not assess
<i>Pooled precision-weight random effects model</i>						
Biondi-Zoccai <i>et al.</i> (2003) Italy <i>Resuscitation</i>	Is vasopressin superior to adrenaline or placebo in the management of cardiac arrest?	Systematic review	Pigs, rats and humans	33 (plus 2 human studies)	Not reported	<i>Effect estimates:</i> Risk differences <i>Heterogeneity:</i> Assess (not suspected), but no details on how <i>Synthesis method:</i> Random effects meta-analysis in Easy MA for two outcomes <i>Subgroup analyses:</i> By type of cardiac arrest <i>Assess publication bias:</i> Do not mention

Study	Setting	Origin of data	Species	Number studies	Study quality	Methodology
Horn <i>et al.</i> (2001) Netherlands <i>Stroke</i>	Use of nimodipine in acute ischemia stroke	Systematic review	Rats, rabbits, cats and gerbils	10 (7 for one outcome, 3 for another)	Yes	Effect estimates: Standardised mean differences. Heterogeneity: Assess (heterogeneity suspected so use random effects) but do not give details on how assessed Synthesis method: Random effects inverse-variance weighted average Subgroup analyses: Yes Assess publication bias: Acknowledge the possibility of publication bias and state that “ if such a publication bias for negative animal experiments exists, this would result in an even more ‘negative’ conclusion of the present review”
Macleod <i>et al.</i> 2004 Australia <i>Stroke</i>	Efficacy of nicotinamide in experimental stroke	Systematic review	Rats and mice	71 outcomes from 14 studies	Refer to Horn et al	Effect estimates: Percentage difference in neurological test for experimental compared to control animals Heterogeneity: Chi-square test. Heterogeneity suspected Synthesis method: Dersimonian and Laird random effects inverse-variance weighted mean difference Subgroup analyses: Yes, in attempt to explain heterogeneity Assess publication bias: Smaller effect estimates observed in reports given in abstracts only thus suggesting evidence of publication bias
Macleod <i>et al.</i> 2005 Australia <i>Journal of Pineal Research</i>	Efficacy of the neuroprotective drug melatonin for ischaemic stroke	Systematic review	Rats and mice	13 studies	Refer to Horn et al and Macleod et al 2004 – give quality score	Effect estimates: Proportional reduction in outcome (infarct volume, neurological score or combined score) in experimental compared to control animals Heterogeneity: Chi-squared test. Heterogeneity suspected Synthesis method: Dersimonian and Laird random effects inverse-variance weighted mean difference Subgroup analyses: Yes Assess publication bias: Do not assess, but mention it the possibility it may have an effect

Study	Setting	Origin of data	Species	Number studies	Study quality	Methodology
Macleod <i>et al.</i> 2005 Australia <i>Journal of Cerebral Blood Flow and Metabolism</i>	Efficacy of a drug (FK506) for acute stroke	Systematic review	Monkeys and rodents (not clear)	24 studies = 109 comparisons	Gave studies a score – maximum of 10	Effect estimates: Proportion improvement in outcome for experimental as a percentage of control outcome Heterogeneity: Used stratified meta-analyses and chi-square test for heterogeneity Synthesis method: Random effects Dersimonian and Laird weighted mean difference Subgroup analyses: Yes Assess publication bias: Mention that any publication bias may mean estimate is an over-estimate of effect, but not do assess it
Mapstone <i>et al.</i> (2003) UK <i>Journal of Trauma Injury, Infection and Critical Care</i>	Fluid resuscitation of uncontrolled haemorrhage	Systematic review	Rats, pigs and sheep	44	Not reported	Effect estimates: Risk ratios for mortality Heterogeneity: Assessed using chi-square test. Heterogeneity suspected - explored through stratification and meta-regression Synthesis method: Random effects meta-regression (in STATA) Subgroup analyses: Yes Assess publication bias: Do not mention
Syeda <i>et al.</i> 2004 Austria <i>Journal of Thoracic and Cardiovascular Surgery</i>	Review efficacy of methods of myocardial salvage	Systematic review	Pigs, dogs and possibly others (not clear)	Two meta-analyses: 7 studies and 5 studies	Not reported	Effect estimates: Mean infarct size Heterogeneity: Used restricted maximum likelihood. Suspected heterogeneity so used random effects model Synthesis method: Random effects weighted mean difference Subgroup analyses: No details Assess publication bias: Used Egger's test (when evidence of publication bias for meta-analysis warn about overestimating effect)

Study	Setting	Origin of data	Species	Number studies	Study quality	Methodology
Willmot <i>et al.</i> 2005 UK <i>Free Radical Biology and Medicine</i>	Efficacy of nitric oxide synthase inhibitors in experimental ischemic stroke	Systematic review	Rats (a number of strains), rabbits, cats, mice, gerbils, pigs and lambs	73 relevant	Used STAIR for methodological quality (Horn et al)	<i>Effect estimates:</i> Standardised mean difference <i>Heterogeneity:</i> Chi-square test. Heterogeneity suspected <i>Synthesis method:</i> Random effects model sample size weighted (in RevMan) <i>Subgroup analyses:</i> Yes <i>Assess publication bias:</i> Use Egger's test (suggests no pub bias, but authors say this cannot be ruled out)
Willmot <i>et al.</i> 2005 UK <i>Nitric Oxide</i>	Efficacy of nitric oxide for acute ischaemic stroke	Systematic review	Rat (a number of strains) and rabbits	25 studies	STAIR (Horn et al)	<i>Effect estimates:</i> Standardised mean differences <i>Heterogeneity:</i> Chi-square test. Heterogeneity suspected <i>Synthesis method:</i> Random effects meta-analysis (in RevMan) <i>Subgroup analyses:</i> Yes, to investigate heterogeneity <i>Assess publication bias:</i> Use Egger's test and funnel plot
<i>Pooled precision-weight fixed and random effects model</i>						
Lee <i>et al.</i> (2003) Canada <i>Journal of Cardiac Failure</i>	Effects of endothelin receptor blockade on survival in experimental heart failure	Systematic review	Rats (Sprague-Dawley and Wistar)	9	Yes – adapted from Cochrane guidelines	<i>Effect estimates:</i> Risk ratios for mortality <i>Heterogeneity:</i> Chi-square test where p<0.1 is significant <i>Synthesis method:</i> Mantel-Haenszel stratified by time of drug administration (fixed and, where evidence of heterogeneity, random effects model) <i>Subgroup analyses:</i> By time of therapy <i>Assess publication bias:</i> Mention, but do not assess
Lucas <i>et al.</i> (2002) Netherlands <i>Lasers in medical Science</i>	Review evidence for those unequivocally in favour of low level laser therapy	Systematic review	Not reported	11	Yes	<i>Effect estimates:</i> Standardised mean difference <i>Heterogeneity:</i> Assess, but do not report how <i>Synthesis method:</i> Inverse-variance weighted (cite Der Simonian and Laird 1986 and Cooper and Hedges 1994) <i>Subgroup analyses:</i> Prospectively planned 7 different subgroup analyses; carried out only 5 of them <i>Assess publication bias:</i> Mention and contact recent authors for original data – do not assessment publication bias

Study	Setting	Origin of data	Species	Number studies	Study quality	Methodology
<i>Fixed effects weighted regression model</i>						
Eichacker <i>et al.</i> (2002) USA <i>American Journal of Respiratory and Critical Care Medicine</i>	Risk of death and treatment effect to explain disparate results between the preclinical and clinical sepsis trials of anti-inflammatory agents	Animal data from citation search of relevant human articles: not from a systematic review	Not reported	38 (= 95 experiments)	Not reported	Effect estimates: Odds ratios for survival Heterogeneity: Used Breslow and Day test (significant heterogeneity identified, so looked at subgroup analyses) Synthesis method: Weighted (by the number of animals in each experiment) linear regression of the log of the odds of treatment mortality and log of the odds of control mortality in each study Subgroup analyses: Yes Assess publication bias: Do not mention
Freedman (1994) USA <i>Statistics in Medicine</i>	Fat intake and mammary tumour incidence	Systematic review	Rats (Sprague-Dawley and others)	68 (= 114 experiments)	Not reported	Effect estimates: Proportion of animals developing at least one tumour Heterogeneity: Mention, but do not assess Synthesis method: Fixed effects linear logistic model Subgroup analyses: No details Assess publication bias: Do not mention
<i>Random effects weighted regression model</i>						
Baron <i>et al.</i> (2000) Austria <i>International Journal of Oral and Maxillofacial Implants</i>	Review the experimental peri-implantitis models	Systematic Review	Monkeys, dogs and pigs	9	Not reported	Effect estimates: Pearson's correlation coefficient Heterogeneity: Weighted ANCOVA Synthesis method: Regression with random effects Subgroup analyses: No details Assess publication bias: Do not mention

Study	Setting	Origin of data	Species	Number studies	Study quality	Methodology
Dirx <i>et al.</i> (2003) Netherlands <i>International Journal of Cancer</i>	Energy restriction on spontaneous mammary tumours	Systematic review	Mice	14	Yes	Effect estimates: Risk differences Heterogeneity: Assess, but give no details on how Synthesis method: Meta-regression - inverse-variance weighted using STATA software Subgroup analyses: Yes, to explore heterogeneity Assess publication bias: Used funnel plot and Egger's test
Fay <i>et al.</i> (1997) USA <i>Cancer Research</i>	Fat intake and mammary tumour incidence	Systematic review	Rats and mice	97 (= 146 experiments)	Not reported	Effect estimates: Proportion of animals with a tumour Heterogeneity: No details Synthesis method: Used conditional logistic regression with sandwich estimators (CLRS) method Subgroup analyses: No details Assess publication bias: Do not mention
<i>Unweighted regression model</i>						
Kroll <i>et al.</i> (1993) USA <i>Pacing and Clinical Electrophysiology</i>	Reviewing evidence that defibrillation thresholds are steadily declining	No details	Humans, dogs and pigs	61 (+ 34 human)	Not reported	Effect estimates: Waveforms Heterogeneity: No details Synthesis method: Multivariate model was built by using backward and forward stepwise selection from the full list of variables Subgroup analyses: No Assess publication bias: Do not mention

Study	Setting	Origin of data	Species	Number studies	Study quality	Methodology
<i>Dose-response models</i>						
Benignus (1994)	Construction of dose-response models for the behavioural effects of exposure to carbon monoxide	No details of search strategy given, however inclusion criteria are reported.	Rats and humans	Not reported	Not reported	Effect estimates: Standard Response Metameter (SRM) - expresses the relative size of behavioural impairments due to the exposure Heterogeneity: No details Synthesis method: The model: $SRM_i = 1 - \beta(HbCO_i)^2$ and was iteratively fitted "using the method of nonlinear regression". The estimated pooled β for humans and rats are presented in the article. Subgroup analyses: No details Assess publication bias: Mention possibility of publication bias and warns that low level exposure studies should not be dismissed.
USA <i>Journal of Applied Physiology</i>						
Benignus et al. (1998)	Informing construction of physiologically based pharmacokinetic models for behavioural effects of toluene exposure	No details of literature search given, but inclusion criteria are presented	Rats and humans	6 – although not clear	Not reported	Effect estimates: Behavioural measure Heterogeneity: No details Synthesis method: Number of dose-response models were used to separately fit the human and rat data and these are given in the article. The rat estimates were extrapolated for human exposure and agreed closely with the actual human data Subgroup analyses: No details Assess publication bias: Do not mention
USA <i>Toxicological Sciences</i>						
Bouzom et al. (2000)	Preclinical data on drug S 20342	No details given on how studies identified	Rats (Wistar)	8 (5 pharmacokinetic, 3 toxicokinetic)	Not reported	Effect estimates: A number of variables Heterogeneity: No details Synthesis method: A nonlinear mixed effects model was used to fit the concentration-time data from the animals and incorporated into a two-compartment structural pharmacokinetic model using software NONMEM IV Subgroup analyses: No details Assess publication bias: Do not mention
France <i>Journal of Pharmaceutical Sciences</i>						

Study	Setting	Origin of data	Species	Number studies	Study quality	Methodology
Brown and Strickland (2003) USA <i>Regulatory Toxicology and Pharmacology</i>	Exposure to hydrogen sulphide	Not systematic - no details given of the literature search or inclusion criteria	Rats and mice	23	Not reported	Effect estimates: For each endpoint observed in the studies, the health effect is categorized by severity Heterogeneity: Assess, but no details Synthesis method: Software developed by the US Environmental Protection Agency (CatReg) is used to combine the data using categorical regression. Dose-duration response models were fitted to the individual studies and where there existed statistically homogeneous data, these were pooled using the dose-duration models. Subgroup analyses: No details Assess publication bias: Do not report
Carroll <i>et al.</i> (1994) USA <i>National Institute of Statistical Sciences</i>	Assess all health effects of acute inhalation to tetrachoroethylene	“all available data from published sources, proceedings and technical reports” are included in this database	Rats, mice and humans	Not reported	Not reported	Effect estimates: Severity scores are assigned to each outcome in all primary studies regardless of animal species used. The severity scores relate to no effect, adverse effect, severe effect Heterogeneity: Assessed through stratified analyses Synthesis method: Logistic regression. One model includes study covariates, such as species used and gender. Conclude that the pooled analysis obscures important effects which can be seen when factors such as ‘species used’ are included in the regression. Subgroup analyses: Yes Assess publication bias: Do not mention
Guth <i>et al.</i> (1997) USA <i>Risk Analysis</i>	Central nervous system health effects of acute inhalation to tetrachoroethylene	“all available studies”	Rats, mice and humans	12	Not reported	Effect estimates: Severity scores are assigned to each outcome in all primary studies regardless of animal species used. The severity scores relate to no effect, adverse effect, severe effect. Heterogeneity: Main model assumes data from homogeneous population, so use stratified regression model to allow for subgroup differences. Synthesis method: Regression model Subgroup analyses: Yes Assess publication bias: Do not mention

Study	Setting	Origin of data	Species	Number studies	Study quality	Methodology
Hendry and Roberts (1990) UK <i>Radiation Research</i>	Dose-incidence relationships for marrow failure in terms of radio sensitivity of tissue-rescuing units	No details on how studies identified	Mouse, monkey, dog, pig, sheep and goat	68	Not reported	Effect estimates: Probability of death Heterogeneity: Assessed; only combined homogeneous data Synthesis method: used the model: $-\ln(-\ln P) = D/D_0 - \ln K$; where P = probability of animal death = $\exp(-KS)$, and S = survival fraction. Subgroup analyses: No details Assess publication bias: Do not mention
Valberg and Crouch (1999) USA <i>Environmental Health Perspectives</i>	Lung tumours from lifetime inhalation of diesel-engine exhaust particles (DEPs)	No details given on the search strategy used to identify these studies, but inclusion criteria are reported	Rats (F344 and Wistar)	8 (= 13 experiments)	Not reported	Effect estimates: Incidence of tumours Heterogeneity: Mention but do not assess Synthesis method: A flexible exposure-response model containing a non-negative, but possibly zero, threshold concentration, applied to the data from each study Subgroup analyses: No details Assess publication bias: Do not mention
<i>More specialised methods of synthesis</i>						
Craig <i>et al.</i> (2000) Australia <i>Journal of Nuclear Medicine</i>	Test performance of dimercaptosuccinic acid (DMSA) scintigraphy for the diagnosis of acute pyelonephritis	Systematic review	Rats and pigs	8	Yes	Effect estimates: Sensitivity and specificity Heterogeneity: Mention but do not assess (the number of studies was too small) Synthesis method: SROC curve constructed with individual study points of equal weighting and weighting by the inverse of the variance. Changing the weighting made no difference Subgroup analyses: No details Assess publication bias: Mention and caution that the estimate of test performance given here may be an overestimate. They concede that the number of studies used here is too few to allow an assessment of publication bias

Study	Setting	Origin of data	Species	Number studies	Study quality	Methodology
Crump <i>et al.</i> (1999) USA <i>Annals of the New York Academy of Sciences</i>	Estimate the proportion of liver carcinogens	From National Toxicology Program data archives	Mice and rats	397 bioassays	Not reported	Effect estimates: Incidence of tumour Heterogeneity: No details Synthesis method: Modelling the distribution of p-values obtained from all tests Subgroup analyses: By sex and species Assess publication bias: Do not mention
<i>Correlations on individual animal data</i>						
Preda <i>et al.</i> 2005 US <i>Investigative Radiology</i>	Compare limited MR performance with MR for whole tumour area	“datasets from 3 separate studies, already published” (not further details)	Sprague-Dawley rats	3 studies = 98 tumours	Not reported	Effect estimates: Correlation coefficients Heterogeneity: No details Synthesis method: Combine all data regardless of which study the data came from Subgroup analyses: No details Assess publication bias: Do not mention

Appendix E: Peters *et al.* (2005) JRSS C paper

Bayesian methods for the cross-design synthesis of epidemiological and toxicological evidence

Jaime L. Peters, Lesley Rushton, Alex J. Sutton, David R. Jones, Keith R. Abrams
and Moira A. Mugglestone

University of Leicester, UK

[Received November 2002. Final revision December 2003]

Summary. Systematic review and synthesis (meta-analysis) methods are now increasingly used in many areas of health care research. We investigate the potential usefulness of these methods for combining human and animal data in human health risk assessment of exposure to environmental chemicals. Currently, risk assessments are often based on narrative review and expert judgment, but systematic review and formal synthesis methods offer a more transparent and rigorous approach. The method is illustrated by using the example of trihalomethane exposure and its possible association with low birth weight. A systematic literature review identified 13 relevant studies (five epidemiological and eight toxicological). Study-specific dose–response slope estimates were obtained for each of the studies and synthesized by using Bayesian meta-analysis models. Sensitivity analyses of the results obtained to the assumptions made suggest that some assumptions are critical. It is concluded that systematic review methods should be used in the synthesis of evidence for environmental standard setting, that meta-analysis will often be a valuable approach in these contexts and that sensitivity analyses are an important component of the approach whether or not formal synthesis methods (such as systematic review and meta-analysis) are used.

Keywords: Bayesian models; Cross-design synthesis; Epidemiology; Meta-analysis; Sensitivity analyses; Systematic review; Toxicology

1. Introduction

1.1. *Systematic review and synthesis methods*

As the emphasis on evidence-based medicine increases, systematic review methods are commonly used to compile, to assess the extent and quality of and to summarize the results of research. As in primary research, the target research questions, methods and results should be clearly laid out, offering transparency and making explicit assumptions and decisions that are made in execution. This allows reproducibility and ease in updating (Sutton *et al.*, 2000; Egger and Davey Smith, 2001). As part of a systematic review it may be appropriate to perform a meta-analysis, yielding a quantitative synthesis of results from several studies. Advantages of meta-analysis include greater statistical power than a single study, the potential for more precise estimates, a framework for investigation of possible sources of heterogeneity between studies and the potential to be more generalizable (Fleiss and Gross, 1991; Blettner *et al.*, 1999).

Often information regarding a particular area of interest will come from studies of fundamentally different designs, such as case–control studies and randomized controlled trials. Several researchers have investigated methods for combining results from human studies of different

Address for correspondence: Jaime L. Peters, Department of Health Sciences, University of Leicester, 22–28 Princess Road West, Leicester, LE1 6TP, UK.
E-mail: jlp9@leicester.ac.uk

designs, known as cross-design synthesis (Larose and Dey, 1997; Bhatia *et al.*, 1998; Muller *et al.*, 1999; Prevost *et al.*, 2000; Sutton and Abrams, 2001; Roberts *et al.*, 2002a). Bayesian methods of synthesis are sufficiently flexible to allow, if appropriate, for prior evidence and/or expert judgment (e.g. on the relative appropriateness of certain types of evidence) to be incorporated into the analysis of the observed data (Spiegelhalter *et al.*, 2000a).

In assessing risks to human health from exposure to chemical substances in the environment, relevant evidence comes from both animal and human research. There has been relatively little exploration of methods for quantitative synthesis of evidence from human *and* animal studies, or even of toxicological studies alone (Roberts *et al.*, 2002b; Sandercock and Roberts, 2002). However, DuMouchel and Harris (1983) and DuMouchel and Groer (1989) have investigated alternative Bayesian models for combining dose-response slopes from animal and human studies.

In this paper we explore the use of systematic review and formal methods of synthesis for combining animal and human data for setting standards on levels of human exposure to environmental chemicals. We use the example of the possible association between exposure to trihalomethanes (THMs) in drinking water and low birth weight to illustrate these approaches.

1.2. Use of human and animal data to set environmental exposure standards

Current methods to review and incorporate diverse evidence to inform risk assessments are generally not systematic and lack some degree of quantification (Risk Assessment and Toxicology Steering Committee (1999) and references therein). A single, pivotal, study may be selected as the basis for the risk assessment, e.g. the establishment of a threshold value below which adverse health effects are not observed. It is important that when environmental exposure standards are set there is transparency regarding what data are used, how they are used and the various assumptions that are made. The nature and degree of uncertainty that exists in the estimation of these standards needs to be clear to both setters and users of the standards. Currently, this is not always so; there is a real need for the development of a systematic transparent methodology to combine human and animal data that could formally incorporate biological or mechanistic data, in so far as they exist (Budtz-Jorgensen *et al.*, 2001) and allow an estimation of inherent uncertainty.

2. Example: exposure to trihalomethanes and low birth weight

2.1. Background to the example

THMs are a group of chlorinated by-products (CBPs), formed when chlorine, added to drinking water supplies for disinfection, reacts with organic and inorganic substances that are already present in the water-supply (Fawell *et al.*, 1997). The type and concentration of these CBPs depend on factors such as the amount of chlorine added, the time since the chlorine was added, the water temperature and its level of acidity (Koivusalo and Vartiainen, 1997). The most commonly measured THMs are chloroform, bromodichloromethane (BDCM), dibromochloromethane (DBCM) and bromoform, but often *total* THMs are measured and reported in epidemiological studies. Correlations between the individual and total THM concentrations (Whitaker *et al.*, 2003) suggest that total THMs are a good indicator for the concentration of chloroform in drinking water, but not for the other THMs.

A possible association between exposure to CBPs and the incidence of cancer is reported by Morris *et al.* (1992). The potential for reproductive health effects of exposure to CBPs has also been a focus of concern, in particular risks that are associated with the foetus or newborn babies. Nieuwenhuijsen *et al.* (2000) published a narrative review of human and animal research investigating the potential association between exposure to THMs and adverse reproductive effects.

They suggested that there was a potential link between exposure to THMs and low birth weight, although the quality of the human studies, particularly for assessing exposure levels, was not high and the animal data were not comprehensive. Here we build on the review of Nieuwenhuijsen *et al.* (2000) by systematically collating and quantitatively summarizing the human and animal research investigating oral exposure to THMs and a possible association with low birth weight.

2.2. A systematic search of the literature

We performed a systematic search of the literature on the potential association between oral exposure to THMs and low birth weight following standard procedures (National Health Service Centre for Reviews and Dissemination, 2001). Evidence from both human and animal studies was sought. The search was limited to reports published in English. Lists of the electronic databases and search terms that are used are given in Peters *et al.* (2003). The search was supplemented by an investigation of the references from published reports and reviews (e.g. Reif *et al.* (1996) and Fawell *et al.* (1997)) for further relevant research. 13 studies were found to be relevant (five epidemiological and eight toxicological). In the epidemiological studies the odds ratios (ORs) for low birth weight are adjusted for different covariates in each study; in the toxicological studies, different species and strains of animal are compared. There were many differences between the disciplines. Exposures are reported as parts per billion in the epidemiological studies, but in the toxicological studies as milligrams per kilogram of body weight per day; the toxicology studies report means (and standard deviations) of weight at each dose level, whereas ORs for low birth weight are reported in the epidemiological studies. These differences must be addressed directly so that comparable study-specific estimates can be obtained for each of the 13 studies.

2.3. Obtaining study-specific estimates

We transformed the exposure scale in the epidemiological studies (parts per billion) to that reported in the toxicological studies (milligrams per kilogram of body weight per day), assuming, as in standard risk assessment practice, an average body weight of 60 kg and an average water intake of 21 day⁻¹ (Department of the Environment, 1993; World Health Organization, 1996). Our aim was to obtain a dose-response slope estimate of the natural logarithm of the OR (ln(OR)) for low birth weight associated with exposure to THMs for each study. To transform the mean weights given at each dose level in the toxicological studies to ORs we further assumed that the foetal weights were normally distributed and that 7.6% of the animals in the control (zero dose) group were of low weight. (As there appear to be no generally accepted cut-off values for low weight in different species, the percentage of low birth weight babies (weighing less than 2.5 kg) in England in 1999 (Office for National Statistics, 2000) was used.) Under these assumptions we obtained the number of animals in each dose group that were considered to be of normal and low foetal weight and hence calculated the OR. For each study, we fitted a weighted least squares linear regression of ln(OR) on the natural logarithm of dose (ln(dose)), the weight being inversely proportional to the variance of the ln(OR) estimate at each ln(dose) level:

$$\ln(\text{OR}_{ij}) = \alpha_i + \beta_i \ln(\text{dose}_{ij}) + \varepsilon_i \quad (1)$$

where j is the number of observations in study i ($i = 1, \dots, 13$) and $\varepsilon_i \sim N(0, \phi^2)$. The slope estimates β_i (and corresponding variances) are the study-specific estimates to be used in the subsequent synthesis. Although there are differences in the exposures measured between the

studies (i.e. individual and total THMs), for simplicity we here assume that these exposures are equivalent throughout.

3. Methods of synthesis

3.1. Synthesis of study-specific estimates within the two disciplines

A two-level hierarchical Bayesian model (model 1) was used to combine the study-specific dose-response slope estimates β_i within each of the two disciplines (epidemiology and toxicology). Within this model framework, the data were analysed in several ways: model 1a combined only the epidemiological studies and model 1b combined only the toxicological studies. Both of these models had ‘vague’ prior distributions on the unknown parameters, so that the data would contribute most to the posterior distribution. In Section 5.3 we report sensitivity analyses of various choices of vague prior distributions.

Two further models, models 1c and 1d, also combined the epidemiological and toxicological studies respectively, but with informative prior distributions on the pooled slope parameter μ based on the posterior mean and variance from models 1a and 1b. Thus, the posterior mean and variance from the synthesis of the toxicological studies (model 1b) were used to inform the prior distribution for the synthesis of the epidemiological studies (model 1c). Similarly, the posterior mean and variance from synthesizing the epidemiological studies (model 1a) were used as a basis for the prior distribution in the synthesis of the toxicological studies (model 1d). The structure that was used by all forms of model 1 is

$$\begin{aligned} \beta_i &\sim N(\theta_i, \sigma_i^2), & \mu &\sim N(a, b), \\ \theta_i &\sim N(\mu, \tau^2), & 1/\tau^2 &\sim \text{gamma}(0.001, 0.001) \end{aligned} \tag{2}$$

where θ_i is the true dose-response slope in study i (for models 1a and 1c, $i = 1, \dots, 5$; for models 1b and 1d, $i = 1, \dots, 8$), μ is the pooled dose-response slope and τ^2 is the between-study variance. In all four models, we assume that σ_i^2 are known, and we use the observed variances of the slope estimates in each study to represent them. In models 1a and 1b, $a = 0$ and $b = 10^9$, allowing for a range of plausible values for the dose-response slope estimate of between -2×10^9 and 2×10^9 . However, in model 1c a is the posterior mean from model 1b and b is the posterior variance from model 1b, and in model 1d a is the posterior mean from model 1a and b is the posterior variance from model 1a. In all four models, a vague prior is placed on the between-study precision parameter $1/\tau^2$.

3.2. Synthesis of study-specific estimates across disciplines

A fifth form of model 1, model 1e, combines all 13 studies ($i = 1, \dots, 13$), taking no account of the studies being from two different sources, with vague prior distributions on all unknown parameters. Three further models (models 2, 3a and 3b) do take into account the fact that data from different sources are being combined.

Model 2 is a two-level Bayesian hierarchical model estimating an overall pooled slope estimate μ , but allowing distinct estimates of the between-study variances for the epidemiological studies and the toxicological studies. Model 2 is given by

$$\left. \begin{aligned} \beta_{ij} &\sim N(\psi_{ij}, \sigma_{ij}^2), & \mu &\sim N(0, 10^9), \\ \psi_{ij} &\sim N(\mu, \tau_j^2), & 1/\tau_1^2 &\sim \text{gamma}(0.001, 0.001), \\ & & 1/\tau_2^2 &\sim \text{gamma}(0.001, 0.001) \end{aligned} \right\} \tag{3}$$

where the β_{ij} are the observed slope estimates for the i th study of type j (where $j = 1$ for epidemiological studies and $j = 2$ for toxicological studies), σ_{ij}^2 are the variances of the β_{ij} and ψ_{ij} is the true slope estimate for study i of type j . τ_j^2 is the variance between studies of type j , and μ is the pooled slope estimate. Vague prior distributions were placed on the unknown parameters.

Model 3 is a three-level Bayesian hierarchical model (Prevost *et al.*, 2000; Sutton and Abrams, 2001) which includes a level to account for study type, j . In model 3a, $j = 1, 2$ ($j = 1$ for epidemiological studies and $j = 2$ for toxicological studies). In model 3b, we further divide the toxicological studies in an attempt to account for the different species and strain of animals that were used in the studies; hence $j = 1, \dots, 4$ ($j = 1$ for human studies, $j = 2$ for studies using rabbits, $j = 3$ for studies using Sprague–Dawley rats and $j = 4$ for studies using F344 rats). Model 3 is given by

$$\left. \begin{aligned} \beta_{ij} &\sim N(\psi_{ij}, \sigma_{ij}^2), & \mu &\sim N(0, 10^9), \\ \psi_{ij} &\sim N(\theta_j, \tau_j^2), & 1/\tau_j^2 &\sim \text{gamma}(0.001, 0.001), \\ \theta_j &\sim N(\mu, \nu^2), & 1/\nu^2 &\sim \text{gamma}(0.001, 0.001). \end{aligned} \right\} \quad (4)$$

As well as estimating the variance between studies of type j , τ_j^2 , model 3 estimates the variance between study types, ν^2 , and the pooled dose–response slope for the j th type of study, θ_j . As with models 1 and 2, the pooled slope estimate of the 13 studies is given by μ . Vague prior distributions were placed on the unknown parameters.

WinBUGS (version 1.3) (Spiegelhalter *et al.*, 2000b) was used to estimate parameters for the Bayesian analyses by using Markov chain Monte Carlo simulation. For each model a burn-in of 10000 iterations was followed by a further 200000 updates after which the median and the 2.5- and 97.5-percentiles of the posterior distribution were used to summarize the parameter estimates. All initial values were set to 1 and convergence and model performance were assessed visually from the trace and autocorrelation plots that are available within WinBUGS.

As part of the sensitivity analyses that are reported in Section 5.3, the length of burn-in, number of updates and the initial values given above were changed, and various vague prior distributions were placed on the parameters in models 1e, 3a and 3b to assess the sensitivity of the results to these specifications.

4. Results

4.1. Study-specific dose–response slope estimates

The dose–response slope estimates (medians), β_i and 95% credibility intervals (CIs) that were calculated from each study are shown in Fig. 1. Apart from the study of Ruddick *et al.* (1983) on exposure to DBCM, the toxicological studies all have very similar dose–response slope estimates and are quite precise; the epidemiological study estimates appear more heterogeneous and generally less precise. The estimate (posterior median) of the between-study variance τ^2 for the epidemiological studies is 0.0051, but for the toxicological studies it is 0.0027.

4.2. Within-discipline pooled dose–response slope estimates

In Table 1 the pooled dose–response slope estimates μ and 95% CIs obtained from models 1a–1d are compared with results from a classical random-effects synthesis fitted by using the META command in STATA 7 (StataCorp, 2001).

Results from the Bayesian models with vague prior distributions (models 1a and 1b) and the classical models show that the pooled epidemiological slope estimates μ are slightly larger than the pooled toxicological slope estimates and have much more variability associated with

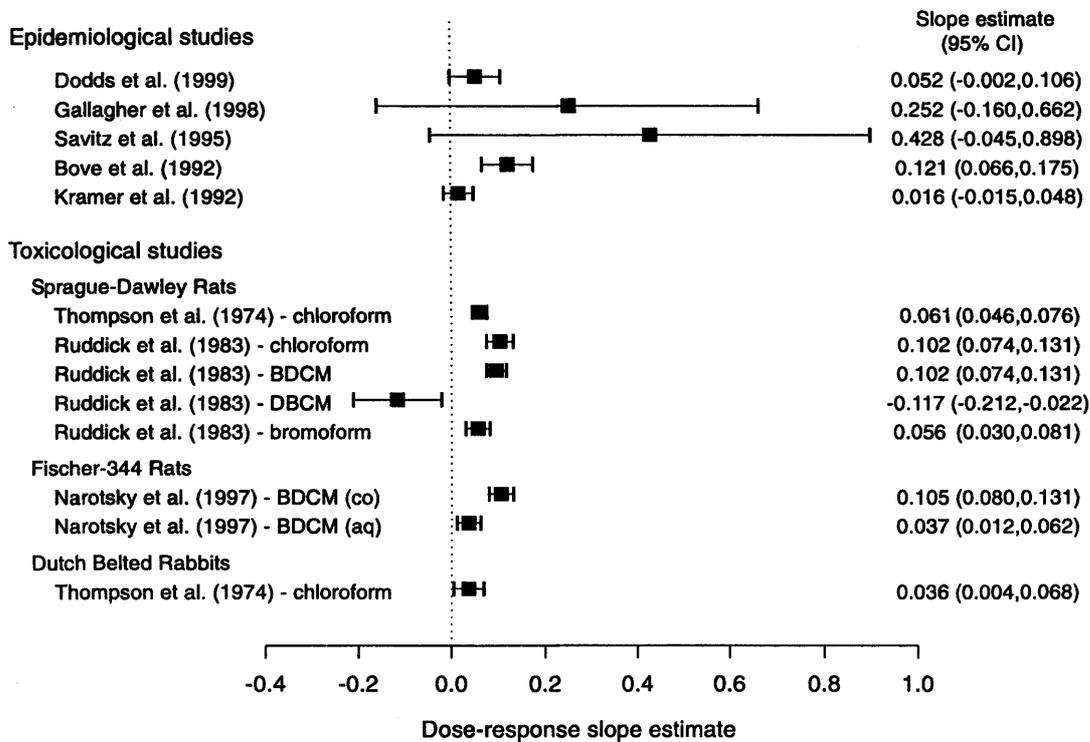


Fig. 1. Study-specific dose-response slope estimates β_i and 95% CIs from $\ln(\text{OR})$ versus $\ln(\text{dose})$ linear model (co, corn oil vehicle; aq, aqueous vehicle)

Table 1. Epidemiological and toxicological pooled slope estimates μ (and 95% CIs or confidence intervals)

Model	μ from epidemiological studies (n = 5)	μ from toxicological studies (n = 8)
Bayesian model with vague prior distribution (models 1a and 1b)	0.077 (-0.017, 0.240)	0.058 (0.006, 0.100)
Bayesian model with informative prior distribution (models 1c and 1d)	0.061 (0.023, 0.099)	0.060 (0.017, 0.100)
Classical random-effects model	0.071 (0.007, 0.135)	0.062 (0.038, 0.086)

them. This reflects what was seen in Fig. 1. As expected, the Bayesian models give pooled slope estimates with wider CIs than the estimates that are obtained from the classical random-effects model, because more variability is taken into account in the Bayesian model. The results from the Bayesian models with informative prior distributions (models 1c and 1d) are very similar to each other and are close to the pooled slope estimate of the toxicological studies from model 1b and the classical random-effects model. Thus, it would appear that in models 1c and 1d the data from the toxicological studies are dominating, regardless of whether they form the prior or the likelihood. This reflects the fact that the slope estimates from the toxicological studies are estimated more precisely than those from the epidemiological studies.

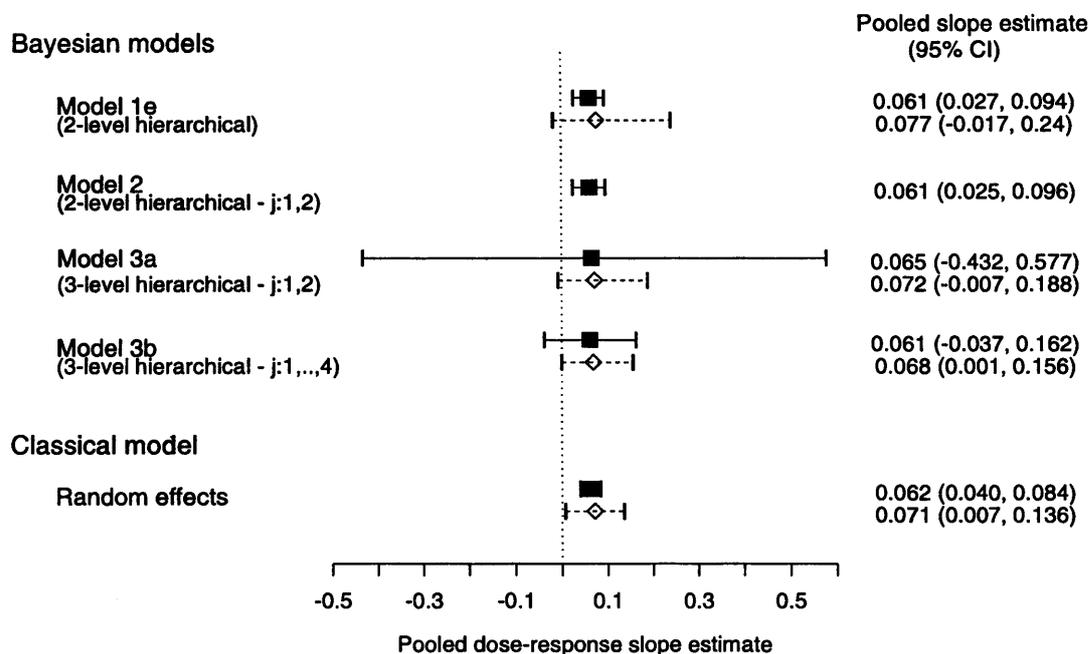


Fig. 2. Pooled dose-response slope estimates μ (and 95% CIs) obtained from the five synthesis models that were used to combine all 13 studies (model 1e, the human epidemiological estimate is μ from model 1a; model 3, the human epidemiological estimate is θ_1): ■, all-species estimate; ◇, human epidemiological estimate

4.3. Overall pooled dose-response slope estimates

The slope estimates μ that were obtained from pooling all 13 studies using the various models are shown in Fig. 2. Again, pooled estimates from the classical random-effects model are given for comparison.

From Fig. 2 we can see that the models give similar pooled dose-response slope estimates μ , but their level of precision varies. The three-level hierarchical model that only distinguishes between the epidemiological and toxicological studies (model 3a) gives a very wide CI compared with the pooled slope estimates from the other models in Fig. 2. The reason for this, as Prevost *et al.* (2000) also noted, is that only two pieces of evidence, in our example the pooled epidemiological slope estimate θ_1 and the pooled toxicological slope estimate θ_2 , are being synthesized to obtain the overall pooled slope estimate μ . Hence, when the different species and strains are taken into account in model 3b, so that there are four sources of evidence, the CI is much narrower.

5. Sensitivity analyses

As has been pointed out, many assumptions were made to obtain comparable dose-response slope estimates from each of the 13 studies and to combine using the synthesis models described. In this section some of these assumptions are assessed for their sensitivity. These include the effect of different dose-response models on the resulting slope estimates β_i , changing the water intake and body weight assumptions that were made in the initial analyses and the assessment of different vague prior distributions that were placed on the unknown parameters in the synthesis models that are defined in Section 3.

5.1. Dose-response models

We investigated the effect on the findings of using different dose-response models to obtain the study-specific slope estimates. The models were as follows:

- (a) a linear model as given in equation (1), but with dose replacing $\ln(\text{dose})$ as the explanatory variable;
- (b) a linear model ($\ln(\text{OR})$ and dose) as in (a), but taking into account the correlation structure between the $\ln(\text{OR})$ s within each study (Greenland and Longnecker, 1992);
- (c) a logit model (on $\ln(\text{dose})$) given by

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta \ln(\text{dose}) + \varepsilon \quad (5)$$

where p is the probability of observing a foetus with a low birth weight;

- (d) a logit model as in (c), but with dose replacing $\ln(\text{dose})$ as the explanatory variable.

Comparing the results of the three dose-response models where 'dose' is used as the explanatory variable (models (a), (b) and (d)), the slope estimates from model (a) are generally larger than those from models (b) and (d). The Greenland and Longnecker (1992) model (b) takes into account the correlation structure of the $\ln(\text{OR})$ estimates within each study, whereas in the logit model (d) the estimates within a study are all independent. This may suggest that the incorrect assumption of independence of the estimates within a study, as made in model (a), increases the resulting slope estimates.

The toxicological slope estimates (and standard errors) are somewhat smaller when dose is used in the model, compared with $\ln(\text{dose})$ (the dose-response slope estimates from the linear dose-response models can be found at <http://www.hs.le.ac.uk/division/epph/projects/epiandtox/>). However, for the epidemiological studies, the models that include dose give slope estimates (and standard errors) that are orders of magnitude greater than those including $\ln(\text{dose})$. The slope estimates are defined as the increase in $\ln(\text{OR})$ per unit increase in dose (milligrams per kilogram per day) and, whereas the range of dose levels that is covered by the toxicological studies reaches $400 \text{ mg kg}^{-1} \text{ day}^{-1}$, the exposures that were reported in the epidemiological studies do not exceed $0.0042 \text{ mg kg}^{-1} \text{ day}^{-1}$.

If a dose-response model is used, the epidemiological studies will each have very little weight in the synthesis of all studies (as they have very low precision), and so the toxicological studies will tend to dominate. However, if the $\ln(\text{dose})$ dose-response model is used, the toxicological studies are less likely to dominate because the precision of the slope estimates is similar for the toxicological and epidemiological studies. The results of pooling the dose-response slope estimates that were obtained from the various linear dose-response models can be found at <http://www.hs.le.ac.uk/division/epph/projects/epiandtox/>.

It is clear that the choice of dose-response model is critical for the synthesis of the studies. To investigate which dose-response model is most appropriate, we assessed the fit of each of the dose-response models by using the Bayes information criterion (Schwarz, 1978). The linear models appear to give a better fit to the data than do the logit models, since the Bayes information criterion values are lower for the linear models. However, results from applying the Bayes information criterion suggest that neither dose nor $\ln(\text{dose})$ in the linear model is more advantageous than the other in terms of model fit. The complete table of Bayes information criterion values for each dose-response model can be found at <http://www.hs.le.ac.uk/division/epph/projects/epiandtox/>.

5.2. Sensitivity of assumptions on body weight, water consumption and low birth weight cut-off values

This section reports the results from the sensitivity analyses of the assumptions that were made in Section 2.3. To convert the measurement of exposure that was reported in the epidemiological studies, we applied standard default body weight and water consumption values. Recent surveys suggest that these values are inaccurate (European Centre for Ecotoxicology and Toxicology of Chemicals, 2001; Environment Agency and Department for Environment, Food and Rural Affairs, 2002). We reanalysed the dose–response slopes β_i from the epidemiological studies using estimates from these surveys in place of the above ‘default’ values. Furthermore, by using a Bayesian model we could also assess whether inclusion of uncertainty about the assumed values affected the dose–response slope estimates that were obtained and the subsequent synthesis of the slopes. These sensitivity analyses were carried out for both the $\ln(\text{dose})$ linear dose–response model and the dose linear dose–response model. A mean of 68.53 kg (standard deviation 13.87 kg) was assumed as the average body weight (Environment Agency and Department for Environment, Food and Rural Affairs, 2002), and 0.9911 day^{-1} (standard deviation 0.0304) for average water intake (European Centre for Ecotoxicology and Toxicology of Chemicals, 2001). The full results can be found at <http://www.hs.le.ac.uk/division/epph/projects/epiandtox/>.

The different assumptions had no effect on the individual and pooled dose–response slope estimates when the $\ln(\text{dose})$ linear model is applied to the data. However, for the dose linear model, changing the assumptions increases the size of the estimates by a multiple of 2 and more. For the $\ln(\text{dose})$ model, if the variability of the assumptions from the survey data is taken into account the estimates and CIs do not change. However, taking into account the variability of the body weight and water intake assumptions in the dose model, larger dose–response slope estimates are obtained and the variability that is associated with these estimates is much greater than when the variability is not taken into account. The model that includes information on the variability of the water intake and body weight assumptions is more appealing as it shows from the outset the amount of variability that is associated with the results. Clearly, as more accurate data are collected, the models can be updated.

Further assumptions were made in defining the percentage of control group animals that were of low birth weight in the toxicological studies so that ORs could be calculated from the reported means (and standard deviations) of weight. We applied four further cut-off levels of low birth weight: 5%, 10%, 20% and taking two standard deviations below the mean for the control group (Foster and Auton, 1995). These different cut-off values had moderate effects on the ORs and slope estimates β_i that were obtained for the toxicological studies.

5.3. Checking of model assumptions and sensitivity of prior distributions

Use of different burn-in lengths, number of iterations and initial values in the Bayesian analyses in WinBUGS suggested that all the models had converged satisfactorily. Changing the vague prior distributions in model 1e has very little effect on the estimate for the pooled dose–response slopes μ and the 95% CIs (full results can be found at <http://www.hs.le.ac.uk/division/epph/projects/epiandtox/>).

The prior distributions on the unknown parameters μ , $1/\tau_j^2$ and $1/\nu^2$ in models 3a and 3b as part of these sensitivity analyses are shown in Table 2.

For brevity we do not discuss all the results here; instead we refer interested readers to Peters *et al.* (2003) for more details. However, in general, the more diffuse prior distribution on μ , $N(0, 10^{11})$, in models 3a and 3b makes no difference to the pooled slope estimates μ and 95% CIs that were obtained in the initial analyses.

Table 2. Prior distributions used in the sensitivity analyses for models 3a and 3b

Prior for pooled slope estimate μ	Prior for between-study (within type) precision, $1/\tau_j^2$	Prior for between-study-type precision, $1/\nu^2$
$N(0, 10^9)$	gamma(0.001,0.001)	gamma(0.001,0.001)
$N(0, 10^{11})$	gamma(0.1,0.1)	gamma(0.1,0.1)
	gamma(1,1)	gamma(1,1)
	$N(0, 10^6) \dagger$ for $\tau_j > 0$	$N(0, 10^6) \ddagger$ for $\nu > 0$
	$N(0, 10^4) \dagger$ for $\tau_j > 0$	$N(0, 10^4) \ddagger$ for $\nu > 0$
	$N(0, 10^2) \dagger$ for $\tau_j > 0$	$N(0, 10^2) \ddagger$ for $\nu > 0$

\dagger Prior distribution placed on the between-study standard deviation τ_j .

\ddagger Prior distribution placed on the between-study-type standard deviation ν .

The prior distributions on the precision components $1/\tau_j^2$ and $1/\nu^2$ in models 3a and 3b do have an effect on the estimated pooled dose-response slopes μ and 95% CIs. The half-normal prior distributions on $1/\tau_j^2$ give very similar estimates and 95% CIs to each other, but the gamma prior distributions give slightly different estimates and 95% CIs. Larger slope estimates and CIs are obtained when the gamma(1, 1) prior distribution is applied since larger values for the variance components can be sampled. This general pattern can be seen for models 3a and 3b (see Peters *et al.* (2003)).

These sensitivity analyses demonstrate that the choice of prior distributions that are placed on the unknown precision parameters in model 3 may have a very important influence on the results that are obtained. As Lambert *et al.* (2003) discussed in their simulation study of the effect of prior distributions in a meta-analysis scenario, it is difficult to specify truly vague priors, particularly for variance parameters. This is more problematic when there are, in effect, only two pieces of evidence from which to estimate the variance, as is the case with model 3a. However, none of the pooled slope estimates μ that are obtained from the analyses of model 3 suggested a significant increase or decrease in the risk of low birth weight with exposure to THMs, so our conclusions are robust to different choices of vague prior.

Using a Bayesian model, it is possible to include judgments on the relevance of the toxicological data to the setting of standards of safe exposure to humans. In Section 3.1 the posterior mean and variance from the synthesis of the toxicological studies is used to form the prior distribution for the synthesis of the epidemiological studies. Prevost *et al.* (2000) and Sutton and Abrams (2001) have looked at how the weight of one source of evidence may be changed in accordance with prior beliefs concerning the relevance of that evidence in the synthesis. A figure showing the pooled estimates μ corresponding to the weight that is given to the toxicological studies can be found at <http://www.hs.le.ac.uk/division/epph/projects/epiandttox/>. Using different weights for the toxicological data to inform the synthesis of the epidemiological studies does impact substantially on the results of the synthesis. Thus, expert judgment can be incorporated to assess the relevance of the toxicological data to the epidemiological data in setting environmental exposure standards. Ibrahim and Chen (2000) explored the use of a power transform prior, whereby the likelihood function of historical evidence is raised to the power of ρ , where $0 \leq \rho \leq 1$. Such an approach could be applied to our example and possibly extended to include more sources of evidence.

6. Discussion

We have introduced the potential for systematic review and formal synthesis models to be used in the context of assessing risks to human health from exposures to chemicals in the environment. The application of this methodology to the illustrative example of exposure to THMs and the risk of delivering a baby with a low birth weight has required many assumptions to be made. So also do more traditional approaches to analysing and interpreting evidence on health outcomes of environmental exposures. However, the use of systematic review and meta-analysis methods forces and facilitates explicit acknowledgement and description of these assumptions. In this example, the sensitivity of many of these assumptions has been investigated; depending on the dose-response model, the assumptions that are made can have a critical effect on the estimates that are obtained.

Our analyses have also demonstrated the necessity of checking the appropriateness of the dose-response model chosen. For simplicity, we initially assumed a linear relationship between $\ln(\text{OR})$ for low birth weight and $\ln(\text{dose})$. However, the logit model is among the most commonly applied dose-response models in toxicology, together with the probit and Weibull models (Crump, 1984; Covello and Merkhofer, 1994; Lovell and Thomas, 1997). In the example that we have used, the application of the Bayes information criterion has suggested that the linear model provides a better fit to the data than the logit model. The use of $\ln(\text{dose})$ seems more sensible than dose, as it reflects a multiplicative effect with increasing exposure. In fact, in toxicology and pharmacology, $\ln(\text{dose})$ is generally used (Rang *et al.*, 1999). In this example, at least for the purpose of synthesis, $\ln(\text{dose})$ is more appealing as the epidemiological slope estimates from the $\ln(\text{dose})$ dose-response model have a magnitude and precision that are comparable with those from the toxicological studies.

In assessing which synthesis model to recommend, account must be taken of the fact that, in this example, two different sources of evidence are being combined; hence we do not advocate the use of model 1e to synthesize all 13 studies. The synthesis models that appear to be the most advantageous, in terms of taking into account the two different sources of evidence, are those that used informative prior distributions (models 1c and 1d). Furthermore, by assessing the relevance of a source of evidence and allowing it to have more (or less) influence in the synthesis the effect of this particular source of evidence can be quantitatively explored. For both the dose-response and the synthesis models, methods of model comparison could be extended to the use of Bayes factors and averaging over models (Kass and Raftery, 1995).

7. Conclusion

Systematic review and formal synthesis methods ensure that all relevant evidence is included and that data are combined in a transparent and reproducible manner. We believe that this initial attempt to apply systematic and transparent methodology to the setting of environmental exposure standards is, in itself, a significant step forward in the chemical risk assessment and exposure standard setting process. Many assumptions had to be made and documented in transforming and analysing the data in this example for this to be achieved. However, these assumptions are parallel to those which are currently made in the risk assessment process, and our approach forces them to be made *explicit*. We recommend that systematic review methods be used in the risk assessment process as they provide a structured, transparent means of identifying and assessing the body of research concerning a particular area of interest. Sensitivity analyses should also be considered, not just in the development of this methodology, but in more general standard setting processes.

Since the main focus of this paper is to illustrate the potential benefits of applying systematic review and, if appropriate, meta-analysis methods in the risk assessment process, we do not, at this point, attempt to conclude whether there is an association between exposure to THMs and low birth weight, nor at what level of exposure the risk of such an effect may increase.

The next stages in the development of our approach will be to investigate another, quite different, example using these methods of systematic review and meta-analysis, to help to identify those assumptions and issues that are of general importance and those that are only relevant to specific examples. Furthermore, we want to build on the models that are presented in this paper to incorporate additional relevant information. We have already begun to investigate this by assessing the effect of changing the relevance of the animal data to the human exposures (Section 5.3). However, further information such as the different routes of exposure and different types of exposure (for instance individual chemicals *versus* mixtures of chemicals) could also be taken into account, together with available data on biological effects and mechanisms, and a Bayesian hierarchical approach offers an appealing and flexible framework for doing so.

Acknowledgements

We thank Ian Pate (Syngenta), Philip Holmes (Medical Research Council Institute for Environment and Health) and Frances Pollitt, Michael Wearing and Bob Maynard (UK Department of Health) for helpful discussions concerning this work. The work described in this paper was performed in part while David Jones was on study leave from the University of Leicester.

The Medical Research Council Institute for Environment and Health received funding for this work from the Department of Health (contract IEH/00/1); the views expressed in this publication are those of the authors and not necessarily those of the Department of Health.

References

- Bhatia, R., Lopipero, P. and Smith, A. H. (1998) Diesel exhaust exposure and lung cancer. *Epidemiology*, **9**, 84–91.
- Blettner, M., Sauerbrei, W., Schlehofer, B., Scheuchenpflug, T. and Friedenreich, C. (1999) Traditional reviews, meta-analyses and pooled analyses in epidemiology. *Int. J. Epidemiol.*, **28**, 1–9.
- Bove, F. J., Fulcomer, M. C., Klotz, J. B., Esmart, J., Dufficy, E. M. and Zagraniski, R. T. (1992) *Report on Phase IV-A: Public Drinking Water Contamination and Birthweight, Fetal Deaths, and Birth Defects*. Trenton: New Jersey Department of Health.
- Budtz-Jørgensen, E., Keiding, N. and Grandjean, P. (2001) Benchmark dose calculation from epidemiological data. *Biometrics*, **57**, 698–706.
- Covello, V. T. and Merkhofer, M. W. (1994) *Risk Assessment Methods: Approaches for Assessing Health and Environmental Risks*. New York: Plenum.
- Crump, K. S. (1984) A new method for determining allowable daily intakes. *Fundam. Appl. Toxicol.*, **4**, 854–871.
- Department of the Environment (1993) *Risk Assessment of Existing Substances*. London: Department of the Environment.
- Dodds, L., King, W., Woolcott, C. and Pole, J. (1999) Trihalomethanes in public water supplies and adverse birth outcomes. *Epidemiology*, **10**, 233–237.
- DuMouchel, W. and Groër, P. G. (1989) A Bayesian methodology for scaling radiation studies from animals to man. *Health Phys.*, **57**, suppl. 1, 411–418.
- DuMouchel, W. H. and Harris, J. E. (1983) Bayes methods for combining the results of cancer studies in humans and other species. *J. Am. Statist. Ass.*, **78**, 293–308.
- Egger, M. and Davey Smith, G. (2001) Principles of and procedures for systematic reviews. In *Systematic Reviews in Health Care: Meta-analysis in Context* (eds M. Egger, D. G. Altman and G. Davey Smith), pp. 23–42. London: British Medical Journal Books.
- Environment Agency and Department for Environment, Food and Rural Affairs (2002) *The Contaminated Land Exposure Assessment Model (CLEA): Technical Basis and Algorithms*. Bristol: Environment Agency.
- European Centre for Ecotoxicology and Toxicology of Chemicals (2001) *Exposure Factors Sourcebook for European Populations (with Focus on UK Data)*. Brussels: European Centre for Ecotoxicology and Toxicology of Chemicals.

- Fawell, J., Robinson, D., Bull, R. J., Birnbaum, L., Boorman, G. A., Butterworth, B., Daniel, P., Galal-Gorchev, H., Hauchman, F., Julkunen, P., Klaassen, C., Krasner, S., Orme-Zavaleta, J., Reif, J. and Tardiff, R. G. (1997) Disinfection by-products in drinking water: critical issues in health effects research. *Environ. Hlth Perspect.*, **105**, 108–109.
- Fleiss, J. L. and Gross, A. J. (1991) Meta-analysis in epidemiology, with special reference to studies of the association between exposure to environmental tobacco smoke and lung cancer: a critique. *J. Clin. Epidemiol.*, **44**, 127–139.
- Foster, P. M. and Auton, T. R. (1995) Application of benchmark dose risk assessment methodology to developmental toxicity: an industrial view. *Toxicol. Lett.*, **82–83**, 555–559.
- Gallagher, M. D., Nuckols, J. R., Stallones, L. and Savitz, D. A. (1998) Exposure to trihalomethanes and adverse pregnancy outcomes. *Epidemiology*, **9**, 484–489.
- Greenland, S. and Longnecker, M. P. (1992) Methods for trend estimation from summarised dose–response data, with applications to meta analysis. *Am. J. Epidemiol.*, **135**, 1301–1309.
- Ibrahim, J. G. and Chen, M.-H. (2000) Power prior distributions for regression models. *Statist. Sci.*, **15**, 46–60.
- Kass, R. E. and Raftery, A. E. (1995) Bayes Factors. *J. Am. Statist. Ass.*, **90**, 773–795.
- Koivusalo, M. and Vartiainen, T. (1997) Drinking water chlorination by-products and cancer. *Rev. Environ. Hlth*, **12**, 81–90.
- Kramer, M. D., Lynch, C. F., Isacson, P. and Hanson, J. W. (1992) The association of waterborne chloroform with intrauterine growth retardation. *Epidemiology*, **3**, 407–413.
- Lambert, P. C., Sutton, A. J., Burton, P. B., Abrams, K. R. and Jones, D. R. (2003) How vague is vague?: a simulation study of the impact of the use of vague prior distributions in MCMC. *Technical Report 03-01*. Department of Epidemiology and Public Health, University of Leicester, Leicester.
- Larose, D. T. and Dey, D. K. (1997) Grouped random effects models for Bayesian meta-analysis. *Statist. Med.*, **16**, 1817–1829.
- Lovell, D. P. and Thomas, G. (1997) Quantitative risk assessment. In *Food Chemical Risk Analysis* (ed. D. R. Tennant), pp. 57–86. London: Blackie.
- Morris, R. D., Audet, A.-M., Angelillo, I. F., Chalmers, T. C. and Mosteller, F. (1992) Chlorination, chlorination by-products, and cancer: a meta-analysis. *Am. J. Publ. Hlth*, **82**, 955–963.
- Muller, P., Parmigiani, G., Schildkraut, J. and Tardella, L. (1999) A Bayesian hierarchical approach for combining case-control and prospective studies. *Biometrics*, **55**, 858–866.
- Narotsky, M. G., Pegram, R. A. and Kavlock, R. J. (1997) Effect of dosing vehicle on the developmental toxicity of bromodichloromethane and carbon tetrachloride in rats. *Fundam. Appl. Toxicol.*, **40**, 30–36.
- National Health Service Centre for Reviews and Dissemination (2001) *NHS Centre for Reviews and Dissemination—Undertaking Systematic Reviews of Research on Effectiveness*. York: University of York. (Available from <http://www.york.ac.uk/inst/crd/report4.htm>.)
- Nieuwenhuijsen, M. J., Toledano, M. B., Eaton, N. E., Fawell, J. and Elliott, P. (2000) Chlorination disinfection by-products in water and their association with adverse reproductive outcomes: a review. *Occup. Environ. Med.*, **57**, 73–85.
- Office for National Statistics (2000) *Health Statistics Quarterly*, 7. London: Office for National Statistics. (Available from <http://www.statistics.gov.uk/downloads/theme.health/HSQ7Book.pdf>.)
- Peters, J. L., Rushton, L., Sutton, A. J., Jones, D. R., Abrams, K. R. and Mugglestone, M. A. (2003) Bayesian methods for the cross design synthesis of epidemiological and toxicological evidence. *Technical Report 03-02*. Department of Epidemiology and Public Health, University of Leicester, Leicester.
- Prevost, T. C., Abrams, K. R. and Jones, D. R. (2000) Hierarchical models in generalized synthesis of evidence: an example based on studies of breast cancer screening. *Statist. Med.*, **19**, 3359–3376.
- Rang, H. P., Dale, M. M. and Ritter, J. M. (1999) *Pharmacology*. Edinburgh: Churchill Livingstone.
- Reif, J. S., Hatch, M. C., Bracken, M., Holmes, L. B., Schwetz, B. A. and Singer, P. C. (1996) Reproductive and developmental effects of disinfection by-products in drinking water. *Environ. Hlth Perspect.*, **104**, 1056–1061.
- Risk Assessment and Toxicology Steering Committee (1999) *Risk Assessment Approaches used by UK Government for Evaluating Human Health Effects of Chemicals*. Leicester: Medical Research Council Institute for Environment and Health, Risk Assessment and Toxicology Steering Committee. (Available from <http://ie.ac.uk/ieh>.)
- Roberts, K. A., Dixon-Woods, M., Fitzpatrick, R., Abrams, K. R. and Jones, D. R. (2002a) Factors affecting uptake of childhood immunisation: a Bayesian synthesis of qualitative and quantitative evidence. *Lancet*, **360**, 1596–1599.
- Roberts, I., Kwan, I., Evans, P. and Haig, S. (2002b) Does animal experimentation inform human healthcare?: observations from a systematic review of international animal experiments on fluid resuscitation. *Br. Med. J.*, **324**, 474–476.
- Ruddick, J. A., Villeneuve, D. C. and Chu, I. (1983) A teratological assessment of four trihalomethanes in the rat. *J. Environ. Sci. Hlth B*, **18**, 333–349.
- Sandercock, P. and Roberts, I. (2002) Systematic reviews of animal experiments. *Lancet*, **360**, 586.
- Savitz, D. A., Andrews, K. W. and Pastore, L. M. (1995) Drinking water and pregnancy outcome in Central North Carolina: source, amount and trihalomethane levels. *Environ. Hlth Perspect.*, **103**, 592–596.
- Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.

- Spiegelhalter, D. J., Myles, J. P., Jones, D. R. and Abrams, K. R. (2000a) Bayesian methods in health technology assessment. *Hlth Technol. Assessmnt*, **4**, 1–122.
- Spiegelhalter, D. J., Thomas, A. and Best, N. (2000b) *WinBugs Version 1.3: User Manual*. Cambridge: Medical Research Council Biostatistics Unit.
- StataCorp (2001) *Stata Statistical Software: Release 7.0*. College Station: Stata Corporation.
- Sutton, A. J. and Abrams, K. R. (2001) Bayesian methods in meta-analysis and evidence synthesis. *Statist. Meth. Med. Res.*, **10**, 277–303.
- Sutton, A. J., Abrams, K. R., Jones, D. R., Sheldon, T. A. and Song, F. (2000) *Methods for Meta-analysis in Medical Research*. Chichester: Wiley.
- Thompson, D. J., Warner, S. D. and Robinson, V. B. (1974) Teratology studies on orally administered chloroform in the rat and rabbit. *Toxicol. Appl. Pharmacol.*, **29**, 348–357.
- Whitaker, H., Nieuwenhuijsen, M. J., Best, N., Fawell, J., Gowers, A. and Elliot, P. (2003) Description of trihalo-methane levels in three UK water supplies. *J. Expo. Anal. Environ. Epidem.*, **13**, 17–23.
- World Health Organization (1996) *Guidelines for Drinking-water Quality*. Geneva: World Health Organization.

Appendix F: Search strategy for articles

List of electronic databases used in the systematic literature review

MEDLINE

EMBASE

Web of Science

TOXLINE

CANCERLIT

BIOSIS

SCISEARCH

ENVIROLINE

POLLUTION ABSTRACTS

List of terms used in the electronic database searches

[\$ indicates truncation, such that 'pregnan\$' retrieves pregnant, pregnancy, pregnancies and so on]

Drinking (adjacent to) Water

or Water-Pollutants-Chemicals

or Water (adjacent to) Supply

or Water (adjacent to) Pollution or Pollutants

or Water-Management

or Water-Quality

AND

Water/ Chlorin\$

or Trihalomethan\$

or Haloace\$
or Disinfection
or Disinfectant
or Disinfectant-Agent
or Byproducts or By (adjacent to) Products

AND

Genital-System-Function
or Reproductive-Toxicity
or Gestation
or Pregnan\$
or Infant Newborn
or Pregnancy Outcome
or Abnormalities
or Birth
or Fetus
or Reproduction
or Reproductive

Appendix G: Search strategy and list of included references for Mn articles

List of terms used in the Medline search

Manganese (in title, abstract, medical subject headings)
Manganese (adjacent to) ore\$1 (in title, abstract, medical subject headings)
Pyrolusite (in title, abstract, medical subject headings)
Pyrochroite (in title, abstract, medical subject headings)
Cianciulliite (in title, abstract, medical subject headings)
Manganese (adjacent to) oxide\$1 (in title, abstract, medical subject headings)
MNO (in title, abstract, medical subject headings)
Manganese (adjacent to) II (adjacent to) oxide (in title, abstract, medical subject headings)
Manganese (adjacent to) IV (adjacent to) oxide (in title, abstract, medical subject headings)
MNO₂ (as keyword)
MNO (adjacent to) 2 (in title, abstract, medical subject headings)
Polianite (in title, abstract, medical subject headings)
Ramsdellite (in title, abstract, medical subject headings)
Manganese (adjacent to) III (adjacent to) Oxide (in title, abstract, medical subject headings)
MN₂O₃ (as keyword)
Braunite (in title, abstract, medical subject headings)
MN (adjacent to) 2 (adjacent to) O (adjacent to) 3
MN₃O₄ (as keyword)
MN (adjacent to) 3 (adjacent to) O (adjacent to) 4
Hausmannite (in title, abstract, medical subject headings)
MN₃O₇ (as keyword)
MN (adjacent to) 3 (adjacent to) O (adjacent to) 7
MN₅O₈ (as keyword)
MN (adjacent to) 5 (adjacent to) O (adjacent to) 8
NA₃MNO₄ (as keyword)
NA (adjacent to) 3 (adjacent to) MNO (adjacent to) 4
Sodium (adjacent to) manganate (in title, abstract, medical subject headings)

BAMNO4 (as keyword)
Barium (adjacent to) manganate (in title, abstract, medical subject headings)
MNSO4 (as keyword)
MNSO (adjacent to) 4
Manganese (adjacent to) sulphate (in title, abstract, medical subject headings)
Manganese (adjacent to) II (adjacent to) sulphate (in title, abstract, medical subject headings)
MNCL2 (as keyword)
Manganese (adjacent to) chloride (in title, abstract, medical subject headings)
Manganese (adjacent to) II (adjacent to) chloride (in title, abstract, medical subject headings)
MN (adjacent to) NO3 (adjacent to) 2
Manganese (adjacent to) nitrate (in title, abstract, medical subject headings)
Manganese (adjacent to) II (adjacent to) nitrate (in title, abstract, medical subject headings)
Manganous (adjacent to) salt\$1 (in title, abstract, medical subject headings)
MN2 (adjacent to) SO4 (adjacent to) 3
Manganese (adjacent to) III (adjacent to) sulphate (in title, abstract, medical subject headings)
MN (adjacent to) SO4 (adjacent to) 2
Manganese (adjacent to) IV (adjacent to) sulphate (in title, abstract, medical subject headings)
FEMN (in title, abstract, medical subject headings)
Ferromanganese (in title, abstract, medical subject headings)
SIMN (in title, abstract, medical subject headings)
Silicamanganese (in title, abstract, medical subject headings)
FESIMN (in title, abstract, medical subject headings)
Ferrosilicamanganese (in title, abstract, medical subject headings)
Manganese with steel (in title, abstract, medical subject headings)
Potassium (adjacent to) permanganate (in title, abstract, medical subject headings)
KMNO4 (as keyword)
KMNO (adjacent to) 4
Potassium (adjacent to) manganate (adjacent to) VII (in title, abstract, medical subject headings)
Potassium (adjacent to) VII (adjacent to) manganate (in title, abstract, medical subject headings)
Manganese-compounds (exploded)
Manganese (exploded)

AND

Neurotoxicity-syndromes (exploded)

Toxicology (exploded)

Toxic\$7 (in title, abstract, medical subject headings)

Nervous-system (exploded)

Nervous-system-diseases (exploded)

Toxicity-tests (exploded)

Manganese-poisoning (exploded)

55 Human studies

- Abd El Naby S, Hussanein M (1965) Neuropsychiatric manifestations of chronic manganese poisoning. *Journal of Neurology, Neurosurgery and Psychiatry* 28:282-287.
- Amr M, Allam M, Osmaan AL, el-Batanouni M, el Samra G, Halim Z (1993) Neurobehavioral changes among workers in some chemical industries in Egypt. *Environmental Research* 63:295-300.
- Barrington WW, Angle CR, Willcockson NK, Padula MA, Korn T (1998) Autonomic function in manganese alloy workers. *Environmental Research* 78:50-58.
- Beuter A, Edwards R, deGeoffroy A, Mergler D, Hudnell K (1999) Quantification of neuromotor function for detection of the effects of manganese. *Neurotoxicology* 20:355-366.
- Beuter A, Mergler D, de Geoffroy A, Carriere L, Belanger S, Varghese L, Sreekumar J, Gauthier S (1994) Diadochokinesimetry: a study of patients with Parkinson's disease and manganese exposed workers. *Neurotoxicology* 15:655-664.
- Brown DSO, Wills CE, Yousefi V, Nell V (1991) Neurotoxic effects of chronic exposure to manganese dust. *Neuropsychiatry, Neuropsychology and Behavioral Neurology* 4:238-250.
- Camerino D, Cassitto MG, Gilioli R (1993) Prevalence of abnormal neurobehavioural scores in populations exposed to different industrial chemicals. *Environmental Research* 61:251-257.
- Chia SE, Foo SC, Gan SL, Jeyaratnam J, Tian CS (1993) Neurobehavioral functions among workers exposed to manganese ore. *Scandinavian Journal of Work, Environment and Health* 19:264-270.
- Chia SE, Gan SL, Chua LH, Foo SC, Jeyaratnam J (1995) Postural stability among manganese exposed workers. *Neurotoxicology* 16:519-526.
- Crump KS, Rousseau P (1999) Results from eleven years of neurological health surveillance at a manganese oxide and salt producing plant. *Neurotoxicology* 20:273-286.
- Deschamps FJ, Guillaumot M, Raux S (2001) Neurological effects in workers exposed to manganese. *Journal of Occupational and Environmental Medicine* 43:127-132.
- Dietz MC, Ihrig A, Wrazidlo W, Bader M, Jansen O, Triebig G (2001) Results of magnetic resonance imaging in long-term manganese dioxide-exposed workers. *Environmental Research* 85:37-40.
- Discalzi G, Pira E, Herrero Hernandez E, Valentini C, Turbiglio M, Meliga F (2000) Occupational Mn parkinsonism: magnetic resonance imaging and clinical patterns following CaNa₂-EDTA chelation. *Neurotoxicology* 21:863-866.
- EC (1997) *Study of toxic effects in the central and peripheral nervous system of workers in the ferro-alloy industry*. Health and Safety at Work, European Commission, Brussels, Belgium.

- Emara AM, El-Ghawabi SH, Madkour OI, El-Samra G (1971) Chronic manganese poisoning in the dry battery industry. *British Journal of Industrial Medicine* 28:78-82.
- Flinn RH, Neal PA, Fulton WB (1941) Industrial manganese poisoning. *Journal of Industrial Hygiene and Toxicology* 23:374-387.
- Gibbs JP, Crump KS, Houck DP, Warren PA, Mosley WS (1999) Focused medical surveillance: a search for subclinical movement disorders in a cohort of US workers exposed to low levels of manganese dust. *Neurotoxicology* 20:299-314.
- Hochberg F, Miller G, Valenzuela R, McNelis S, Crump KS, Covington T, Valdivia G, Hochberg B, Trustman JW (1996) Late motor deficits of Chilean manganese miners: a blinded control study. *Neurology* 47:788-795.
- Hua M-S, Huang C-C (1991) Chronic occupational exposure to manganese and neurobehavioural function. *Journal of Clinical and Experimental Neuropsychology* 13:495-507.
- Huang C-C, Chu N-S, Lu C-S, Chen Rs, Calne DB (1998) Long term progression in chronic manganism: ten years of follow-up. *Neurology* 50:698-700.
- Huang C-C, Chu N-S, Lu C-S, Wang J-D, Tsai J-L, Tzeng J-L, Wolters EC, Calne DB (1989) Chronic manganese intoxication. *Archives of Neurology* 46:1104-1106.
- Huang C-C, Lu C-S, Chu N-S, Hochberg F, Lilienfeld D, Olanow W, Calne DB (1993) Progression after chronic manganese exposure. *Neurology* 43:1479-1483.
- Huang Q, Liu W, Pan C (1990) The neurobehavioral changes of ferromanganese smelting workers. *Occupational Epidemiology* 329-332.
- Iregren A (1990) Psychological test performance in foundry workers exposed to low levels of manganese. *Neurotoxicology and Teratology* 12:673-675.
- Jonderko G, Kujawsa A, Langauer-Lewowicka H (1971) Problems of chronic manganese poisoning on the basis of investigations of workers at a manganese alloy foundry. *Internationales Archiv fuer Arbeitsmedizin* 28:250-264.
- Kaji H, Ohsaki Y, Rokujo C, Higashi T, Fujino A, Kamada T (1993) Determination of blood and urine manganese (Mn) concentrations of static sensography as the indices of Mn-exposure among Mn-refinery workers. *Journal of the University of Occupational and Environmental Health* 15:287-296.
- Kawamura R, Ikuta H, Fukuzumi S, Yamada R, Tsubaki S, Kodama T, Kurata S (1941) Intoxication by manganese in well water. *Kisasato Archives of Experimental Medicine* 18:145-169.
- Kilburn CJ (1987) Manganese, malformations and motor disorders: findings in a manganese-exposed population. *Neurotoxicology* 8:421-430.
- Kilburn KH (1998) Neurobehavioral impairment and symptoms associated with aluminium remelting. *Archives of Environmental Health* 53:329-335.

- Kilburn KH (1999) Neurobehavioral and respiratory findings in jet engine repair workers: a comparison of exposed and unexposed volunteers. *Environmental Research* 80:244-252.
- Kim Y, Kim KS, Yang JS, Park IJ, Kim E, Jin Y, Kwon K-R, Chang KH, Kim J-W, Park S-H, Lim H-S, Cheong H-K, Shin YC, Park J, Moon Y (1999) Increase in signal intensities on T1-weighted magnetic resonance images in asymptomatic manganese-exposed workers. *Neurotoxicology* 20:901-907.
- Kondakis XG, Makris N, Leotsinidis M, Prinou M, Papapetropoulos T (1989) Possible health effects of high manganese concentration in drinking water. *Archives of Environmental Health* 44:175-178.
- Lucchini R, Apostoli P, Perrone C, Placidi D, Albin E, Migliorati P, Mergler D, Sassine M-P, Palmi S (1999) Long term exposure to "low levels" of manganese oxides and neurofunctional changes in ferroalloy workers. *Neurotoxicology* 20:287-298.
- Lucchini R, Bergamaschi E, Smargiassi A, Festa D, Apostoli P (1997) Motor function, olfactory threshold, and hematological indices in manganese-exposed ferroalloy workers. *Environmental Research* 73:175-180.
- Lucchini R, Selis L, Folli D, Apostoli P, Mutti AV, O., Iregren A, Alessio L (1995) Neurobehavioral effects of manganese in workers from a ferroalloy plant after temporary cessation of exposure. *Scandinavian Journal of Work, Environmental and Health* 21:143-149.
- Mena I, Marin O, Fuenzalida S, Cotzias GC (1967) Chronic manganese poisoning: clinical picture and manganese turnover. *Neurology* 17:128-136.
- Mergler D, Baldwin M, Belanger S, Larribe F, Beuter A, Bowler R, Panisset M, Edwards R, de Geoffroy A, Sassine M-P, Hudnell K (1999) Manganese neurotoxicity, a continuum of dysfunction: Results from a community based study. *Neurotoxicology* 20:327-342.
- Mergler D, Huel G, Bowler R, Iregren A, Bélanger S, Baldwin M, Tardif R, Smargiassi A, Martin L (1994) Nervous system dysfunction among workers with long-term exposure to manganese. *Environmental Research* 64:151-180.
- Rodier J (1955) Manganese poisoning in Moroccan mines. *British Journal of Industrial Medicine* 12:21-35.
- Roels H, Eslava MIO, Ceulemans E, Robert A, Lison D (1999) Prospective study on the reversibility of neurobehavioral effects in workers exposed to manganese dioxide. *Neurotoxicology* 20:255-271.
- Roels H, Lauwerys R, Buchet J-P, Genet P, Sarhan MJ, Hanotiau I, de Fays M, Bernard A, Stanescu D (1987) Epidemiological survey among workers exposed to manganese: effects on lung, central nervous system and some biological indices. *American Journal of Industrial Medicine* 11:307-327.
- Roels H, Sarhan MJ, Hanotiau I, de Fays M, Genet P, Bernard A, Buchet J-P, Lauwerys R (1985) Preclinical toxic effects of manganese in workers from a Mn salts and oxides producing plant. *Science of the Total Environment* 42:201-206.

- Roels HA, Ghyselen P, Buchet J-P, Ceulemans E, Lauwerys RR (1992) Assessment of the permissible exposure level to manganese in workers exposed to manganese dioxide dust. *British Journal of Industrial Medicine* 49:25-34.
- Saric M, Markicevic A, Hrustic O (1977) Occupational exposure to manganese. *British Journal of Industrial Medicine* 34:114-118.
- Schuler P, Oyanguren H, Maturana V, Valenzuela A, Cruz E, Plaza V, Schmidt E, Haddad R (1957) Manganese poisoning: environmental and medical study at a Chilean mine. *Industrial Medicine and Surgery* 26:167-173.
- Sjögren B, Gustavsson P, Hogstedt C (1990) Neuropsychiatric symptoms among welders exposed to neurotoxic metals. *British Journal of Industrial Medicine* 47:704-707.
- Sjögren B, Iregren A, Frech W, Hagman M, Johansson L, Tesarz M, Wennberg A (1996) Effects on the nervous system among welders exposed to aluminium and manganese. *Occupational and Environmental Medicine* 53:32-40.
- Stýblová V, Bencko V, Drobný M, Chumchal O, Rimská V, Žlab L (1979) Clinical and epidemiological study in workers exposed to manganese. *Activ nerv suppl* 21:290-291.
- Vieregge P, Heinzow B, Korf G, Teichert H-M, Schleifenbaum P, Mosinger H-U (1995) Long term exposure to manganese in rural well water has no neurological effects. *Canadian Journal of Neurological Sciences* 22:286-289.
- Wang J-D, Huang C-C, Hwang Y-H, Chiang J-R, Lin J-M, Chen J-S (1989) Manganese induced parkinsonism: an outbreak due to unrepaired ventilation control system in a ferromanganese smelter. *British Journal of Industrial Medicine* 46:856-859.
- Wennberg A, Hagman M, Johansson L (1992) Preclinical neurophysiological signs of parkinsonism in occupational manganese exposure. *Neurotoxicology* 13:271-274.
- Wennberg A, Iregren A, Struwe G, Cizinsky G, Hagman M, Johansson L (1991) Manganese exposure in steel smelters: a health hazard to the nervous system. *Scandinavian Journal of Work, Environment and Health* 17:255-262.
- Whitlock CMJ, Amuso SJ, Bittenbender JB (1966) Chronic neurological disease in two manganese steel workers. *American Industrial Hygiene Association Journal* 27:454-459.

37 Animal experiments

- Bird ED, Anton AH, Bullock B (1984) The effect of manganese inhalation on basal ganglia dopamine concentrations in rhesus monkey. *Neurotoxicology* 5:59-66.
- Bonilla E (1984) Chronic manganese intake induces changes in the motor activity of rats. *Experimental Neurology* 84:696-700.
- Calabresi P, Ammassari-Teule M, Gubellini P, Sancesario G, Morello M, Centonze D, Marfia GA, Saulle E, Passino E, Picconi B, Bernardi G (2001) A synaptic mechanism underlying the behavioral abnormalities induced by manganese intoxication. *Neurobiology of Disease* 8:419-432.
- Chandra AV, Ali MM, Saxena DK, Murthy RC (1981) Behavioural and neurochemical changes in rats simultaneously exposed to manganese and lead. *Archives of Toxicology* 49:49-56.
- Chandra SV (1972) Histological and histochemical changes in experimental manganese encephalopathy in rabbits. *Archives of Toxicology* 29:29-38.
- Chandra SV (1983) Psychiatric illness due to manganese poisoning. *Acta Psychiatrica Scandinavica* 67:49-54.
- Chandra SV, Shukla GS (1978) Manganese encephalopathy in growing rats. *Environmental Research* 15:28-37.
- Derevenco P, Vaida A, Stoica N, Gabor S, Ivanof L, Botoc M, Alex L, Baci I (1988) Neurobehavioral and biochemical response to manganese and selenium exposure in rats. *Physiologie* 25:111-118.
- Dorman DC, Struve MF, Vitarella D, Byerly FL, Goetz J, Miller R (2000) Neurotoxicity of manganese chloride in neonatal and adult CD rats following subchronic (21-day) high-dose exposure. *Journal of Applied Toxicology* 20:179-187.
- Eriksson H, Gillberg P-G, Aquilonius S-M, Hedstroem K-G, Heilbronn E (1992a) Receptor alterations in manganese intoxicated monkeys. *Archives of Toxicology* 66:359-364.
- Eriksson H, Maegiste K, Plantin L-O, Fonnum F, Hedstroem K-G, Theodorsson-Norheim E, Kristensson K, Staalberg E, Heilbronn E (1987) Effects of manganese oxide on monkeys as revealed by a combined neurochemical, histological and neurophysiological evaluation. *Archives of Toxicology* 61:46-52.
- Eriksson H, Tedroff J, Thuomas KA, Aquilonius S-M, Hartvig P, Fasth K-J, Bjurling P, Laangstroem B, Hedstroem K-G, Heilbronn E (1992b) Manganese induced brain lesions in *Macaca fascicularis* as revealed by positron emission tomography and magnetic resonance imaging. *Archives of Toxicology* 66:403-407.
- Gray LEJ, Laskey JW (1980) Multivariate analysis of the effects of manganese on the reproductive physiology and behavior of the male house mouse. *Journal of Toxicology and Environmental Health* 6:861-867.
- Gupta SK, Murthy RC, Chandra SV (1980) Neuromelanin in manganese-exposed primates. *Toxicology Letters* 6:17-20.

- Ingersoll RT, Montgomery EBJ, Aposhian HV (1995) Central nervous system toxicity of manganese. Part 1. Inhibition of spontaneous motor activity in rats after intrathecal administration of manganese chloride. *Fundamental and Applied Toxicology* 27:106-113.
- Jonderko G, Wegiel A (1966) Reduced glutathione of erythrocytes in acute experimental manganese poisoning. *Polish Medical Journal* 5:944-946.
- Khan KN, Andress JM, Smith PF (1997) Toxicity of subacute intravenous manganese chloride administration in beagle dogs. *Toxicologic Pathways* 25:344-350.
- Komura J, Sakamoto M (1991) Short-term oral administration of several manganese compounds in mice: physiological and behavioral alterations caused by different forms of manganese. *Bulletin of Environmental Contamination and Toxicology* 46:921-928.
- Komura J, Sakamoto M (1992) Effects of manganese forms on biogenic amines in the brain and behavioral alterations in the mouse: long term oral administration of several manganese compounds. *Environmental Research* 57:34-44.
- Kristensson K, Eriksson H, Lundh B, Plantin L-O, Wachtmeister L, el Azazi M, Morath C, Heilbronn E (1986) Effects of manganese chloride on the rat developing nervous system. *Acta Pharmacologica et Toxicologica* 59:345-348.
- Laskey JW, Edens FW (1985) Effects of chronic high-level manganese exposure on male behavior in the Japanese quail (*Coturnix coturnix japonica*). *Poultry Science* 64:579-584.
- Lown BA, Morganti JB, Agostino RB, Stineman CH, Massaro EJ (1984) Effects of the postnatal development of the mouse of preconception, postconception and/or suckling exposure to manganese via maternal inhalation exposure to MnO₂ dust. *Neurotoxicology* 5:119-129.
- Morganti JB, Lown BA, D'agostino RB, Stineman CH, Massaro EJ (1982) Uptake, pathology and behavioral effects of inhalation exposure to manganese (MnO₂) in the mouse. *Neurotoxicology* 3:148-149.
- Murthy RC, Lal S, Saxena DK, Shukla GS, Ali MM, Chandra SV (1981) Effect of manganese and copper interaction on behavior and biogenic amines in rats fed a 10% casein diet. *Chemico-biological Interactions* 37:299-308.
- Mustafa SJ, Chandra SV (1971) Levels of 5-hydroxytryptamine, dopamine and norepinephrine in whole brain of rabbits in chronic manganese toxicity. *Journal of Neurochemistry* 18:931-933.
- Nachtman JP, Tubbem RE, Commissaris RL (1986) Behavioral effects of chronic manganese administration in rats: locomotor activity studies. *Neurobehavioral Toxicology and Teratology* 8:711-715.
- Neff NH, Barrett RE, Costa E (1969) Selective depletion of caudate nucleus dopamine and serotonin during chronic manganese dioxide administration to squirrel monkeys. *Experientia* 25:1140-1141.

- Newland MC, Weiss B (1992) Persistent effects of manganese on effortful responding and their relationship to manganese accumulation in the primate *globus pallidus*. *Toxicology and Applied Pharmacology* 113:87-97.
- Nishiyama K, Suzuki Y, Fujii N, Yano H, Ohnishi K, Miyati T (1977) Biochemical changes and manganese distribution in monkeys exposed to manganese dioxide dust. *Tokushima Journal of Experimental Medicine* 24:137-145.
- Olanow CW, Good PF, Shinotoh H, Hewitt KA, Vingerhoets F, Snow BJ, Beal MF, Calne DB, Perl DP (1996) Manganese intoxication in the rhesus monkey: a clinical, imaging, pathologic and biochemical study. *Neurology* 46:492-498.
- Pappas BA, Zhang D, Davidson CM, Crowder T, Park GAS, Fortin T (1997) Perinatal manganese exposure: behavioral, neurochemical and histopathological effects in the rat. *Neurotoxicology and Teratology* 19:17-25.
- Roels H, Meiers G, Delos M, Ortega I, Lauwerys R, Buchet J-P, Lison D (1997) Influence of the route of administration and chemical form (MnCl₂, MnO₂) on the absorption and cerebral distribution of manganese in rats. *Archives of Toxicology* 71:223-230.
- Senturk UK, Oner G (1996) The effect of manganese-induced hypercholesterolemia on learning in rats. *Biology of Trace Element Research* 51:249-257.
- Shinotoh H, Snow BJ, Hewitt KA, Pate BD, Doudet D, Nugent R, Perl DP, Olanow W, Calne DB (1995) MRI and PET studies of manganese-intoxicated monkeys. *Neurology* 45:1199-1204.
- Subhash MN, Padmashree TS (1991) Effect of manganese on biogenic amine metabolism in regions of the rat brain. *Food and Chemical Toxicology* 29:579-582.
- Suzuki Y, Mouri T, Suzuki Y, Nishiyama K, Fujii N, Yano H (1975) Study of subacute toxicity of manganese dioxide in monkeys. *The Tohoku Journal of Experimental Medicine* 22:5-10.
- Yu IJ, Song KS, Chang HK, Han JH, Kim KJ, Chung YH, Maeng SH, Park SH, Han KT, Chung KH, Chung HK (2001) Lung fibrosis in Sprague-Dawley rats, induced by exposure to manual metal acr-stainless steel welding fumes. *Toxicological Sciences* 63:99-106.

Appendix H: Peters *et al.* (2006) JAMA paper

Comparison of Two Methods to Detect Publication Bias in Meta-analysis

Jaime L. Peters, MSc

Alex J. Sutton, PhD

David R. Jones, PhD

Keith R. Abrams, PhD

Lesley Rushton, PhD

SYSTEMATIC REVIEWS AND META-analyses are commonly used to identify and evaluate evidence about interventions or exposures in human health. Even when conducted thoroughly, systematic reviews and meta-analyses can be subject to publication bias—studies being less likely to be published, hence less likely to be included in a systematic review or meta-analysis because of the size and/or statistical significance of their estimate of effect.¹ If publication bias occurs, the subsequent systematic review or meta-analysis of published literature may be misleading.

Of the methods available to researchers for the detection of publication bias, one of the simplest is the funnel plot.² This is a scatterplot of the estimate of effect from each study in the meta-analysis against a measure of its precision, usually $1/SE$ (FIGURE 1A). In the absence of bias, the plot should resemble a “funnel shape,” as smaller, less precise studies are more subject to random variation than larger studies when estimating an effect. In the presence of publication bias, some smaller studies reporting negative results will be missing, leading to an asymmetrical funnel plot. Of course, publication bias is not the only possible explanation for observed (or tested) funnel plot asymmetry.³ Between-study heterogeneity and small-study effects (the tendency for

Context Egger's regression test is often used to help detect publication bias in meta-analyses. However, the performance of this test and the usual funnel plot have been challenged particularly when the summary estimate is the natural log of the odds ratio (lnOR).

Objective To compare the performance of Egger's regression test with a regression test based on sample size (a modification of Macaskill's test) with lnOR as the summary estimate.

Design Simulation of meta-analyses under a number of scenarios in the presence and absence of publication bias and between-study heterogeneity.

Main Outcome Measures Type I error rates (the proportion of false-positive results) for each regression test and their power to detect publication bias when it is present (the proportion of true-positive results).

Results Type I error rates for Egger's regression test are higher than those for the alternative regression test. The alternative regression test has the appropriate type I error rates regardless of the size of the underlying OR, the number of primary studies in the meta-analysis, and the level of between-study heterogeneity. The alternative regression test has comparable power to Egger's regression test to detect publication bias under conditions of low between-study heterogeneity.

Conclusion Because of appropriate type I error rates and reduction in the correlation between the lnOR and its variance, the alternative regression test can be used in place of Egger's regression test when the summary estimates are lnORs.

JAMA. 2006;295:676-680

www.jama.com

smaller studies to show greater effects than larger studies) are also possible explanations.³ However, when the study summary estimates are odds ratios (ORs), there is a correlation between the natural log of OR (lnOR) and its SE, since the variance is a function of lnOR.⁴ This correlation is stronger the further the estimated OR is from unity.

Thus, some asymmetry observed in a funnel plot may be due to this correlation rather than publication bias. The effect of this correlation can be avoided by plotting effect size estimates against sample size, rather than precision. The meta-analysis plotted in Figure 1A uses data simulated from a model with no publication bias. However, it appears that some small negative studies could be missing from the bottom left-hand corner, which could

be interpreted as indicating publication bias. When these data are plotted against sample size (Figure 1B), the funnel plot looks more symmetrical. Although Figures 1A and 1B are not remarkably different, since only the y-axis has changed, the impact on Egger's regression test can be quite striking, especially if the underlying OR is far from null.

Author Affiliations: Centre for Biostatistics and Genetic Epidemiology, Department of Health Sciences, University of Leicester, Leicester, England (Drs Sutton, Jones, and Abrams, and Ms Peters); MRC Institute for Environment and Health, Leicester, England (Dr Rushton). Dr Rushton is now with the Department of Epidemiology and Public Health, Imperial College London, London, England.

Corresponding Author: Jaime Peters, MSc, Centre for Biostatistics and Genetic Epidemiology, Department of Health Sciences, University of Leicester, 22-28 Princess Rd W, Leicester, LE1 6TP, England (jip9@leicester.ac.uk).

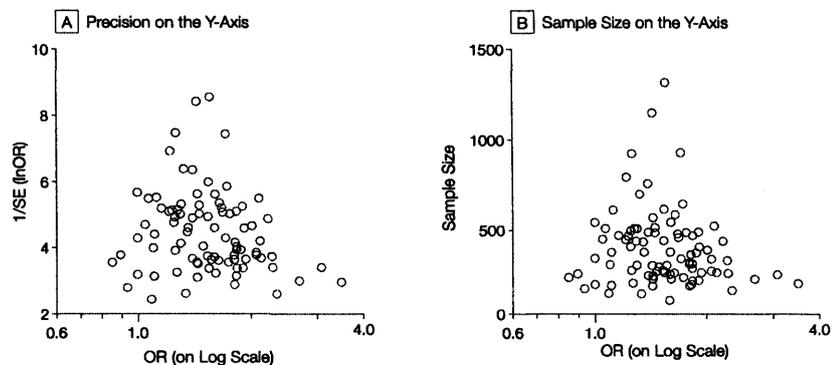
Statistical tests have been developed to provide more formal assessments for publication bias than the inspection of funnel plots. Egger's regression test⁵ is widely used (eg, as of January 11, 2006, the Web of Knowledge⁶ included 819 articles citing this article), is implemented in a number of software packages,⁷⁻¹⁰ and has become a standard procedure (eg, of 43 meta-analyses published in *JAMA* since 1997 in which an assessment of publication bias was made, 13 reported using Egger's regression test). Since it is based directly on the funnel plot, where the standardized effect estimate (*effect/SE*) is regressed on a measure of precision ($1/SE$), Egger's regression test is also subject to the effects of the correlation when using ORs.

In fact, Egger's regression test has been challenged because of its high type I error rates (the proportion of false-positive results) when ORs are used,^{3,11,12} a probable symptom of this correlation. As almost one third of the *JAMA* articles reviewed above used Egger's regression test when the summary estimates were ORs, this needs investigation. Using simulation analyses, we confirm that Egger's regression test is indeed inappropriate for ORs, particularly when the ORs are large and there is considerable between-study heterogeneity.^{3,4,11} We describe a simple alternative (a modified version of Macaskill's test,⁴ which is little used in practice) for detecting funnel plot asymmetry that avoids this correlation.

METHODS

We assessed the performance of 8 regression tests for funnel plot asymmetry, including Egger's regression test, using simulation methods. The tests¹³ differ in terms of the independent variable used, the weighting used, and whether random effects were included. In this article we compare the performance of Egger's regression test and the test found to have the most desirable properties compared with the remaining regression tests (results for all tests can be found in Peters et al¹³). Other modified tests are also being developed.¹⁴

Figure 1. Funnel Plots of a Meta-analysis Simulated With No Publication Bias



lnOR indicates natural log of the odds ratio; SE, standard error.

Characteristics of the simulated meta-analyses were based on a systematic review of meta-analyses of animal experiments,¹⁵ but the findings can be applied generally. Meta-analyses of 6, 16, 30, and 90 primary studies with underlying ORs of 1, 1.2, 1.5, 3, and 5 were simulated. The control group event rate was allowed to vary for each primary study. It was sampled from a uniform distribution with lower and upper limits of 0.3 and 0.7, respectively, representing a fairly common event in the control group. The treatment group event rate was calculated from this and the assumed underlying OR. The number of subjects in the control group in each study was based on the exponential of the normal distribution with a mean of 5 and variance of 0.3. The ratio of control to treated/exposed subjects was 1. The median sample size was around 300 in each simulated meta-analysis. Fixed- and random-effects models were used to simulate the meta-analyses. Since between-study heterogeneity is often found in meta-analyses,^{16,17} an understanding of the performance of tests for publication bias in such situations is essential in practice. The between-study heterogeneity parameter was calculated as a percentage of the average within-study variance estimate. From the fixed-effects model, the average within-study variance was calculated and between-study heterogeneity was then

defined to be 20%, 150%, and 500% of the within-study variance estimate. This reflects scenarios ranging from modest to considerable between-study heterogeneity. These levels of between-study heterogeneity corresponds to values of I^2 , the percentage of total variation across studies that is due to heterogeneity rather than chance,¹⁸ of 16.7%, 60%, and 83.3%, respectively.

Performance of the regression tests was assessed in the absence and presence of induced funnel plot asymmetry. Asymmetry was induced in 2 ways. First, it was induced on the basis of the *P* value associated with a study's effect size^{19,20} (the larger the *P* value the more likely that study was excluded from the meta-analysis). Since a study estimate is more likely to be statistically significant when the underlying OR is large, little publication bias is actually induced for the larger underlying ORs. Therefore, publication bias was also induced on the basis of study effect size.²¹ Studies with the most extreme negative effect sizes were excluded from the meta-analysis. Results are based on 1000 replications. The maximum SE of estimates for the type I error rates and power in the simulations is 1.7%. All simulations and analyses were carried out in Stata 8.2.⁷ For ease of presentation, only results for the underlying ORs of 1, 1.5, and 5 are given in the Figures (findings for underlying ORs of 1.2 and 3 follow the same general trend¹³).

RESULTS

An ideal test has the desired type I error rate (eg, 10% when statistical significance is specified from a 2-tailed test at $P < .10$, as is advocated for these tests³) and good power to detect asymmetry when it exists. In FIGURE 2, Egger's regression test exceeds the appropriate type I error rate of 10% for large underlying ORs. As the amount of between-study heterogeneity and number of primary studies increases, the type I error rates also increase, even for moderate ORs (ie, $OR = 1.5$). We also

observed an imbalance in the tail probability areas for Egger's 2-tailed test,¹³ as previously demonstrated.^{4,11}

In the presence of funnel plot asymmetry, Egger's regression test appears reasonably powerful to detect this asymmetry (FIGURE 3), especially as the underlying OR and number of studies in the meta-analysis increase.

However, in assessing practical use of the test, power must be considered in light of the type I error rates (so that false-positive results are not mistaken for true-positive results). This

trade-off between power and type I error rates is similar to that between the sensitivity and specificity of a diagnostic test. Our findings and those of others^{3,4,11} lead us to have serious concerns over the practical use of Egger's test to identify funnel plot asymmetry for lnORs.

Of the 7 further regression models assessed, one model stands out in that its performance is superior to all the others, including Egger's regression test.¹³ This model and the simulated results from it are now discussed.

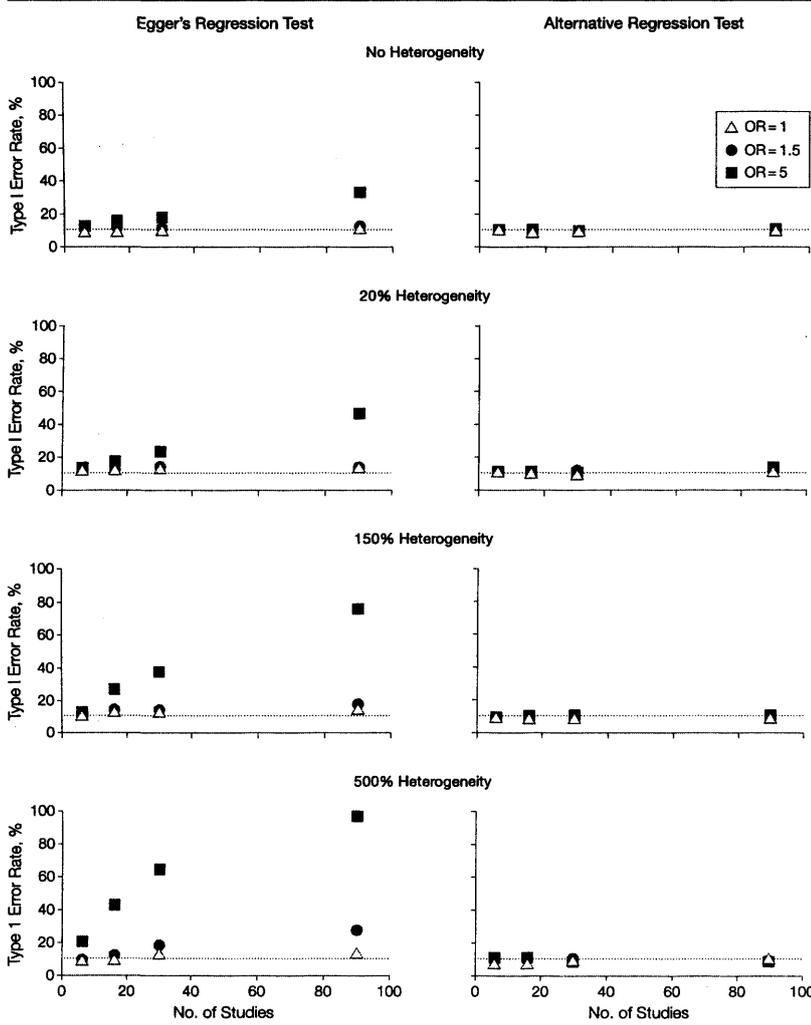
An Alternative to Egger's Regression Test

In preference to Egger's regression test, we recommend a simple weighted linear regression with lnOR as the dependent variable and the inverse of the total sample size as the independent variable. This is a minor modification of Macaskill's test,⁴ with the inverse of the total sample size as the independent variable rather than total sample size. Our results indicate that use of the inverse of the total sample size gives more balanced type I error rates in the tail probability areas than where there is no transformation of sample size.¹³ Use of sample size reduces the correlation between the lnOR and its SE.^{4,13} It also avoids violating an assumption of regression models that Egger's regression test does not avoid, as the independent variable, SE, is subject to random error (so that Egger's regression test is affected by regression dilution bias²²).

The weighting given to each study by the alternative regression test is based on the assumption that the null hypothesis is true, ie, the underlying $OR = 1$. Choice of this weighting helps to reduce the correlation between the lnOR and the weight given to each study when the standard inverse variance weighting is used. Thus, appropriate type I error rates and balance in the tail probabilities are achieved.

Further explanation of the implications of this choice of weighting can be found in Macaskill et al⁴ and details of the weighting given to each study are given in Peters et al.¹³ Figure 2 shows

Figure 2. Type I Error Rates for Egger's Regression Test and the Alternative Regression Test



OR indicates odds ratio. Dotted line indicates the expected type I error rate (ie, 10%).

that the type I error rates for this alternative regression test are approximately 10%, as expected, regardless of the size of the underlying OR, the number of studies in the meta-analysis, and the amount of between-study heterogeneity, unlike those for Egger's regression test (Figure 2).

When there is little between-study heterogeneity, the alternative regression test and Egger's regression test appear to have moderate power to detect asymmetry when it is induced on the basis of *P* value (Figure 3) and high power when asymmetry is induced on the magnitude of the effect (data not shown).

When there is considerable heterogeneity (Figure 3), Egger's regression test is more powerful than the alternative regression test, however as discussed it is difficult to disentangle the high type I error rates of Egger's regression test from power.

COMMENT

Neither Egger's regression test nor the alternative regression test are particularly powerful in all scenarios. However, a test that may not be optimal, but performs well in all situations, is needed. Thus, although the alternative regression test is no more powerful than Egger's regression test, we recommend that the alternative be routinely used rather than Egger's regression test because it reduces the correlation between $\ln OR$ and its SE^4 through the choice of weighting and has appropriate type I error rates. The alternative regression test can easily be run in any software package allowing weighted linear regression. (Details on implementing this test in Stata⁷ are available from the authors.) In fact, applying this test to the meta-analysis illustrated in Figure 1 gives a nonsignificant result ($P = .18$), as one would expect since the data were simulated with no publication bias; Egger's regression test yields $P = .07$.

The alternative regression test is analogous to a funnel plot based on sample size. Thus, although contrary to the recommendations of Sterne and Egger²³

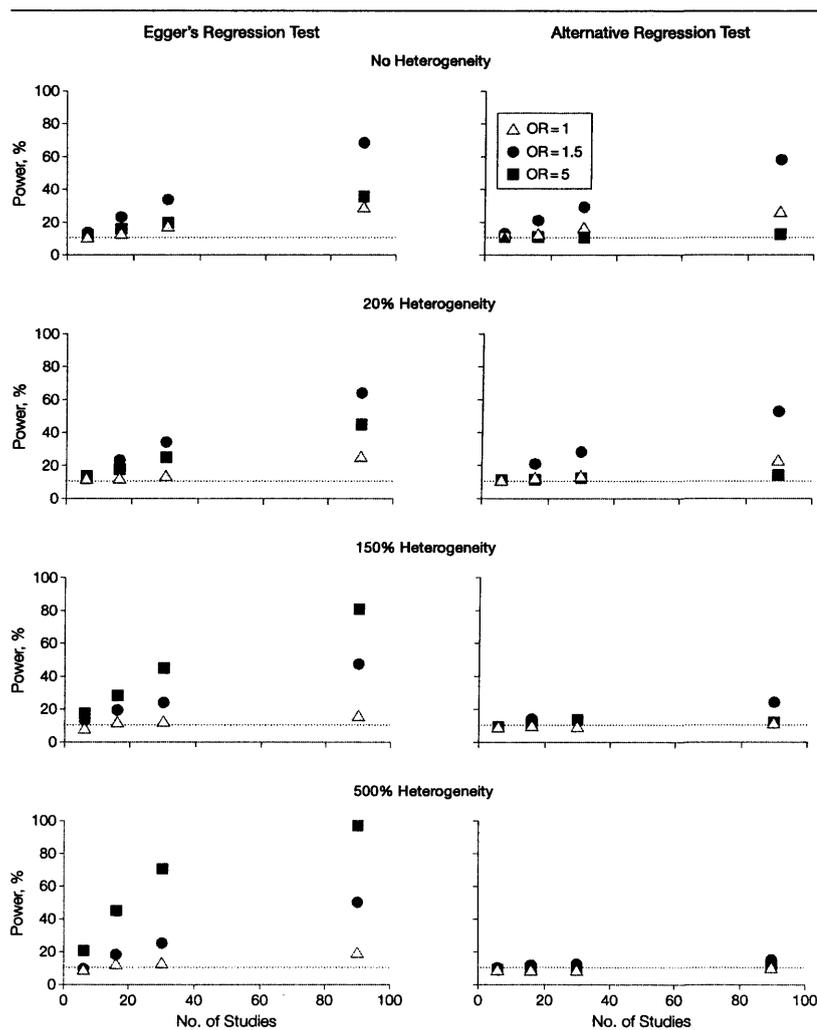
for choice of funnel plot axis, we advocate use of sample size¹³ for $\ln OR$ s.

We have also assessed use of the permutation test to obtain the *P* value for each test. The permutation test has been advocated for use in meta-regression to deal with inflated type I error rates.²⁴ Preliminary findings do not necessarily suggest better performance of tests based on *P* values from the permutation test compared with the usual *t* test.¹³ Extensions to, and performance of, these regression tests when some of the between-study heterogeneity can be

explained by a study-level covariate is ongoing work.

Our results, like those of some others,^{3,4} only concern synthesis of ORs. Findings of an investigation of Egger's regression test using relative risks (RRs) suggests a similar result: excessive type I error rates.¹¹ Although more work is needed on the performance of both tests when the summary estimate is not the OR, it is likely that other relative summary estimates (eg, RRs and risk differences) will be subject to effects similar to the correlation described above for the

Figure 3. Power of Egger's Regression Test and the Alternative Regression Test to Detect Publication Bias Induced by *P* Value



OR indicates odds ratio. Dotted line indicates the expected type I error rate (ie, 10%).

OR, thus suggesting Egger's regression test may not be appropriate. We did not consider meta-analyses of rare events. Evidence suggests the type I error rates for Egger's regression test are particularly high in these situations,^{3,11} but performance of the alternative regression test needs exploring.

Simply testing for the presence of asymmetry does not help obtain an unbiased estimate from the meta-analysis, particularly as there is overreliance on these tests (eg, a nonsignificant *P* value being taken as evidence that publication bias is not an issue). Our review of 43 meta-analyses published in *JAMA* since 1997 found that a number of approaches are taken when publication bias is suspected (as in 11 of the 43 meta-analyses). These include

acknowledging possible publication bias, but giving no detail on the extent or impact of such bias; discussing the possible implications of suspected publication bias and advising caution in the interpretation of the pooled estimate; and attributing inconsistent findings to the possible existence of publication bias. Other possible approaches include the trim and fill method²⁵ and best evidence synthesis approach.^{26,27} None of these approaches is adequate; while better methods of detecting and dealing with publication bias are being developed, we recommend that authors draw their conclusions cautiously, keeping the possibility of sensitivity to publication and related biases in mind.

Author Contributions: Ms Peters had full access to all of the data in the study and takes responsibility for

the integrity of the data and the accuracy of the data analysis.

Study concept and design: Peters, Sutton, Jones, Abrams, Rushton.

Analysis and interpretation of data: Peters, Sutton, Jones, Abrams, Rushton.

Drafting of the manuscript: Peters, Sutton, Jones, Abrams, Rushton.

Critical revision of the manuscript for important intellectual content: Peters, Sutton, Jones, Abrams, Rushton.

Statistical analysis: Peters, Sutton, Jones, Abrams, Rushton.

Obtained funding: Peters, Sutton, Jones, Abrams, Rushton.

Administrative, technical, or material support: Peters, Sutton, Jones, Abrams, Rushton.

Study supervision: Sutton.

Financial Disclosures: None reported.

Funding/Support: Ms Peters is funded through a UK Department of Health Evidence Synthesis Award.

Role of the Sponsor: The funding source had no role in any aspect of the study.

Acknowledgment: We are pleased to thank Petra Macaskill, PhD (School of Public Health, Sydney, Australia), for her comments on an earlier draft and suggestions for its improvement. Dr Macaskill did not receive any compensation.

REFERENCES

- Song F, Eastwood AJ, Gilbody S, Duley L, Sutton AJ. Publication and related biases. *Health Technol Assess*. 2000;4:1-115.
- Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. *Methods for Meta-analysis in Medical Research*. Chichester, England: Wiley; 2000.
- Sterne JAC, Gavaghan D, Egger M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *J Clin Epidemiol*. 2000; 53:1119-1129.
- Macaskill P, Walter SD, Irwig L. A comparison of methods to detect publication bias in meta-analysis. *Stat Med*. 2001;20:641-654.
- Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ*. 1997;315:629-634.
- Web of Knowledge. Available at: <http://wok.mimas.ac.uk>. Accessed September 20, 2005.
- Stata Statistical Software, Release 8.2. College Station, Tex: Stata Corp; 2004.
- Borenstein M, Rothstein H. Comprehensive meta-analysis: a computer program for research synthesis. 1999. Available at: <http://www.meta-analysis.com>. Accessibility verified January 6, 2006.
- Rosenberg MS, Adams DC, Gurevitch J. *Metawin: Statistical Software for Meta-analysis: Version 2.0*. Sunderland, Mass: Sinauer Association; 1999.
- StatsDirect Statistical Software. Available at: <http://www.statsdirect.com>. Accessibility verified January 6, 2006.
- Schwarzer G, Antes G, Schumacher M. Inflation of type I error rate in two statistical tests for the detection of publication bias in meta-analyses with binary outcomes. *Stat Med*. 2002;21:2465-2477.
- Irwig L, Macaskill P, Berry G, Glasziou P. Bias in meta-analysis detected by a simple, graphical test: graphical test is itself biased. *BMJ*. 1998;316:470.
- Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. *Performance of Tests and Adjustments for Publication Bias in the Presence of Heterogeneity: Technical Report 05-01*. Leicester, England: Dept of Health Sciences, University of Leicester; 2005.
- Harbord RM, Egger M, Sterne JA. A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints [published online ahead of print December 12, 2005]. *Stat Med*. doi: 10.1002/sim.2380. Accessed January 18, 2006.
- Peters JL, Sutton AJ, Jones DR, Rushton L, Abrams KA. *A Review of the Use of Systematic Review and Meta-analysis Methods to Evaluate Animal Toxicology Studies: Technical Report 04-02*. Leicester, England: Dept of Health Sciences, University of Leicester; 2004.
- Engels EA, Schmid CH, Terrin N, Olkin I, Lau J. Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. *Stat Med*. 2000;19:1707-1728.
- Villar J, Mackey ME, Carroli G, Donner A. Meta-analyses in systematic reviews of randomized controlled trials in perinatal medicine: comparison of fixed and random effects models. *Stat Med*. 2001;20:3635-3647.
- Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med*. 2002;21:1539-1558.
- Hedges LV, Vevea JL. Estimating effect size under publication bias: small sample properties and robustness of a random effects selection model. *J Educ Behav Stat*. 1996;21:299-332.
- Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. *Biometrics*. 1994;50:1088-1101.
- Duval S, Tweedie RL. A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *J Am Stat Soc*. 2000;95:89-98.
- Irwig L, Glasziou P, Wilson A, Macaskill P. Estimating an individual's true cholesterol level and response to intervention. *JAMA*. 1991;266:1678-1685.
- Sterne JAC, Egger M. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. *J Clin Epidemiol*. 2001;54:1046-1055.
- Higgins JPT, Thompson SG. Controlling the risk of spurious findings from meta-regression. *Stat Med*. 2004;23:1663-1682.
- Duval S, Tweedie R. Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*. 2000;56:455-463.
- Slavin RE. Best-evidence synthesis: an alternative to meta-analytic and traditional reviews. *Educ Res*. 1986;15:5-11.
- Slavin RE. Best evidence synthesis: an intelligent alternative to meta-analysis. *J Clin Epidemiol*. 1995; 48:9-18.

Appendix I: Additional simulation results for Chapter 7

Figure I.1 Power of the rank correlation test to detect 'moderate' publication bias induced by p-value (top row) and effect size (bottom row)

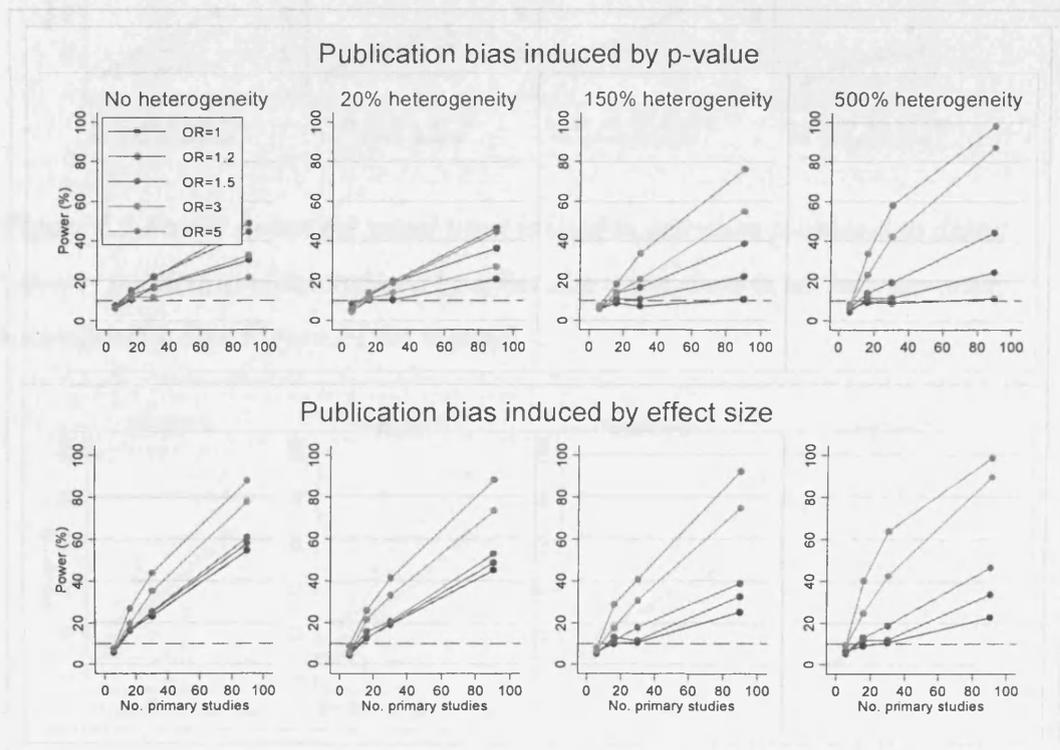


Figure I.2 Power (when the usual *t*-test is used to calculate *p*-values) to detect 'moderate' publication bias induced by *p*-value when there is no between-study heterogeneity (see Figure I.1 for legend)

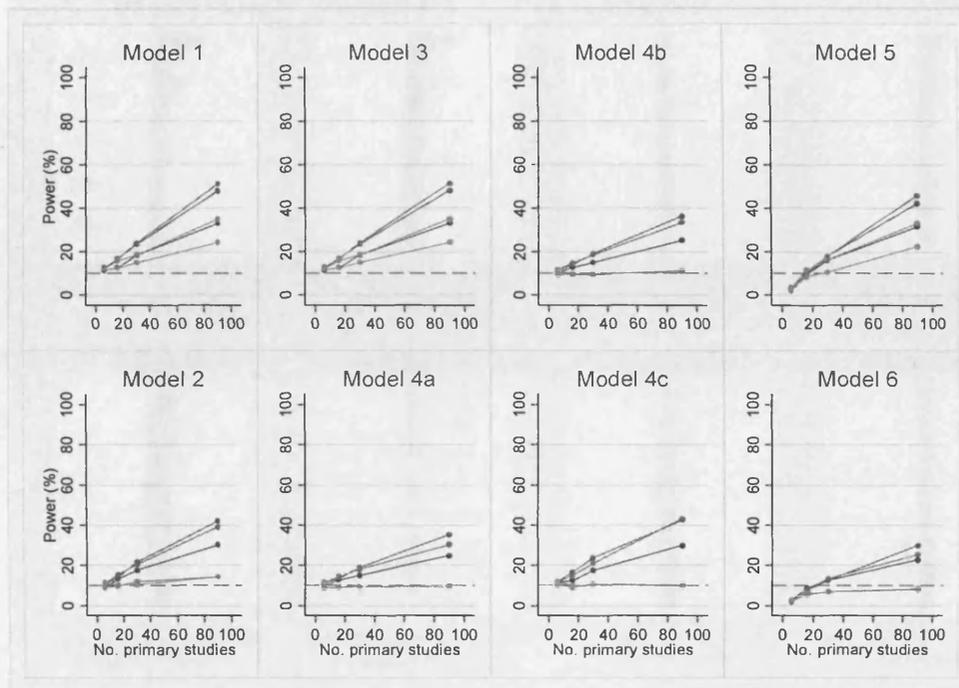


Figure I.3 Power (when the usual *t*-test is used to calculate *p*-values) to detect 'severe' publication bias induced by effect size when there is no between-study heterogeneity (see Figure I.1 for legend)

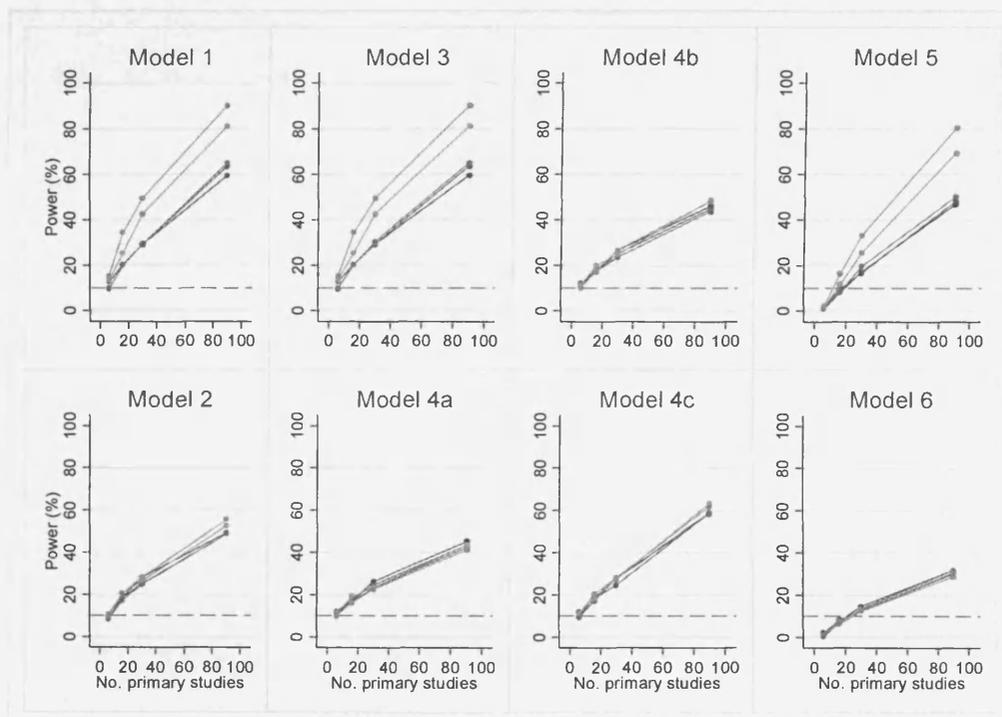


Figure I.4 Type I error rates for eight regression models when the permutation test is used to calculate the p -values and there is no between-study heterogeneity

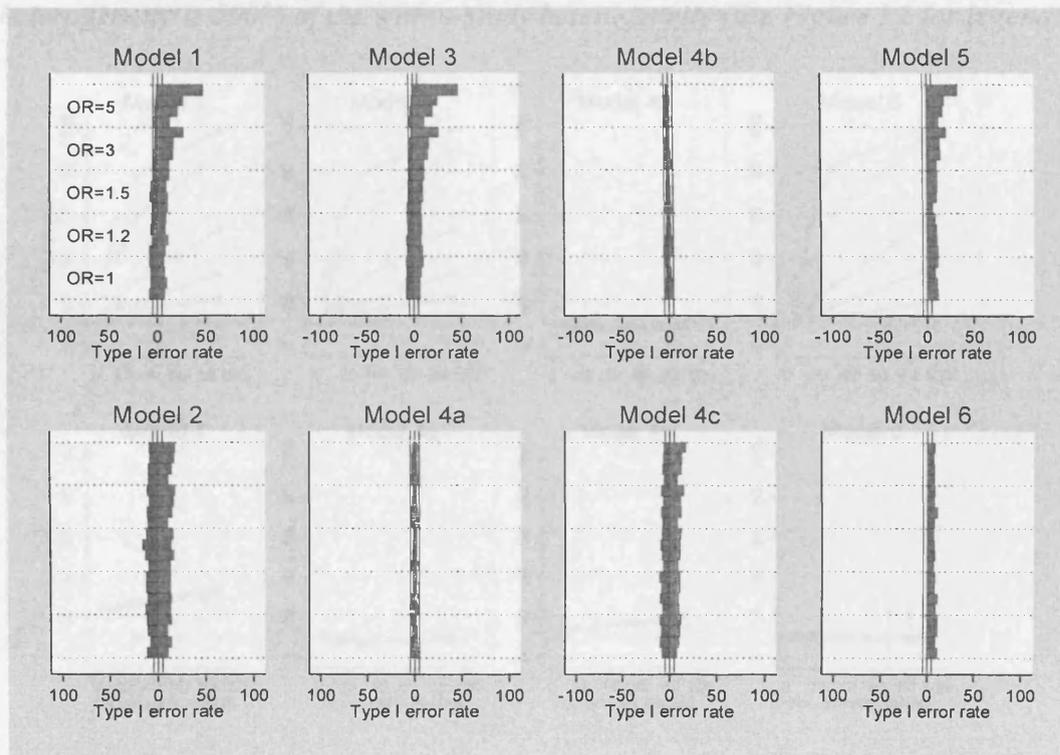


Figure I.4 Type I error rates for eight regression models when the permutation test is used to calculate the p -values and there is no between-study heterogeneity

Figure I.5 Power (when the permutation tests is used to calculate p-values) to detect 'severe' publication bias induced by p-value when between-study heterogeneity is 500% of the within-study heterogeneity (see Figure I.1 for legend)

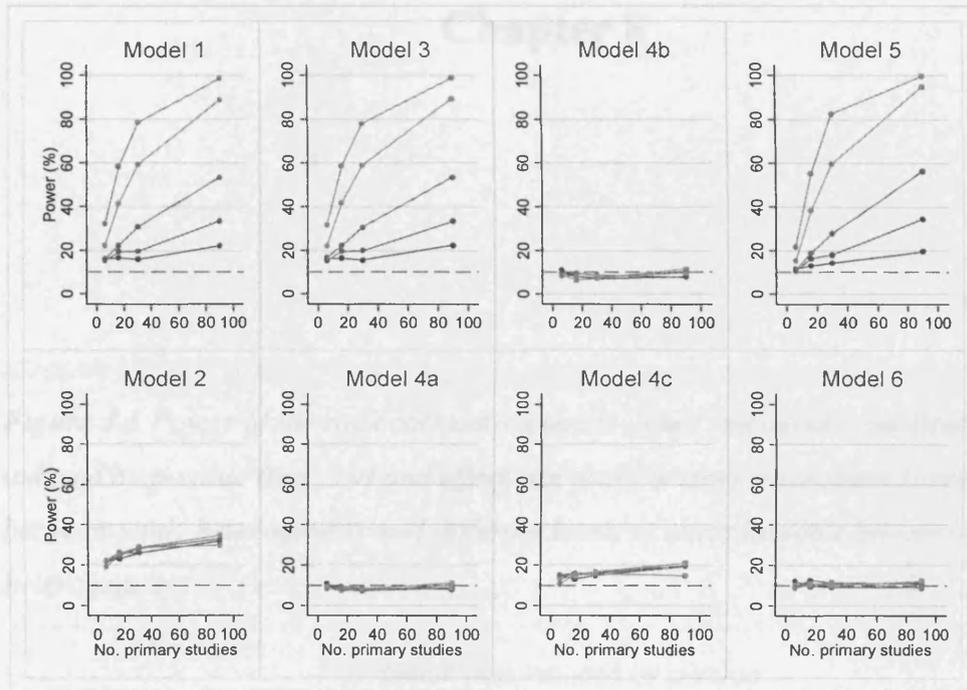
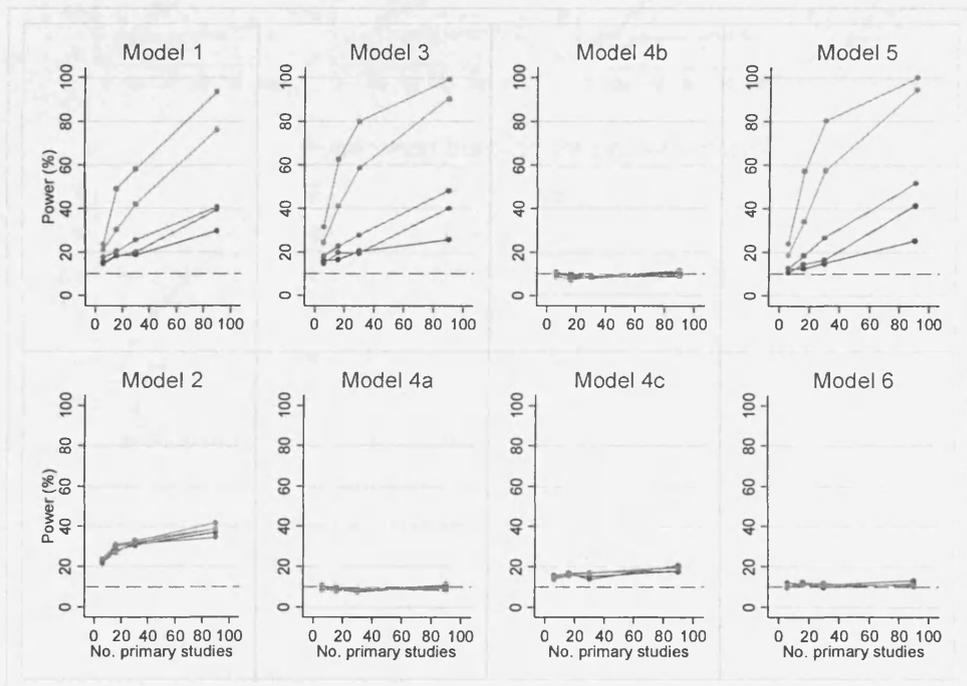


Figure I.5 Power (when the permutation tests is used to calculate p-values) to detect 'severe' publication bias induced by effect size when between-study heterogeneity is 500% of the within-study heterogeneity (see Figure I.1 for legend)



Appendix J: Additional simulation results for Chapter 8

Figure J.1 Power of the rank correlation test to detect 'moderate' publication bias induced by p-value (top row) and effect size (bottom row) when there is explainable between-study heterogeneity and differing levels of unexplainable between-study heterogeneity

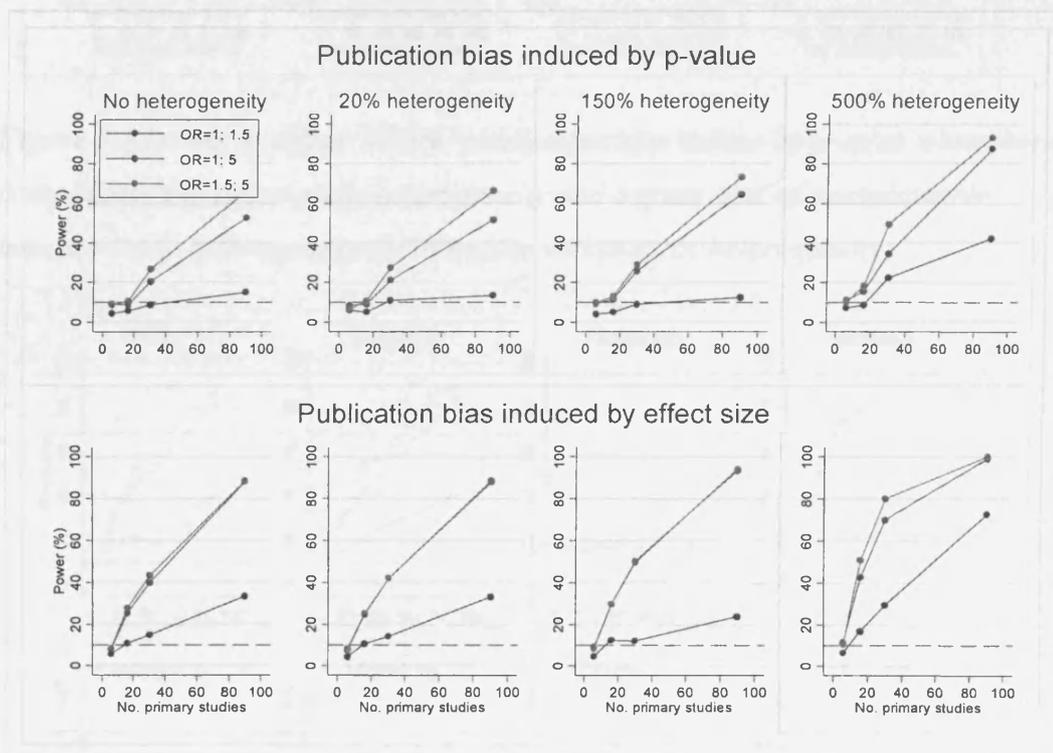


Figure J.2 Type I error rates when there is unexplainable between-study heterogeneity, and a great deal of unexplained between-study heterogeneity (500% of within study heterogeneity)

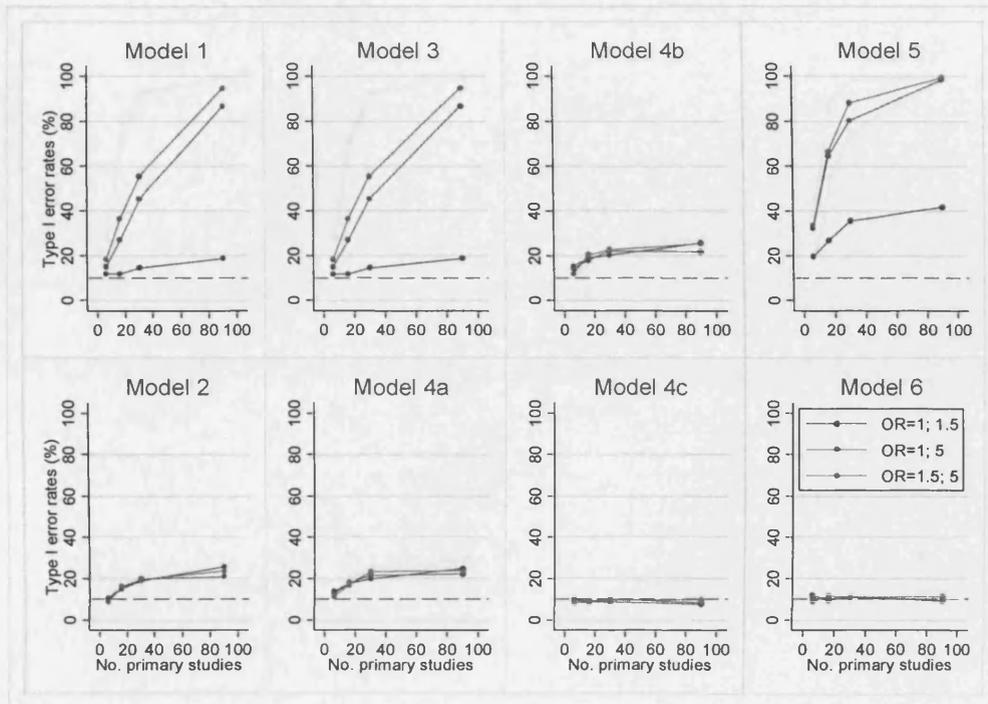


Figure J.3 Power to detect 'severe' publication bias induce by p-value when there is explainable between-study heterogeneity and a great deal of unexplainable between-study heterogeneity (500% of the within-study heterogeneity)

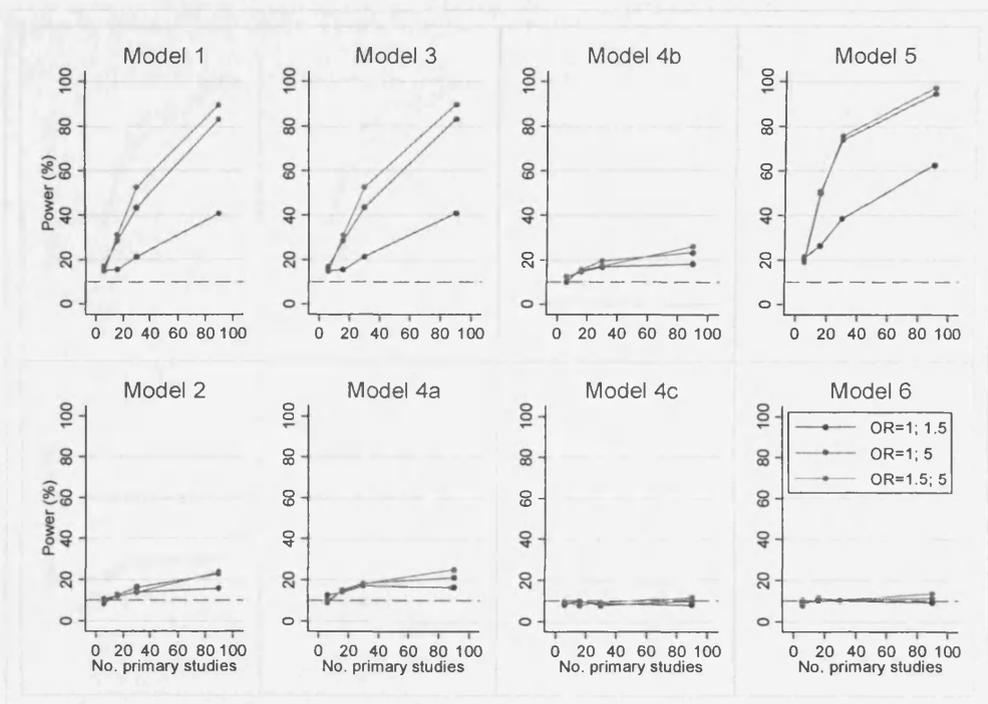


Figure J.4 Power to detect 'severe' publication bias induced by effect size when there is explainable between-study heterogeneity and a great deal of unexplainable between-study heterogeneity (500% of the within-study heterogeneity)

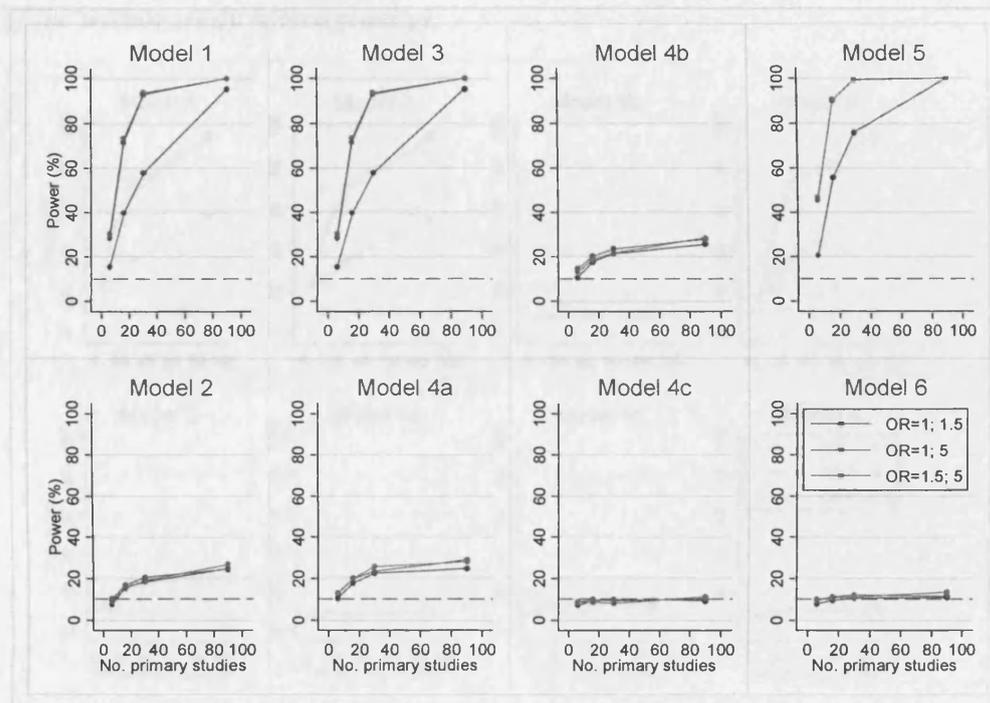


Figure J.5 Type I error rates (where the permutation tests is used to calculate p -values) when there is explainable between-study heterogeneity and a great deal of unexplained between-study heterogeneity (500% of the within-study heterogeneity)

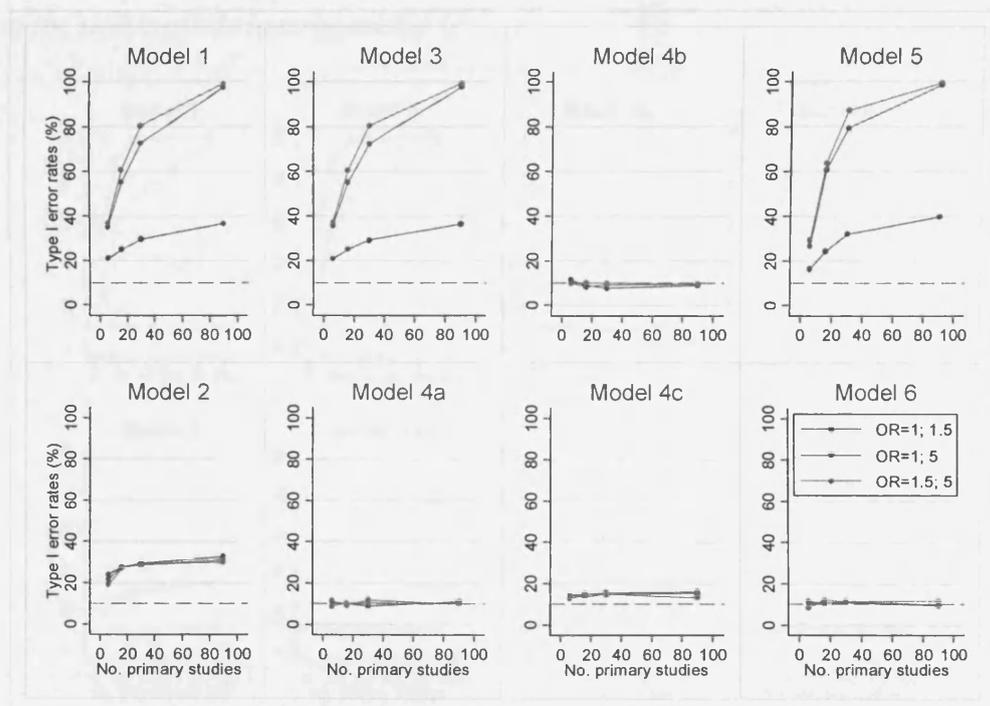


Figure J.6 Power (*p*-values calculated by permutation test) to detect 'severe' publication bias induced by *p*-value when there is explainable between-study heterogeneity and a great deal of unexplained between-study heterogeneity (500% of the within-study heterogeneity)

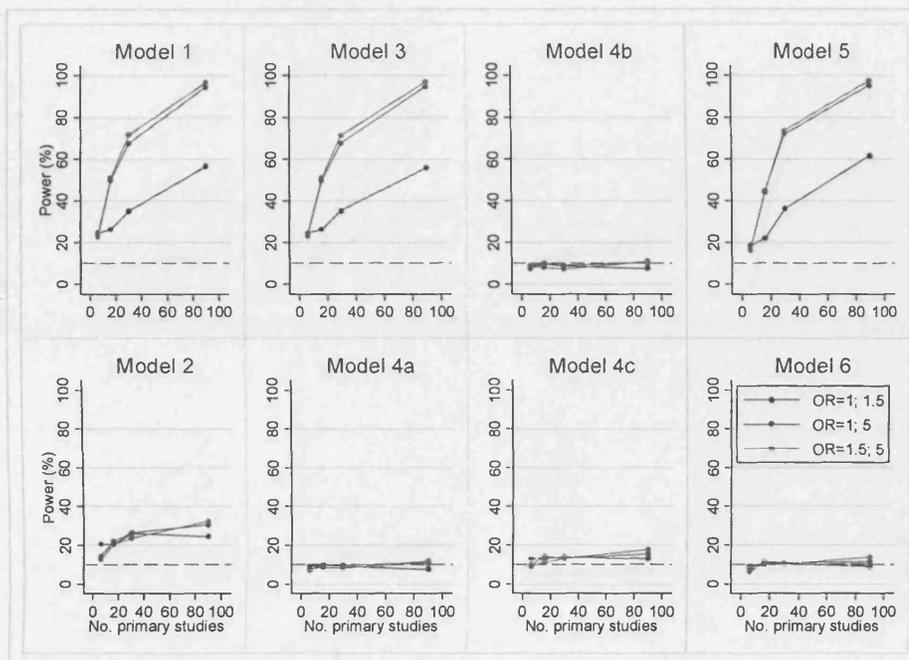


Figure J.7 Power (*p*-values calculated by permutation test) to detect 'severe' publication bias induced by effect size when there is explainable between-study heterogeneity and a great deal of unexplained between-study heterogeneity (500% of the within-study heterogeneity)

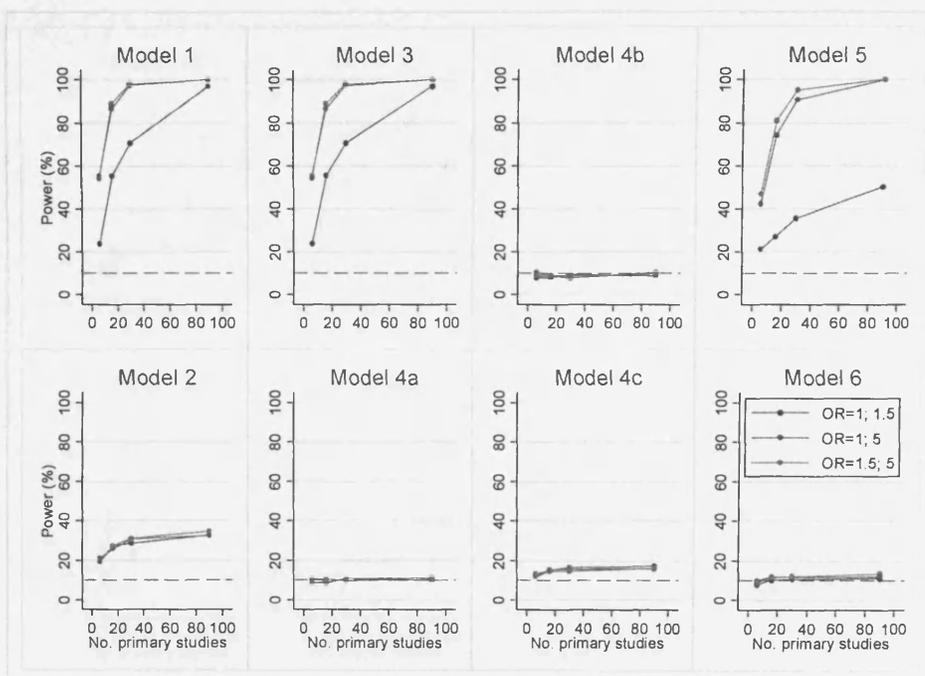


Figure J.8 Power (p -values calculated by the usual t -test) of extended regression tests to detect 'severe' publication bias induced by p -value when there is explainable between-study heterogeneity and a great deal of unexplainable between-study heterogeneity (500% of the within study heterogeneity)

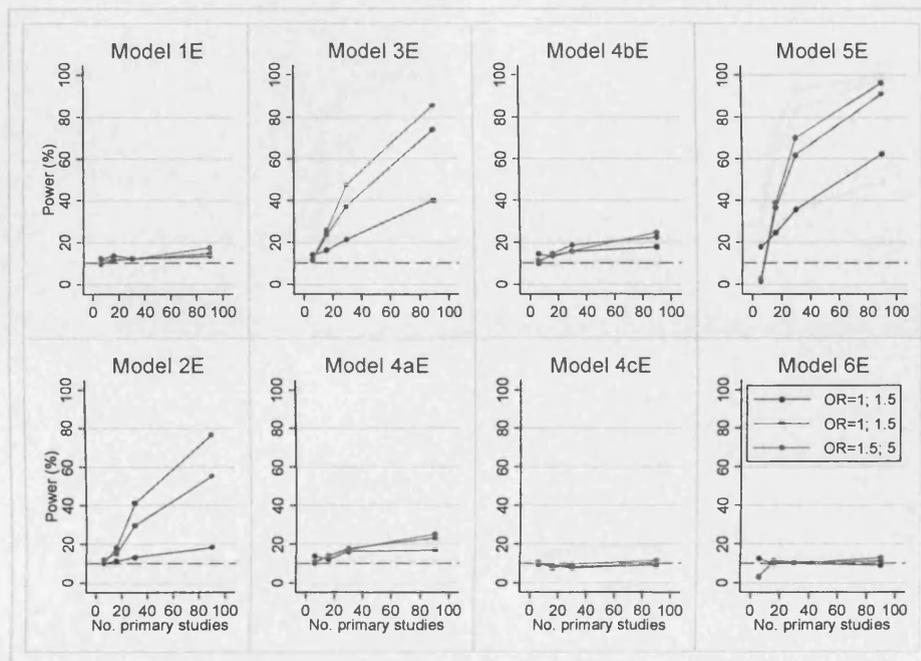


Figure J.9 Power (p -values calculated by the usual t -test) of extended regression tests to detect 'severe' publication bias induced by effect size when there is explainable between-study heterogeneity and a great deal of unexplainable between-study heterogeneity (500% of the within study heterogeneity)

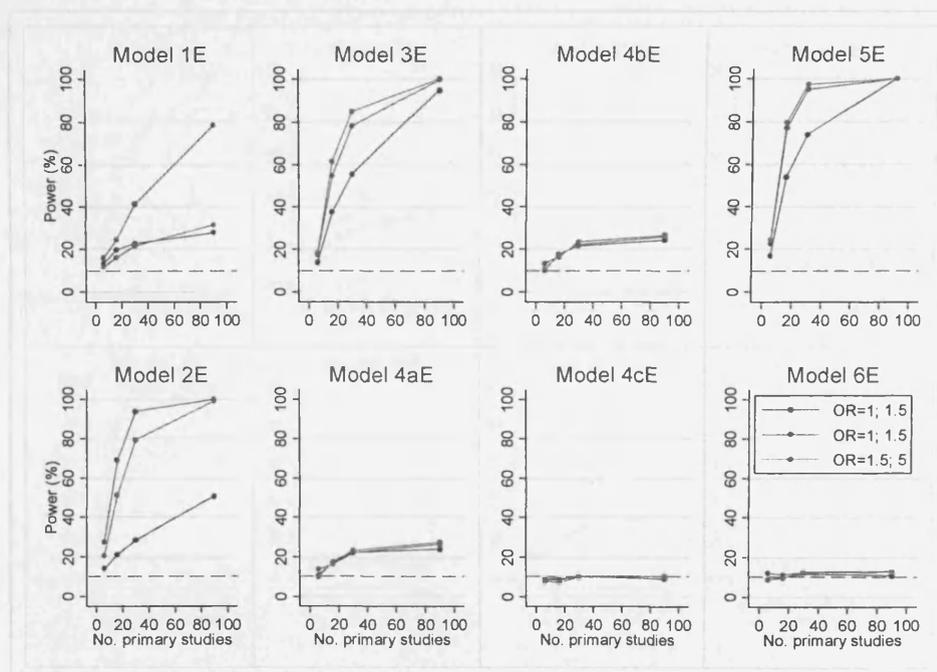


Figure J.10 Power (*p*-values calculated by the permutation test) of extended regression tests to detect 'severe' publication bias induced by effect size when there is explainable between-study heterogeneity and a great deal of unexplainable between-study heterogeneity (500% of the within study heterogeneity)

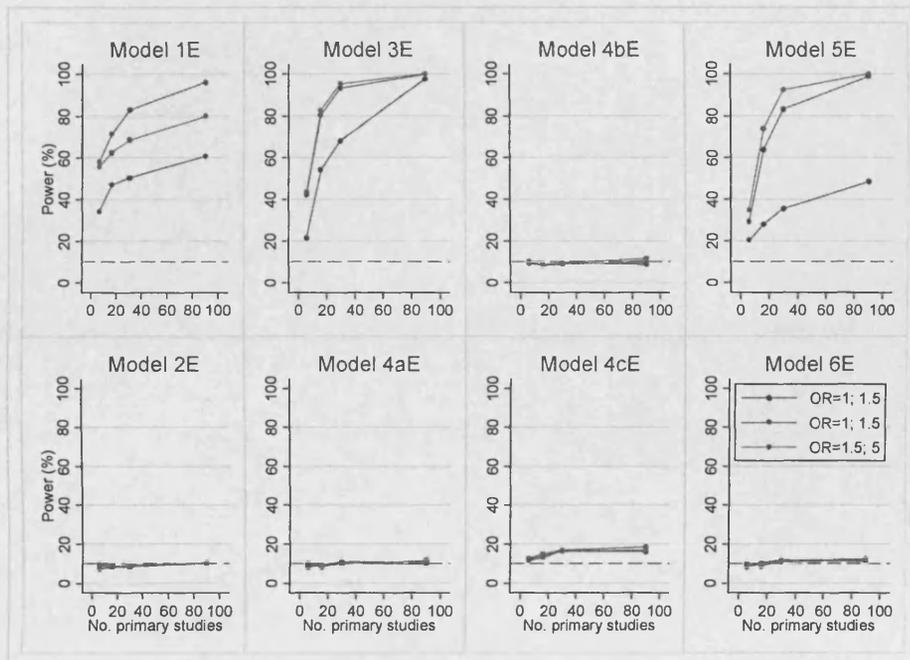


Figure J.11 Power (*p*-values calculated by the permutation test) of extended regression tests to detect 'severe' publication bias induced by *p*-value when there is explainable between-study heterogeneity and no unexplainable between-study heterogeneity

