EVIDENCE SYNTHESIS OF
PROGNOSTIC MARKER STUDIES

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

by

Richard David Riley BSc. MSc.

Department of Health Sciences
University of Leicester

November 2004
EVIDENCE SYNTHESIS OF PROGNOSTIC MARKER STUDIES

Richard David Riley B.Sc. M.Sc.

Abstract

Prognostic markers are important clinical tools because they help to identify patients with different risks of outcome (e.g. death) and thereby facilitate the most appropriate treatment strategies. Evidence-based results signifying the most relevant markers for clinical practice are therefore highly desirable, especially as (sometimes conflicting) results regarding markers are commonly published across a large number of studies.

In this thesis I identify and demonstrate the methodological difficulties of obtaining evidence-based prognostic marker results by performing a large-scale systematic review in neuroblastoma involving 13 different markers across 260 primary studies. The reporting in these studies was often inadequate, with considerable heterogeneity in many important clinical and statistical factors. These problems restricted the extraction and subsequent meta-analysis of the results from the primary studies, thus severely limiting the formation of clinically relevant evidence-based conclusions.

In response to this, guidelines for the reporting of primary prognostic marker studies are developed in this thesis; in particular, improved reporting of both summary statistics and clinical information, combined with the availability of individual patient data, are all recommended to facilitate future evidence-based results. To complement this initiative, I also introduce, develop and illustrate bivariate meta-analysis methods that utilise the correlation between overall and disease-free survival statistics to limit the problems of unreported outcomes and dissemination bias, all of which hinder evidence synthesis of prognostic markers. Where missing data can be assumed ‘missing (completely) at random’, bivariate random-effects models are shown capable of more reliable evidence-based results than a standard univariate meta-analysis. Furthermore, when data is ‘not missing at random’, I propose how the bivariate framework can be used alongside assessments of funnel plot asymmetry in order to evaluate the robustness of evidence-based conclusions to the potential impact of dissemination bias.

This thesis therefore makes a positive contribution toward more reliable and clinically relevant evidence-based results for prognostic markers in the future. However, numerous methodological issues remain and so further research priorities are outlined to help ensure the move toward an evidence-based use of prognostic markers continues.
ACKNOWLEDGMENTS

This thesis was made possible by funding from the NHS Health Technology Assessment (HTA) programme [Project no. 97/15/03] and the Department of Health, who awarded me a Training Fellowship in Health Services Research to complete this work.

Firstly, I would like to sincerely thank Keith Abrams for providing me with the opportunity to undertake this research and for his supervision, encouragement and continual guidance during the whole thesis. I am particularly grateful for how he has helped develop this career path for me, and for the large amount of time he has spent reading various drafts and providing constructive feedback over the last few months. I also want to greatly thank Paul Lambert and Alex Sutton, who have similarly helped guide and advise, and have always been willing to give-up their valuable time to listen and encourage. Furthermore, I thank David Jones and John Thompson who provided helpful direction and motivation at some important stages of the thesis. I am also especially thankful to Sue Burchill and David Heney for providing invaluable clinical input throughout, and to Doug Altman for his encouragement and feedback about various aspects of my work. Furthermore, I thank Martin Perkins, Lucy Smith and the numerous others at Princess Road West who have continually encouraged me.

My love and thanks naturally also go to my family for their support and encouragement throughout the duration of the thesis. Yes, this is what I have been doing all this time! In particular, I would like to say a big thanks to Mum, Dad, Pete (and the PES series), Sian, Katie, Ellie, Joe, Gran, Grandad, Paul and Wendy. You have all helped see me through, as have Tim, Barry, and other friends.

Now to Lorna – where do I begin? This thesis would not have been possible without you. Love and thanks in abundance, for all you are and all you do for me every day. I am very blessed to have you. Finally, I thank God for giving me strength, purpose and overseeing every area of my life - all the best bits of the thesis are yours.
## CONTENTS

### CHAPTER 1: AIMS AND BACKGROUND FOR THE THESIS

1. **Aims of the thesis**  
2. **Prognostic markers and prognostic marker studies**  
3. **Prognostic markers in the clinical management of cancer patients**
   - 1.3.1 *What is cancer?*
   - 1.3.2 *Limiting the impact of cancer using prognostic markers*
   - 1.3.3 *The example of prognostic tumour markers in neuroblastoma*
4. **Multiple studies, multiple markers and inconsistent evidence**  
5. **The growth of formal evidence synthesis in the medical literature**  
6. **Systematic reviews and the threat of dissemination bias**  
7. **Methods for meta-analysis in medical research**
   - 1.7.1 *Fixed-effects meta-analysis*
   - 1.7.2 *Random-effects meta-analysis*
   - 1.7.3 *Meta-regression for explaining the between-study heterogeneity*
   - 1.7.4 *An example: a meta-analysis to assess lactate dehydrogenase as a possible prognostic marker in patients with Ewing's Sarcoma*
8. **Bayesian approaches to meta-analysis**
   - 1.8.1 *Bayesian statistics*
   - 1.8.2 *Markov Chain Monte Carlo methods*
   - 1.8.3 *A Bayesian random-effects meta-analysis model to assess LDH*
   - 1.8.4 *Benefits and limitations of a Bayesian approach to meta-analysis*
9. **Rationale and structure for the thesis**

### CHAPTER 2: A FEASIBILITY STUDY OF A SYSTEMATIC REVIEW AND META-ANALYSIS OF PROGNOSTIC MARKER STUDIES

1. **Rationale and methods for a systematic review of markers in neuroblastoma**
   - 2.1.1 *Neuroblastoma*
   - 2.1.2 *Search strategy*
   - 2.1.3 *Inclusion criteria*
   - 2.1.4 *Exclusion criteria*
2.1.5 Information extracted 42
2.1.6 Meta-analysis of the prognostic marker studies 42

2.2 Results from the literature search and data extraction 47
2.2.1 Literature search results 47
2.2.2 Data extraction of prognostic marker results 47

2.3 Problems limiting meta-analysis 48
2.3.1 Poor reporting of primary studies 48
2.3.2 Heterogeneity of clinical and statistical factors 51
2.3.3 Publication bias and other dissemination bias issues 53

2.4 Should I proceed with meta-analysis? 54
2.4.1 Limited evidence-based results are possible 54
2.4.2 Meta-analysis results 55
2.4.3 Clinical interpretation of the meta-analysis results 60

2.5 Discussion and generalisations from the empirical investigation 61
2.5.1 Appraisal of the systematic review and data extraction 61
2.5.2 Generalisations of the problems to other prognostic markers 63
2.5.3 The need to improve the quality and reporting of primary studies 65

2.6 Suggestions for improving the quality of prognostic marker studies 66
2.6.1 Study quality issue 1: the purpose of prognostic marker studies 66
2.6.2 Study quality issue 2: the type of prognostic marker studies 68
2.6.3 Summary of improving study quality in prognostic marker studies 70

2.7 Clinical requirements from a prognostic marker study 71

2.8 Why do reporting problems exist in prognostic marker studies? 75
2.8.1 Poorly designed and poorly targeted studies 76
2.8.2 Lack of collaborative research 76
2.8.3 Biased reporting of results 76
2.8.4 Lack of appreciation of evidence-based research 77
2.8.5 Lack of statistical knowledge 78

2.9 Towards guidelines for improved reporting of primary prognostic marker studies to facilitate future evidence-based reviews 79
2.9.1 IPD and evidence synthesis of prognostic marker studies 83
2.9.2 IPD and the problem of heterogeneous cut-off levels 86
2.9.3 IPD and the problem of heterogeneous adjustment factors 87
2.9.4 IPD and the evidence synthesis of sub-populations 88
2.9.5 Limitations of an IPD approach to evidence synthesis 89
2.9.6 Implementation of the guidelines and recommendations in practice 91
2.9.7 Prospectively planned pooled analyses – the ultimate solution? 92

2.10 Summary and rationale for subsequent chapters 93

CHAPTER 3: BIVARIATE RANDOM-EFFECTS META-ANALYSIS AND ITS POTENTIAL ROLE IN THE EVIDENCE SYNTHESIS OF PROGNOSTIC MARKER STUDIES

3.1 Rationale for developing bivariate meta-analysis methods in this thesis 94
3.2 Missing and correlated OS and DFS estimates for marker MYCN 97
3.3 Multiple summary statistics and multivariate meta-analysis models 101
3.4 A hierarchy for an IPD bivariate random-effects meta-analysis 103
   3.4.1 The hierarchy for an IPD univariate random-effects meta-analysis 103
   3.4.2 The hierarchy for an IPD bivariate random-effects meta-analysis 106
3.5 Applications of an IPD bivariate meta-analysis in practice 109
3.6 Bivariate random-effects meta-analysis using summary statistics: a general framework and literature review
   3.6.1 A general bivariate random-effects meta-analysis model 112
   3.6.2 The problem of unknown within-study correlation values 113
   3.6.3 A literature review of multivariate meta-analysis in practice 114
3.7 Summary and rationale for subsequent chapters 120

CHAPTER 4: AN ANALYTICAL ASSESSMENT OF MODEL A, WITH CRITICAL COMPARISON TO UNIVARIATE RANDOM-EFFECTS META-ANALYSIS

4.1 Estimation using restrictive iterative generalised least squares 124
4.2 Fitting Model A in SAS Proc Mixed using Cholesky Decomposition 126
4.3 Analytic solutions for the unknown parameters in Model A 128
4.4 Weighting of each study in the analytic solutions from Model A 130
4.5 Interpretation of the analytic solutions for the pooled estimates 131
4.6 Empirical Bayes or shrunken study estimates from Model A 134
4.7 Application of Model A to the Berkey dataset
   4.7.1 Results for the original Berkey dataset 136
   4.7.2 Results for the modified Berkey dataset 138
   4.7.3 Results for \( \tilde{\beta}_1 - \tilde{\beta}_2 \) 139

Contents
4.7.4 Results for the empirical Bayes estimates

4.8 Further comparison of the parameter estimates from Model A to those from a univariate random-effects meta-analysis

4.8.1 Remark 1: The precision of $\tilde{\beta}_j$ is smallest in Model A on average

4.8.2 Remark 2: The mean-square error of $\tilde{\beta}_j$ is smallest in Model A on average

4.8.3 Remark 3: Remarks 1 and 2 may be true for $\tau_j^2$ in Model A

4.8.4 Remark 4: Situations when Model A is equivalent to two independent URMAs

4.8.5 Remark 5: The potential benefits of Model A given missing data

4.8.6 Remark 6: Estimation of $\text{var}(\hat{\beta}_1 - \hat{\beta}_2)$ using Model A compared to a URM

4.9 Summary and rationale for subsequent chapters

CHAPTER 5: A SIMULATION STUDY TO ASSESS THE BENEFITS AND LIMITATIONS OF MODEL A FOR COMPLETE-CASE DATA

5.1 Previous simulation studies of bivariate meta-analysis models

5.2 Simulations of Model A for complete-case data from $n = 50$ studies

5.2.1 The simulation procedure for $n = 50$ studies

5.2.2 Degrees of freedom for the coverage of the pooled estimates

5.2.3 Simulation results and assessment of URMA and Model A for $n = 50$

5.3 Simulations of Model A for complete-case data from $n = 5$ studies

5.3.1 The simulation procedure for $n = 5$ studies

5.3.2 URMA simulation results and assessment for $n = 5$ studies

5.3.3 Model A simulation results and assessment for $n = 5$ studies

5.4 Simulations of complete-case data from $n = 5$ and $n = 50$ studies for low and high values of $\tau_1^2$ and $\tau_2^2$

5.4.1 The simulation procedure for investigating various ratios of $\tau_j^2$ to $s_j^2$

5.4.2 Simulation results for low $\tau_1^2$ and low $\tau_2^2$ in relation to the $s_j^2$'s

5.4.3 Simulation results for low $\tau_1^2$ and high $\tau_2^2$ in relation to the $s_j^2$'s

5.4.4 Simulation results for high $\tau_1^2$ and high $\tau_2^2$ in relation to the $s_j^2$'s

5.5 Main conclusions from the complete-case data simulations of Model A

5.6 A critical assessment of the reasons for and the impact of a between-study correlation estimate of 1 or -1

5.6.1 Reason (I) why $\rho_B$ is often estimated as 1 or -1

Contents
5.6.2 Reason (II) why $\rho_B$ is often estimated as 1 or -1
5.6.3 The impact of Reasons (I) and (II) on the other parameter estimates
5.6.4 The dangers of making evidence-based conclusions from Model A when $\rho_B$ is estimated as 1 or -1

5.7 Summary and rationale for subsequent chapters

CHAPTER 6: A CRITICAL ASSESSMENT OF MODEL A FOR MISSING DATA AND OTHER EXTENSIONS

6.1 Using Model A given missing summary statistics across studies
6.2 Types of missing data
6.3 Using Model A when the type of missing data is unknown
   6.3.1 Are the missing OS and DFS summary statistics missing at random?
   6.3.2 Using Model A as a sensitivity analysis
6.4 A simulation study to assess Model A for missing data
   6.4.1 Data is missing completely at random for outcome $j = 2$
   6.4.2 Data is missing at random for outcome $j = 1$
6.5 Main conclusions about using Model A to estimate $\beta_j$ given missing data
6.6 The clear benefit of Model A for estimating $(\beta_1 - \beta_2)$ and $\text{var}(\tilde{\beta}_1 - \tilde{\beta}_2)$ for both complete-case data and missing data
6.7 Extensions of Model A to other situations
   6.7.1 Negative correlation
   6.7.2 Relation of findings to bivariate fixed-effects meta-analysis
   6.7.3 Trivariate and other higher order multivariate meta-analyses
   6.7.4 Extension of Model A to bivariate random-effects meta-regression
6.8 Summary and rationale for subsequent chapters

CHAPTER 7: AN ASSESSMENT OF THE BENEFITS AND LIMITATIONS OF A BAYESIAN APPROACH TO MODEL A

7.1 Including all parameter uncertainty and prior knowledge about $\rho_B$
7.2 Formulating the Bayesian approach to Model A in WinBUGS
7.3 The influence of 'vague' prior distributions in Model A-Bayes
   7.3.1 Application of Model A-Bayes to the Berkey dataset
   7.3.2 Application of Model A-Bayes to the modified Berkey dataset
   7.3.3 The dangers of using 'vague' prior distributions in Model A-Bayes

Contents
7.4 The tension between accounting for uncertainty and the influence of prior distributions

7.5 The benefit of incorporating external information in Model A-Bayes
   7.5.1 Results for the complete-case dataset
   7.5.2 Results for the missing completely at random dataset
   7.5.3 Results for the missing at random dataset
   7.5.4 The difficulty of eliciting strong external information

7.6 Summary and rationale for subsequent chapters

CHAPTER 8: BIVARIATE RANDOM-EFFECTS META-ANALYSIS WHEN THE WITHIN-STUDY CORRELATIONS ARE UNKNOWN

8.1 Options suggested by the current literature for applying Model A or Model A-Bayes when the within-study correlations are unknown
   8.1.1 Option 1: apply Model A assuming $\rho_w = 0$ for all studies
   8.1.2 Option 2: sensitivity analysis using a range of $\rho_w$ values
   8.1.3 Option 3: Utilising prior knowledge to narrow the range of possible $\rho_w$ values
   8.1.4 Option 4: Placing a prior distribution on $\rho_w$ within Model A-Bayes
   8.1.5 Summary of current approaches for dealing with unknown $\rho_w$ values

8.2 Introducing an alternative model for bivariate random-effects meta-analysis
   8.2.1 Modelling directly the overall correlation between the $\tilde{Y}_i$,s and $\tilde{Y}_{i2}$
   8.2.2 RIGLS analytic solutions for Model B, with comparison to Model A
   8.2.3 Fitting Model B in STATA, with application to the Berkey data

8.3 A simulation study to assess the benefits and limitations of Model B
   8.3.1 Complete-case data simulation results for $n = 50$ studies in relation to $\hat{\beta}$
   8.3.2 Complete-case data results for $n = 5$ studies in relation to $\hat{\beta}$
   8.3.3 Complete-case data results for $n = 5$ and $n = 50$ studies in relation to $(\hat{\beta}_1 - \hat{\beta}_2)$
   8.3.4 Missing completely at random simulation results for $\hat{\beta}_j$ and $(\hat{\beta}_1 - \hat{\beta}_2)$
   8.3.5 Missing at random simulation results for $\hat{\beta}_j$ and $(\hat{\beta}_1 - \hat{\beta}_2)$
   8.3.6 Some negative correlation simulation results for Model B

8.4 Discussion of the major benefits and limitations of Model B
   8.4.1 The major benefits of Model B identified by the simulations

Contents
8.4.2 The major limitations of Model B identified by the simulations 292
8.4.3 Discussion on the choice of -0.95 to 0.95 as a 'safe' range for $\hat{\rho}$ 293

8.5 Application of Model B to datasets from the neuroblastoma prognostic marker review and other clinically relevant situations

8.5.1 Joint modelling of OS and DFS hazard ratios 295
8.5.2 Joint modelling of two correlated prognostic markers 299
8.5.3 Joint modelling of sensitivity and specificity 300
8.5.4 Joint modelling of the log-odds for a treatment and control group 301
8.5.5 Joint modelling of two log-odds ratios 303
8.5.6 Issues arising from applying Model B in the five examples 304

8.6 A Bayesian approach to Model B 305

8.7 Further discussion of Model B in relation to Model A, with appraisal of my research and suggestions for additional work

8.7.1 Benefits of my research 308
8.7.2 Using Model A and Model B in practice, with suggestions for further research 309
8.7.3 The use of Model A and Model B as a sensitivity analysis 315

8.8 Summary and rationale for subsequent chapters 316

CHAPTER 9: USING THE BIVARIATE META-ANALYSIS FRAMEWORK TO HELP ASSESS DISSEMINATION BIAS IN META-ANALYSIS OF PROGNOSTIC MARKER STUDIES

9.1 Missing summary statistics and the threat of dissemination bias 318
9.2 Assessing the potential impact of dissemination bias on the results of a univariate random-effects meta-analysis

9.2.1 Visual inspection of the funnel plots 322
9.2.2 Egger's test for the presence of dissemination bias 323
9.2.3 The Trim and Fill method 324
9.2.4 Other possible causes of funnel plot asymmetry beside dissemination bias 325

9.3 A method for assessing the potential impact of dissemination bias on the results of a bivariate random effects meta-analysis

9.3.1 The 'three-step approach' 328
9.3.2 Critical discussion of the 'three-step approach' 336
9.4 The use of selection models to assess dissemination bias
9.4.1 Selection models
9.4.2 Two examples of the use of selection models in meta-analysis
9.4.3 The use of selection models when two correlated summary statistics are needed for a meta-analysis
9.4.4 Two common obstacles for using selection models in practice

9.5 Summary and final thoughts on sensitivity analyses in meta-analysis

CHAPTER 10: DISCUSSION AND SUGGESTIONS FOR FURTHER RESEARCH
10.1 Benefits of the neuroblastoma review and the reporting guidelines developed
10.2 Benefits of the research into bivariate random-effects meta-analysis
10.3 Benefits of the research into how the bivariate framework can be used to help assess dissemination bias
10.4 Where next for evidence synthesis of prognostic markers? Identification of the future research priorities toward clinically relevant evidence-based results

10.4.1 Further development of sensitivity methods in meta-analysis
10.4.2 Formal demonstration of the benefits of IPD over summary statistics
10.4.3 Development of methods to weigh benefit against cost for an evidence synthesis using IPD rather than summary statistics
10.4.4 Development of meta-analysis methods to combine IPD and summary statistics
10.4.5 Development of methods to weigh benefit against cost for an evidence synthesis rather than one new, high quality primary study
10.4.6 Initiation of prospectively planned pooled analyses

10.5 A possible strategy for future prognostic marker research
10.6 Conclusion

APPENDIX A
APPENDIX B
APPENDIX C
APPENDIX D

BIBLIOGRAPHY
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRMA</td>
<td>Bivariate random-effects meta-analysis</td>
</tr>
<tr>
<td>Ch17q</td>
<td>Chromosome 17q</td>
</tr>
<tr>
<td>Ch1p</td>
<td>Chromosome 1p</td>
</tr>
<tr>
<td>CI</td>
<td>Confidence interval</td>
</tr>
<tr>
<td>DFS</td>
<td>Disease-free survival</td>
</tr>
<tr>
<td>GLS</td>
<td>Generalised least-squares</td>
</tr>
<tr>
<td>HR</td>
<td>Hazard ratio</td>
</tr>
<tr>
<td>HVA</td>
<td>Homovanillic acid</td>
</tr>
<tr>
<td>IGLS</td>
<td>Iterative generalised least-squares</td>
</tr>
<tr>
<td>IPD</td>
<td>Individual patient data</td>
</tr>
<tr>
<td>LDH</td>
<td>Lactate dehydrogenase</td>
</tr>
<tr>
<td>MAR</td>
<td>Missing at random</td>
</tr>
<tr>
<td>MCAR</td>
<td>Missing completely at random</td>
</tr>
<tr>
<td>MDR</td>
<td>Multi-drug resistance</td>
</tr>
<tr>
<td>MSE</td>
<td>Mean-square error</td>
</tr>
<tr>
<td>NMAR</td>
<td>Not missing at random</td>
</tr>
<tr>
<td>NSE</td>
<td>Neuron-specific enolase</td>
</tr>
<tr>
<td>OS</td>
<td>Overall survival</td>
</tr>
<tr>
<td>RCT</td>
<td>Randomised controlled trial</td>
</tr>
<tr>
<td>REML</td>
<td>Restricted maximum likelihood</td>
</tr>
<tr>
<td>RIGLS</td>
<td>Restrictive iterative generalised least-squares</td>
</tr>
<tr>
<td>S.E.</td>
<td>Standard error</td>
</tr>
<tr>
<td>UKCCSG</td>
<td>United Kingdom Children's Cancer Study Group</td>
</tr>
<tr>
<td>URMA</td>
<td>Univariate random-effects meta-analysis</td>
</tr>
<tr>
<td>VMA</td>
<td>Vanillylmandelic acid</td>
</tr>
</tbody>
</table>
Chapter 1

AIMS AND BACKGROUND FOR THE THESIS

Chapter overview

In this first chapter I will introduce the key aims and rationale for this thesis on evidence synthesis of prognostic marker studies. In particular, I will describe the background to the use of prognostic markers in healthcare and I will explain why formal methods for evidence synthesis are needed to facilitate evidence-based use of markers in practice. This chapter will also introduce frequentist and Bayesian approaches to meta-analysis, and it will conclude by providing the outline of subsequent chapters and the rationale for the research that will be undertaken in each.

1.1 Aims of the thesis

Prognostic markers are important clinical tools because they help to identify patients with different risks of outcome (e.g. recurrence of disease) and thereby facilitate the most appropriate treatment strategies and aid patient counselling. Evidence-based results regarding the most relevant prognostic markers for clinical practice are therefore highly desirable. The primary aim for this thesis is to identify and formally demonstrate the current problems and issues associated with undertaking a systematic review and meta-analysis of prognostic marker studies, and to then develop guidelines and methods that will allow clinically relevant evidence-based results to be obtained in the future. In particular, I will aim to develop guidelines about how to report results from primary prognostic studies in order to improve current standards and allow more pertinent information to be made available for meta-analysis. Furthermore, I will aim to identify methodological issues and then consider the development and application of methods to address these. In particular, I will develop meta-analysis methods that utilise the correlation between overall survival
and disease-free survival statistics to specifically limit the serious problems of unreported outcomes, biased within-study reporting and dissemination bias, all of which commonly hinder evidence synthesis of prognostic markers. By achieving these aims this thesis should complement a growing movement toward better primary and evidence-based research of prognostic markers, and should thereby facilitate the most appropriate use of prognostic markers in the clinical management and treatment of patients.

1.2 Prognostic markers and prognostic marker studies

From this point until Section 1.9 I will provide a detailed background to the use of prognostic markers and methods for evidence synthesis in healthcare, and this will lay the foundations for the research in subsequent chapters of the thesis.

Prognostic markers (also called prognostic factors or prognostic variables) are essentially any patient characteristic or measurement that can be used to predict the most likely clinical outcome of that patient. They can include simple measures such as age, sex, stage of disease or size of tumour, but can also include more complex factors such as abnormal levels of proteins or catecholamines, and unusual genetic mutations. Such markers are used primarily to: (i) help indicate the molecular pathogenesis of a disease, (ii) help decide the most appropriate treatment strategy, (iii) predict the response to therapy, and (iv) aid patient counselling. The main focus for this thesis is in (ii), (iii) and (iv), where prognostic markers are used for informing clinical practice and aiding clinicians in their treatment and management of patients. In particular, I will be focusing heavily on using markers that predict, either or both of, overall survival (OS) and disease-free survival (DFS) in patients with cancer (OS and DFS will be defined in detail in Section 1.7) [1]. However, prognostic markers can be used in many other clinical areas, such as in the management of patients with heart disease [2], head injuries [3], or dementia [4], and for many other outcomes such as surgical and hospitalisation complications [5], brain damage [3], and arthritis [6].
Given the clinical importance of prognostic markers, there is a continual need for individual primary studies to identify the most suitable markers that can be used in practice. There are a number of purposes of prognostic marker studies, including those that seek to improve understanding of a disease process, those that seek to improve the design and analysis of trials, and also those which try to develop prognostic models or strategies to be implemented in practice [7]. However, the vast majority of prognostic marker studies have the purpose to establish the association between a single putative marker of interest and one or more clinical outcomes, and it is these studies that I will be particularly be focusing on in this thesis. The main research area where hundreds of prognostic marker studies are published each year is in oncology, and as a real motivating example to the study of prognostic markers I will now consider prognostic markers of cancer in detail.

1.3 Prognostic markers in the clinical management of cancer patients

1.3.1 What is cancer?

'Cancer' has become a very common word in our society today, and it is rare that a family is not touched in some way by cancer. There are many forms of cancer, with those of the breast, lung, bowel and prostate by far the most incident in the United Kingdom (see Office for National Statistics at http://www.statistics.gov.uk/). One-third of people in the Western world will get some form of cancer in their lifetime and it is now the cause of more than one-fifth of all deaths in Western countries [8]. It is now recognized that cancer, in its simplest form, is a genetic disease, or more precisely, a disease of abnormal gene expression [9]. The disease is characterised by the abnormal behaviour of cells, in which they grow and multiply abnormally, and tend to spread outside their normal 'patch', no longer responding to normal regulatory mechanisms. Starting from a single cell that has become malignant, a lump develops which may have the capacity to destroy cells around it, to invade surrounding tissues and to break into blood vessels which will then carry malignant cells to other parts of the body [10].
The media reports daily on the latest drugs or treatment that is available for cancer patients, whilst giving advice on the life-styles, food and drink that may help prevent the disease. However, few people actually feel confident that they know what cancer is[8], and this lack of knowledge and understanding amongst the general public is not surprising when one considers the uncertainty still present at the scientific level. Thousands of epidemiological, biological and clinical studies have been performed across the world hoping to obtain a more in-depth comprehension of the disease. Whilst many questions have been answered and parts of the cancer jigsaw are in place, the whole picture has yet to be put together[8]. One factor that is known to be associated with cancer incidence is age. Most malignancies are diagnosed in patients over the age of 65, making age an important risk factor for development of many types of cancer[11-14]. Other factors that have been attributed to cancer include family history, diet, alcohol, sexual behaviour and hereditary factors such as mutated chromosomes and genes. However, the importance of each is hard to assess because so many of these known and unknown factors are related with one another and so results are easily confounded. Hence, although studies and intensive research to find the mechanisms and components that cause cancer are ongoing their success is often limited.

1.3.2 Limiting the impact of cancer using prognostic markers

As the causes of cancer are not easy to understand, it is very difficult to prevent the disease from occurring in the first place and it therefore becomes even more important to limit the impact of cancer when it has developed. The fight against the disease at this level is progressing in many ways. Screening programmes have been initiated in order to identify cancer cases earlier than they would normally be detected, and so enable treatment to proceed at a much earlier stage of disease[15;16]. New ways to actually treat patients with cancer, in order to cure or limit the disease, also form an on-going research area. For
example, the use of surgery and chemotherapy, either individually or in combination [17], provides hope to patients with neuroblastoma that their tumour can be removed or destroyed. Drugs and methods to improve quality of life and prolong survival time are continually developed and evaluated [18], as are methods of detecting and treating recurrences of cancer [19,20]. There are also continual improvements in the clinical management of patients with cancer, in which treatment options and procedures can be tailored specifically for individuals depending on their tumour status and other known characteristics that can be related to prognosis. The identification of such, so-called *prognostic markers* is therefore very important to the clinician who is working to cure, contain or delay cancer progression in patients with the disease. The most common prognostic marker studies in oncology investigate the use of tumour markers.

1.3.3 *The example of prognostic tumour markers in neuroblastoma*

It is hard to produce an all-encompassing definition of a tumour marker but it essentially refers to anything in the body that can be used to distinguish types of patients with cancer and therefore aid clinical practice. Tumour markers are the result of either genetic and/or biochemical changes that can be identified by the use of recently developed technology in the fields of immunohistochemistry, flow cytometry, cancer genetics and molecular biology, and they may reflect both the tumour burden and tumour biology [21]. They include abnormal levels of proteins and catecholamines, and unusual gene aberrations or changes – basically anything measurable in the body that becomes abnormal when a tumour is present. This essentially distinguishes it from other non-abnormal characteristics that may also aid clinical practice such as age or sex. Prognostic tumour markers may be binary by definition (e.g. gain or deletion of chromosome 1p in neuroblastoma patients [22]) but many are continuous measures (e.g. serum levels of lactate dehydrogenase (LDH) in patients with Ewing’s sarcoma [23]). However, most authors dichotomise continuous markers into a binary measure by using a cut-off level to identify ‘amplified’ and ‘non-
amplified' levels. This can be seen clearly in Figure 1.1 and Table 1.1 for continuous levels of tumour markers MYCN and LDH, amongst others, in relation to overall survival.

**Figure 1.1:** Example of study showing that high amplification (+ve group) of MYCN is an indicator of poor outcome in patients with neuroblastoma [24]

![Survival Curve for MYCN](image)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Alive</th>
<th>Dead</th>
<th>Total</th>
<th>Hazard Ratio (HR)</th>
<th>95% CI</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>TH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-ve</td>
<td>4</td>
<td>12</td>
<td>16</td>
<td>2.40</td>
<td>1.19 to 4.84</td>
<td>0.014</td>
</tr>
<tr>
<td>+ve</td>
<td>4</td>
<td>29</td>
<td>33</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NSE status</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;100</td>
<td>2</td>
<td>9</td>
<td>11</td>
<td>1.45</td>
<td>0.64 to 3.28</td>
<td>0.38</td>
</tr>
<tr>
<td>&gt;=100</td>
<td>4</td>
<td>16</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing</td>
<td>2</td>
<td>16</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LDH Status</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;1500</td>
<td>4</td>
<td>9</td>
<td>13</td>
<td>11.11</td>
<td>3.30 to 37.4</td>
<td>0.0001</td>
</tr>
<tr>
<td>&gt;=1500</td>
<td>0</td>
<td>11</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing</td>
<td>3</td>
<td>22</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ferritin Status</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;150</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>1.92</td>
<td>0.65 to 5.66</td>
<td>0.24</td>
</tr>
<tr>
<td>&gt;=150</td>
<td>3</td>
<td>21</td>
<td>24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing</td>
<td>3</td>
<td>16</td>
<td>19</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-2</td>
<td>3</td>
<td>10</td>
<td>13</td>
<td></td>
<td></td>
<td>0.003</td>
</tr>
<tr>
<td>2-3</td>
<td>0</td>
<td>12</td>
<td>12</td>
<td>4.09</td>
<td>1.58 to 10.62</td>
<td>0.003</td>
</tr>
<tr>
<td>3-5</td>
<td>4</td>
<td>12</td>
<td>16</td>
<td>0.80</td>
<td>0.34 to 1.87</td>
<td>0.34</td>
</tr>
<tr>
<td>&gt;5</td>
<td>1</td>
<td>7</td>
<td>8</td>
<td>0.88</td>
<td>0.33 to 2.32</td>
<td>0.33</td>
</tr>
<tr>
<td>NMYC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-ve</td>
<td>5</td>
<td>21</td>
<td>26</td>
<td>1.38</td>
<td>0.65 to 2.93</td>
<td>0.41</td>
</tr>
<tr>
<td>+ve</td>
<td>1</td>
<td>10</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing</td>
<td>2</td>
<td>10</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1.1: Results from a typical prognostic marker study in neuroblastoma; in particular estimates of the hazard ratio are provided in relation to overall survival for each of a number of different prognostic markers [24]. N.B. The hazard ratio will be introduced in Section 1.7.
Tumour markers are currently used in four areas in cancer generally: screening, diagnosis, prognosis and monitoring of patients [25;26]. Prognostic tumour markers are essentially any such abnormal characteristics that are indicative of future outcome and risk. For example in neuroblastoma, the most common extracranial solid tumour of childhood, a large number of genetic and biological markers have been investigated in recent years to find markers that would be more predictive of prognosis or response to treatment. Current narrative reviews suggest these include MYCN copy number, ploidy, deletion or loss of heterozygosity of chromosome 1p and gain of chromosome 17q, all of which have been linked with prognosis [1]. In particular, nearly all patients with amplification of MYCN (also referred to as n-myc, N-MYC, NMYC, or MYC-N in the literature) experience rapid tumour progression and a poor prognosis after conventional therapy (Figure 1.1, Table 1.1). MYCN is a proto-oncogene normally expressed in the developing nervous system and selected other tissues and is associated predominantly with an advanced stage of disease at diagnosis. The deletion or allelic loss of the short arm (p) of chromosome 1 has also been associated with a poor prognosis in neuroblastoma. However, most cases also have MYCN amplification, and so loss of 1p may not be a truly independent prognostic factor. There is also increasing evidence that gain of 17q genetic material is perhaps the commonest genetic abnormality in primary neuroblastoma. This region of the chromosome is likely to contain a gene (or genes) that contributes to neuroblastoma tumorigenesis when present in increased copy number. Gain of chromosome 17q has been shown to be associated with an unfavourable outcome and in some cases may act independently to MYCN [27-30]. More intensive therapeutic strategies, such as radiotherapy combined with chemotherapy and bone-marrow transplantation, may therefore be necessary where there is gain of chromosome 17q and/or high levels of MCYN [31].

In an attempt to explain the heterogeneity of neuroblastoma, Brodeur proposed a prognostic model based on composite clinical and genetic features, including tumour
markers [32]. However, no single factor appeared to predominate from his research and this emphasised the complex nature and incomplete understanding of the neuroblastoma tumour, combined with a lack of clear knowledge about the most appropriate prognostic markers to use in practice when managing neuroblastoma patients.

1.4 Multiple studies, multiple markers and inconsistent evidence

A large number of individual primary studies about prognostic markers are published each year, with the majority providing results about the relationship of one or more prognostic markers with outcome(s), and as biotechnology improves the number of such studies is likely to increase yet further. Unfortunately, within any given disease area, the results of these different prognostic marker studies are often inconsistent and contradictory [33]. Furthermore, many studies have small numbers of patients and therefore a low statistical power of detecting either treatment benefits or survival benefits arising out of prognostic markers. There may also be heterogeneity across studies in what exactly was assessed; for example, different studies investigate different sets of markers, different subgroups of patients, different outcomes, different types of treatment and even use different cut-off levels to dichotomise the continuous marker levels [7].

This wealth of conflicting and heterogeneous evidence makes it extremely difficult for the clinician to ascertain the overall evidence about specific prognostic markers and, even for those markers known to be important, how to use them in practice (e.g. to which selection of patients, using which cut-off levels and applying which treatment regimen). One can only feel sorry for the clinician in this position. as he or she may feel an increasing burden to put the ever growing prognostic marker literature into practice but will find it almost impossible to determine the most suitable way forward. The wealth of information to be critically synthesised compiled with large variations in study design and study results
makes an assessment of the overall evidence an onerous task, and one that will exceed the
capacity of most clinicians, whose time is limited as it is. Clinicians therefore need help in
establishing evidence-based results about prognostic markers, and need guidance about
what the overall evidence suggests is the most fitting way to incorporate them into clinical
practice. The application of formal methods of obtaining and synthesising evidence is
therefore greatly needed in the prognostic marker field in order to achieve equity for the
individual patient, otherwise there will continue to be large uncertainty and subjective,
perhaps inappropriate use of prognostic markers in practice.

1.5 The growth of formal evidence synthesis in the medical literature
Since about 1990 there has been an ever growing movement toward evidence-based
clinical practice and the formation of evidence-based clinical and public health policies
[34]. This is essentially a recognition that guidance and regulations for public health
together with the clinical management and treatment of patients should be based on the
overall evidence across all available (published and non-published) clinical trials and other
informative studies (e.g. observational studies), and should not be based, if at all possible,
on either the results from a single study or on subjective, non-reproducible reviews of the
literature (e.g. non-systematic narrative reviews) [35;36]. However, as discussed for
prognostic markers, the large and continually expanding medical literature makes it very
difficult for clinicians and policy makers to identify all the relevant studies and synthesise
the evidence for themselves; hence, although desirable, evidence-based practice is often
not easily attained.

In response to this problem there has been a surge toward developing and applying formal
and explicit methods of identifying, collating and then synthesising the overall evidence.
Indeed, the use of such methods more than doubled from the early to the late 1990s [37].
The Cochrane Collaboration is perhaps the most significant example of the movement towards formal and reproducible reviews in healthcare. This organisation is an international body specialising in reviews of randomised controlled trials (RCTs) of healthcare interventions and is named after the British epidemiologist Archie Cochrane, who wrote one of the first books on evidence-based medicine [38]. The focus of the Cochrane Collaboration is synthesising RCTs, but formal evidence synthesis methods in healthcare are equally often applied to observational studies, which include the majority of prognostic marker studies [39]. Indeed, there is arguably an even greater need for formal methods of evaluating evidence from observational studies (e.g. cohort studies, case-control studies, cross-sectional studies) as inconsistent and heterogeneous results are more likely across individual observational studies than across RCTs [39]. Formal evidence synthesis methods have been specifically developed to help evaluate the possible causes of heterogeneity across studies [40], and the causes may include, for example, different study designs, variation in treatment methods, and also differences in study inclusion criteria.

1.6 Systematic reviews and the threat of dissemination bias

Given the potential benefits, a formal evidence synthesis is increasingly important across all types of healthcare studies. The most commonly applied evidence synthesis approach is a systematic review, which is a transparent framework for the collection, critical appraisal and synthesis of the current evidence from published and unpublished studies [34]. Indeed the Cochrane Collaboration focus on evidence synthesis using systematic reviews, and they disseminate their findings in the Cochrane Database of Systematic Reviews, part of the Cochrane Library. The term ‘systematic’ comes from the fact that the review process is performed systematically and, given clear documentation, this enables the review to be transparent and reproducible. It therefore limits or exposes any subjective elements or biases that may have occurred during the review process, and thus it facilitates the most
sensible and appropriate conclusions being drawn at the end of the systematic review. For example, if a systematic review only included studies conducted in the United Kingdom, then one would know to be cautious about extending the final evidence-based conclusions to clinical practice in other countries. Furthermore, if the review only considered studies that were in the published literature, then one may be concerned about publication bias, where unpublished studies with predominantly non-significant or non-positive results may have been missed unintentionally. Publication bias is one form of dissemination bias, which refers to the various ways the reporting and publishing of an individual study can be influenced by the nature and direction of its results [41;42], and this subject will be considered in detail in Chapter 9.

Numerous systematic reviews are published in the medical literature each month, and they would seem particularly suitable for evidence synthesis of prognostic marker studies, where they could provide a transparent and coherent approach to obtaining and synthesising the vast amounts of evidence available. Indeed, a systematic review holds the most promising way forward to obtaining clinically useful evidence-based results for prognostic markers and it is a potentially very exciting step in the right direction for this research area.

1.7 Methods for meta-analysis in medical research

Meta-analysis is the statistical component of the systematic review, which generally seeks to combine the quantitative evidence from all the individual studies identified in an appropriate statistical model to produce one overall evidence-based result [43]. However, it is important to acknowledge that a systematic review does not necessarily always lead to a meta-analysis; for example, meta-analysis may not be advisable if there was a lack of quantitative data from a small number of heterogeneous studies for which dissemination bias was a concern.
In order to facilitate a meta-analysis approach, it is common for a summary statistic of interest to be chosen and extracted, where possible, from each published (and unpublished) study. For example, for a meta-analysis of studies assessing a specific treatment, one may desire an estimate of the odds ratio from each study and the meta-analysis would produce a single average or 'pooled' estimate of the odds ratio from across studies. Other estimates that might be sought are the relative risk, the difference in proportions, or another summary statistic that can provide a measure of the treatment effect in each study.

For prognostic marker and other such survival studies, it is usual for time to event outcomes to be of interest for one or more groups of patients who are monitored from the start of the study throughout a follow-up period (e.g. 5 years). In this thesis, for meta-analysis I will be particularly focusing on summary statistics that assess the OS and DFS of patients from primary prognostic marker studies. The OS of a group of patients considers the number of patients who have died during follow-up, their time of death, and the number of patients who were alive at the end of their follow-up. The DFS of a group of patients takes into account the number of patients who had a recurrence of disease or had died, their time of recurrence or death (whichever came first), and the number of patients who were alive at the end of their follow-up. For DFS, most often a 'recurrence of disease' refers to the return of cancer following primary treatment (e.g. surgery), from which the patient is initially deemed 'disease-free', but the exact definition of 'recurrence of disease' may vary from study to study (see Sections 2.1.5 and 2.9.1 for further discussion on the implications of this for meta-analysis).

For summarising OS and DFS in a study perhaps the most useful summary statistic is the hazard ratio (HR) as it provides the difference in (instantaneous) risk of an event (e.g. death (OS), recurrence of disease/death (DFS)) between different groups of patients.
defined by prognostic marker levels [44]. It essentially provides the relative survival between (usually two) groups of patients (e.g. ‘high’ versus ‘low’ marker levels) comparing the observed and expected number of events in each group. Furthermore, it is preferred to the relative risk and odds ratio statistics because it takes into account the whole follow-up period, rather than just one specific time-point, and also incorporates those patients who were censored before the end of the follow-up period, i.e. it includes the information up to time \( t \), about those \( i = 1 \) to \( m \) patients who were lost to follow-up after time \( t \). However, the inclusion of censored observations is reliant on the assumption that the censoring is not related to the event in question, but this may not always be the case; for example, the reason some cancer patients withdraw early from a prognostic marker study may be related to a recurrence of disease and their ill-health. A HR estimate also assumes that the underlying risk between the groups in question remains constant throughout the duration of the time-period assessed, but this may also not be valid; for example, a marker may help identify different risk groups during the first year, but may not be beneficial in the latter years of the study. Hence, wherever possible, the assumptions regarding the HR should be critically assessed when it is estimated in a study [44].

Alongside the estimate of the summary statistic chosen, for a meta-analysis one also requires a measure of the uncertainty (e.g. standard error) about the estimate obtained, so that the meta-analysis can give relatively more weight to those estimates with small uncertainty and, conversely, relatively smaller weight to those estimates with large uncertainty. The uncertainty of each estimate is inherently linked to the number of patients in each study, so one can broadly consider that those studies with the greatest number of patients will produce an estimate with smallest uncertainty and will therefore take more weight in the meta-analysis, other things being equal. I will now briefly introduce the most common methods used for meta-analysis in the medical literature.
1.7.1 Fixed-effects meta-analysis

The simplest meta-analysis approach that is commonly used is the fixed-effects meta-analysis model, which is termed ‘fixed-effects’ as it assumes there is no between-study heterogeneity. Essentially this means that the summary statistic estimate \( \hat{Y}_i \) from each study \( (i = 1 \text{ to } n) \) is estimating the same underlying true value \( \beta \) in every study. This assumption may be unrealistic in a meta-analysis of observational studies, and it is arguably more plausible when the studies are RCTs, which will be more similar given the greater control for bias and confounding. The fixed-effects model can be written as follows, where \( s_i^2 \) denotes the \( \text{var}(\hat{Y}_i) \) and is assumed known:

\[
\hat{Y}_i = \beta + e_i, \\
e_i \sim N(0, s_i^2)
\]  

(1.1)

It is common practice to assume \( s_i^2 \) is known in the meta-analysis literature, as this assumption makes little difference in practice [45]. One can estimate the unknown parameter \( \beta \) in equation (1.1) by generalised least-squares (GLS) [46], and one obtains:

\[
\hat{\beta} = \frac{\sum_{i=1}^{n} \frac{\hat{Y}_i}{s_i^2}}{\sum_{i=1}^{n} \frac{1}{s_i^2}} = \frac{\sum_{i=1}^{n} w_i \hat{Y}_i}{\sum_{i=1}^{n} w_i}
\]  

(1.2)

This is a well-known result and \( \hat{\beta} \) is the pooled estimate across studies of the summary statistic of interest, whilst \( w_i \) is the individual weight of each study toward this pooled estimate and is equal to the inverse of \( \text{var}(\hat{Y}_i) \). One can also estimate the variance of \( \hat{\beta} \) using the inverse of Fisher’s information matrix [47], and this turns out to be equal to the inverse of the sum of the individual study weights, i.e.

\[
\text{var}(\hat{\beta}) = \frac{1}{\sum_{i=1}^{n} \frac{1}{s_i^2}} = \frac{1}{\sum_{i=1}^{n} w_i}
\]  

(1.3)
1.7.2 Random-effects meta-analysis

In a random-effects meta-analysis between-study heterogeneity is now assumed to exist and the summary statistic estimate ($\bar{Y}_i$) from each study ($i = 1$ to $n$) is therefore now estimating a different underlying true value ($\theta_i$) in each study. In addition, each $\theta_i$ is assumed to be from a distribution with mean value $\beta$ and variance $\tau^2$. Some authors argue that these model assumptions are unjustified [48;49], while others insist the random-effects model is much more realistic than the fixed-effects model because between-study heterogeneity is likely to exist in practice [40;50]. A fixed-effects meta-analysis may be most plausible when the studies of interest are all RCTs, because such trials are all likely to be well-designed (e.g. patients are randomised to different groups) and will often have very similar characteristics (e.g. in terms of treatment used, patients assessed, definition of outcomes and statistical analyses used). However, for prognostic marker studies a random-effects meta-analysis approach is perhaps more realistic because between-study heterogeneity is likely to exist due to the primary studies having poor design standards [7;33], and there will often be little consistency across studies in important clinical and statistical factors such as the patients and outcomes assessed, the type of treatments used, the method of measuring markers, the cut-off level used to dichotomise continuous markers and the method of statistical analysis, amongst other factors [1].

Again of most interest from the meta-analysis is an estimate of the pooled value $\beta$, but one may now also be interested in the estimate of the between-study variance $\tau^2$. One can extend equation (1.1) to a random-effects meta-analysis model as follows, where $\theta_i = \beta + u_i$ and $s_i^2$ is still assumed known:

$$
\bar{Y}_i = \beta + u_i + e_i \\
u_i \sim N (0, \tau^2) \\
e_i \sim N (0, s_i^2)
$$

(1.4)
This model will be referred to as a *univariate* random-effects meta-analysis (URMA) model from Chapter 3 onwards. To estimate the parameters in equation (1.4) one can again use GLS and the estimate of the pooled value $\beta$ is as follows:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} \frac{\bar{Y}_i}{s_i^2 + \bar{r}^2} = \frac{\sum_{i=1}^{n} w_i \bar{Y}_i}{\sum_{i=1}^{n} \frac{1}{s_i^2 + \bar{r}^2}}}{\sum_{i=1}^{n} w_i}$$  \hspace{1cm} (1.5)

One can see that the study weights of $w_i = \left(\frac{s_i^2 + r^2}{s_i^2 + \bar{r}^2}\right)^{-1}$ in equation (1.5) are different to those study weights ($w'_i$) in the fixed-effects model, and the variance of $\hat{\beta}$ is now:

$$\text{var}(\hat{\beta}) = \frac{1}{\sum_{i=1}^{n} \frac{1}{s_i^2 + \bar{r}^2} \sum_{i=1}^{n} w_i}$$  \hspace{1cm} (1.6)

The change in the weighting formula can lead to potentially large differences between random-effects and fixed-effects meta-analysis results because as $\bar{r}^2$ increases relative to the $s_i^2$'s the weighting becomes more similar for each study. Some authors have severe problems with this because when there is relatively large between-study heterogeneity those studies with large numbers of patients are treated similar to those studies with much smaller numbers of patients [48].

One other important aspect of fitting a random-effects model is that $r^2$ also has to be estimated alongside $\beta$, which is why $\bar{r}^2$ is used in equations (1.5) and (1.6). This makes the estimation procedure iterative (called iterative generalised least squares (IGLS) [46]). so that separate estimates of $\beta$ and $r^2$ are obtained at each iteration until a pre-specified convergence criteria (e.g. $< 10^{-6}$) is reached between successive iterations for both parameters. Equation (1.5) is used to estimate $\beta$ at each iteration, with the value of $\bar{r}^2$ its estimate from the previous iteration, whereas equation (1.7) is used to make the GLS estimate of $r^2$ at each iteration $m$, as follows:

---

*R.D. Riley, Ph.D. Thesis*  
*Chapter 1*  
*16*
\[
\hat{\tau}^2 \text{ at iteration } m = \hat{\tau}^2_m = \frac{\sum_{i=1}^{n} (Y_i - \hat{\beta})^2 - s_i^2}{\sum_{i=1}^{n} (s_i^2 + \hat{\tau}^2_{m-1})}
\] (1.7)

Again the values of \(\hat{\tau}^2_{m-1}\) and \(\hat{\beta}\) on the right of equation (1.7) are the estimates from the previous \((m-1)\)th iteration. For random-effect models, rather using IGLS, it is often recommended to use restricted iterative GLS (RIGLS) or restricted iterative maximum likelihood (REML) estimation [51]. This has no impact on equation (1.5) but does modify equation (1.7) slightly in order to take into account the fact that both \(\hat{\beta}\) and \(\hat{\tau}^2\) are estimated from the data, and this ensures that \(\hat{\tau}^2\) is an unbiased estimate of \(\tau^2\). I will discuss RIGLS and REML further in Chapter 4, and more technical details can be found elsewhere [51]. It is also worth noting here that there is an alternative, perhaps more commonly used approach for estimating \(\tau^2\) which uses methods of moments, as suggested by DerSimonian and Laird, and in practice this approach obtains very similar estimates to the RIGLS or REML approach [50]. Another aspect of the random-effects meta-analysis is that, alongside estimates of \(\beta\) and \(\tau^2\), one can also calculate the ‘shrunken’ or ‘empirical Bayes’ estimates of the underlying true summary statistic value \(\hat{\theta}_i\) in each study by [43]:

\[
\hat{\theta}_i = \hat{\beta} + (Y_i - \hat{\beta}) \left( \frac{\hat{\tau}^2}{(\hat{\tau}^2 + s_i^2)} \right)
\] (1.8)

This statistic may be of importance if the underlying value in one particular study is of interest. One can see from equation (1.8) that \(\hat{\theta}_i\) will always be closer to \(\hat{\beta}\) than \(Y_i\), and hence why the term ‘shrunken’ study estimate is used. In a sense \(\hat{\theta}_i\) ‘borrows strength’ from all the \(\hat{\theta}_i\)’s across studies to make an ‘improved’ estimate of the underlying summary statistic in study \(i\) compared to the original estimate \(Y_i\) (see Section 11.6.1 of Sutton et al. for further details [43]).

R.D. Riley, Ph.D. Thesis  Chapter 1  17
1.7.3 Meta-regression for explaining the between-study heterogeneity

Although a measure of the between-study heterogeneity is important, often of more interest from a clinical perspective are the factors that are causing this variation between studies. One can extend the random-effects framework to include additional covariates that might explain the between-study heterogeneity and therefore reduce the value of $\hat{\tau}^2$, and Thompson suggests that this approach is very important [40;52]. For example, as systematic reviews often contain studies from a broad time period, and therefore across different treatment regimen, one may want to assess whether the year of publication ($X_t$) of the study was a cause of the heterogeneity. If one did not include this covariate in a meta-analysis model, then one could only obtain an estimate for the pooled effect $\beta$, which would relate to the average year of publication. However, by including $X_t$ one could estimate the differences from the pooled effect for different years of publication.

Equation (1.4) can be extended to a meta-regression to allow this assessment as follows:

$$
\bar{Y}_i = \beta + u_i + \xi X_i + e_i
$$

$$
\begin{align*}
    u_i &\sim N (0, \tau^2) \\
    e_i &\sim N (0, s_i^2)
\end{align*}
$$

(1.9)

The term $\xi$ is the average change in $\beta$ between two studies published one year apart, and standard statistical assessments can be made to help ascertain its importance in terms of how much it explains the between-study variance [53]. Further details on meta-regression can be found elsewhere [40;43]. However, Lambert et al. show that individual patient data (IPD), i.e. the raw patient data from each study, is generally required when investigating patient characteristics as effect modifiers in a meta-analysis [54]. For example consider age of patients in each study. From the IPD the age of each patient in each study would be known. However, without the IPD it is usual for only the average age of patients in each study to be known. If this average value is used to explain heterogeneity in a meta-
regression, the findings of Lambert et al. suggest that there would be very low statistical power to detect whether age truly is related to the benefit of the prognostic marker, and indeed one may obtain parameter estimates that are not consistent with those one would have obtained if IPD had been used. Hence, although meta-analysis of summary statistics may be adequate when one is only interested in the overall pooled value (\( \beta \)) or in the study level characteristics (e.g. year of publication), IPD will generally be necessary to perform an appropriate meta-regression when interest lies in investigating patient characteristics across studies [55].

1.7.4 An example: a meta-analysis to assess lactate dehydrogenase as a possible prognostic marker in patients with Ewing’s Sarcoma

As it has a large and well-founded number of methods already available, meta-analysis would seem a very appropriate and sensible way to quantify findings across multiple prognostic marker studies identified by a systematic review. Indeed, some systematic reviews and meta-analyses have already taken place in this field [56;57], and I will now consider one such evidence synthesis in Ewing’s sarcoma to demonstrate the application of the meta-analysis methods introduced in Sections 1.7.1 to 1.7.3.

Tumours of the Ewing’s sarcoma family most frequently occur in children and young adults, with a peak incidence in the second decade of life, and it accounts for 10% to 15% of all primary malignant bone tumours, with the mean annual incidence estimated at 0.6 per million total population [58-60]. These tumours are highly malignant, arising in bone or soft tissue, and are thought to be of primitive neural cell origin. Despite improvements in treatment and outcome, there is still a five-year DFS of only 45% to 60% indicating that relapse is very common [61], and even occurs within 3 years of diagnosis in up to 50% of patients with localised disease [62]. Furthermore, approximately 30% of patients with Ewing’s sarcoma have metastases at presentation, and for these patients OS is 10% to 20%,
compared to 60% in those with localised disease [63]. Given this background, there is therefore a great need to identify and appropriately use prognostic markers for the clinical management of patients with this cancer. In order to obtain evidence-based results, a recent systematic review of tumour markers for Ewing’s sarcoma identified 45 different prognostic marker studies between 1966 and 2000 [56]. The most commonly reported prognostic markers were identified and a meta-analysis was undertaken to clarify the evidence about each of their relationships with OS and DFS. One of the markers evaluated was lactate dehydrogenase (LDH), a tumour marker that is measured in serum of patients with Ewing’s sarcoma, and I will use this marker as my example in this section.

For the meta-analysis of LDH results across studies, the summary statistic desired from each of the published articles was the $\log_e(\text{HR})$ and its variance, and these measures were provided for OS by 5 studies and for DFS by 6 studies (Table 1.2), with non-informative censoring and proportional hazards assumed appropriate if this statistic was available. All the extracted results compared patients with high levels of LDH to those with low levels of LDH, with a (sometimes different) cut-off level used in each study to define the two groups. A fixed-effects meta-analysis, as in equation (1.1), was applied to the OS and DFS estimates separately, and for both outcomes the pooled estimates obtained suggested that patients with high levels of LDH are at an increased risk of poor outcome compared to those patients with low LDH levels (Table 1.3). The pooled estimates and their confidence intervals were far greater than 1 for both OS and DFS ($p < 0.0001$).

As all the studies about LDH were observational studies, there is a strong argument that a random-effects meta-analysis model would be more appropriate than the fixed-effects model just used. Indeed, some heterogeneity may have been caused by the use of different cut-off levels across studies (Table 1.2). A statistical measure of between-study heterogeneity (called the $Q$-statistic, for further details see Sutton et al. [43]) suggested that the heterogeneity was likely to be greater than zero for DFS but not for OS (Table 1.3).
Table 1.2: Extracted DFS and OS LDH data from a systematic review in Ewing's sarcoma, with shrunken study estimates ($\hat{\varphi}$) as obtained from the random-effects meta-analysis in Table 1.3. For references see Riley et al. [56]. The number of events in each study is not reported here because these numbers were difficult to ascertain from the majority of the published literature [56].

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Study Id</th>
<th>Year of publication</th>
<th>No. of patients</th>
<th>Cut-off level</th>
<th>log$_{10}$(HR) ((\hat{Y} ))</th>
<th>var(log$_{10}$(HR))</th>
<th>$\hat{\varphi}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>OS</td>
<td>3</td>
<td>1974</td>
<td>65</td>
<td>170 u/l</td>
<td>1.208</td>
<td>0.367</td>
<td>1.071</td>
</tr>
<tr>
<td></td>
<td>35</td>
<td>1998</td>
<td>64</td>
<td>460 u/l</td>
<td>0.833</td>
<td>0.303</td>
<td>1.071</td>
</tr>
<tr>
<td></td>
<td>43</td>
<td>1981</td>
<td>113</td>
<td>200 u/l</td>
<td>0.999</td>
<td>0.256</td>
<td>1.071</td>
</tr>
<tr>
<td></td>
<td>49</td>
<td>1998</td>
<td>20</td>
<td>300 u/l</td>
<td>1.707</td>
<td>0.582</td>
<td>1.071</td>
</tr>
<tr>
<td></td>
<td>62</td>
<td>1990</td>
<td>88</td>
<td>350 u/l</td>
<td>1.194</td>
<td>0.4</td>
<td>1.071</td>
</tr>
<tr>
<td>DFS</td>
<td>4</td>
<td>1981</td>
<td>66</td>
<td>230 u/l</td>
<td>1.592</td>
<td>0.581</td>
<td>1.309</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>1999</td>
<td>359</td>
<td>240 u/l</td>
<td>1.56</td>
<td>0.093</td>
<td>1.535</td>
</tr>
<tr>
<td></td>
<td>28</td>
<td>1997</td>
<td>98</td>
<td>600 u/l</td>
<td>0.85</td>
<td>0.412</td>
<td>1.073</td>
</tr>
<tr>
<td></td>
<td>34</td>
<td>1987</td>
<td>47</td>
<td>230 u/l</td>
<td>0.031</td>
<td>0.57</td>
<td>0.917</td>
</tr>
<tr>
<td></td>
<td>42</td>
<td>1990</td>
<td>88</td>
<td>350 u/l</td>
<td>1.308</td>
<td>0.405</td>
<td>1.253</td>
</tr>
<tr>
<td></td>
<td>65</td>
<td>1980</td>
<td>76</td>
<td>200 u/l</td>
<td>1.210</td>
<td>0.31</td>
<td>1.213</td>
</tr>
</tbody>
</table>

Table 1.3: LDH meta-analysis results for OS and DFS using RIGLS estimation.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Meta-analysis Method</th>
<th>$\hat{\beta}$</th>
<th>Pooled HR (exp((\hat{\beta} )))</th>
<th>95% CI for pooled HR</th>
<th>Between-study variance ($\tau^2$)</th>
<th>Test for heterogeneity (p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>OS</td>
<td>Fixed-effects</td>
<td>1.071</td>
<td>2.917</td>
<td>2.161, 3.936</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Random-effects</td>
<td>1.071</td>
<td>2.917</td>
<td>2.161, 3.936</td>
<td>0</td>
<td>Q = 2.126 (p = 0.713)</td>
</tr>
<tr>
<td>DFS</td>
<td>Fixed-effects</td>
<td>1.463</td>
<td>4.319</td>
<td>3.668, 5.085</td>
<td>-</td>
<td>Q = 10.475 (p = 0.063)</td>
</tr>
<tr>
<td></td>
<td>Random-effects</td>
<td>1.217</td>
<td>3.376</td>
<td>2.282, 4.993</td>
<td>0.110</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1.2: Forest plot and random-effects meta-analysis results for DFS and OS [56]. The forest plot presents each study's hazard ratio estimate (see the square blocks) separately, with their individual 95% confidence interval indicated by the adjoining horizontal line. Furthermore, the area of each block is proportional to the precision of the hazard ratio estimate. The diamond beneath the individual study estimates shows the pooled estimate and its 95% confidence interval from the random-effects meta-analysis. The vertical line corresponds to a hazard ratio of 1, which is where the risk of death (OS) or death/disease recurrence (DFS) is considered equal in both groups.
The random-effects meta-analysis model of equation (1.4) was applied separately to both OS and DFS, and similar conclusions were found to those from the fixed-effects meta-analysis for both outcomes (Table 1.3, Figure 1.2). However, the DFS pooled estimate was much closer to 1 than before and its confidence interval was slightly wider due to the addition of the $\tau^2$ term in the calculation of $\text{var}(\hat{\beta})$. Of additional interest from the random-effects analyses are the shrunken estimates of the underlying HR in each study (Table 1.2); interestingly, the shrunken estimates for OS are all the same because $\tau^2$ for OS is zero and so $\hat{\theta}_i$ equals $\hat{\beta}$ for all studies (see equation (1.8)).

Given these meta-analysis results, LDH appears to be a potentially valuable prognostic marker for Ewing's sarcoma, and its specific role in the clinical management of patients with this cancer should be investigated further. To facilitate this process it may help to extend the DFS random-effects analysis to a meta-regression and try to explain some of the between-study heterogeneity of the six DFS loge(HR) estimates. For example, one may be interested in whether choice of cut-off level causes variation in the DFS results across studies. To investigate this, a meta-regression was applied for DFS as in equation (1.9) with cut-off level as the additional covariate ($X_j$). The results indicate that cut-off level did not explain a statistically significant amount of the between-study heterogeneity (estimate for cut-off ($\hat{\xi}$) = -0.0010, 95% CI [-0.0031, 0.0010], $p = 0.32$); similarly neither did study size nor year of publication. Of course, for this meta-regression there were only six studies and so there is particularly low statistical power to detect any between-study differences [54], and therefore one cannot make any strong conclusions about whether these factors truly do influence DFS results across studies.
Although LDH appears to be of prognostic value from these meta-analyses, it is difficult to make strong clinical evidence-based conclusions due to the small number of studies and the heterogeneity in cut-off levels and other clinical factors across studies, such as treatment received and stage of disease. Furthermore, there is an additional concern that dissemination bias may have influenced which LDH results were available for the meta-analysis. For example, as none of the studies provided both OS and DFS estimates of the loge(HR), one may be particularly concerned about selective outcome reporting within-studies [64;65]. Hence, if outcomes were often not reported when their loge(HR) was not statistically significant, the LDH meta-analyses are potentially producing biased and misleading pooled estimates. Caution is therefore needed when interpreting the importance of these meta-analysis results for clinical practice. Formal methods to assess dissemination bias and its potential influence on meta-analysis results will be considered in Chapter 9.

1.8 Bayesian approaches to meta-analysis

1.8.1 Bayesian statistics

The example of prognostic marker LDH in Ewing’s sarcoma has illustrated some fundamental meta-analysis methods, all of which used a classical or a 'frequentist' statistical framework. A frequentist approach to probability considers it to be the frequency of events in an infinitely long series of repeated identical situations. However, there is an alternative Bayesian approach which considers probability to be subjective and to be the degree of belief in an event occurring in any situation, which may or may not be one of a number of repeatable such situations [66;67]. For example, the subjective approach to probability could reasonably consider 'the probability that it will rain today', whereas it is more difficult to make a frequentist probability assessment here as the situation itself is not identically repeatable and so the 'long-run' view cannot be taken.
Bayesian statistics is based on a publication by Thomas Bayes, a Presbyterian minister who lived from 1702 to 1761 [68]. This paper introduced what is now known as Bayes' theorem, which can be expressed mathematically as:

\[ p(\theta | y) = \frac{p(y | \theta) p(\theta)}{p(y)} \] (1.10)

, where \( p(\theta | y) \) is called the posterior probability of \( \theta \) given data \( y \), and \( p(\theta) \) is the prior probability of \( \theta \). Furthermore, \( p(y | \theta) \) is the probability (or likelihood) of \( y \) given parameter \( \theta \), and \( p(y) \) is the overall probability of \( y \) occurring. This is an uncontroversial mathematical result about conditional probabilities. However, it is the use of this result to make probabilistic inferences about \( \theta \) that creates tensions with the frequentist approach. Given data \( y \), the frequentist makes inferences about the value of \( \theta \) from the likelihood \( p(y | \theta) \) using an inverse argument, i.e. how plausible are the data given different values of \( \theta \). However, the Bayesian approach is to make inferences about \( \theta \) from the posterior distribution \( p(\theta | y) \), which takes into account both the data \( y \) and the prior beliefs about \( \theta \). Indeed, as \( p(y) \) is a constant, Bayes' theorem in equation (1.10) can be expressed simply in terms of those components including \( \theta \), as follows:

\[ p(\theta | y) \propto p(y | \theta) p(\theta) \] (1.11)

For example, if the data \( y \) is distributed normally with unknown mean \( \theta \), then the frequentist would specify the likelihood \( p(y | \theta) \) using this distribution and, perhaps by maximising the likelihood with respect to \( \theta \), would then make inferences about the most likely value of \( \theta \). However, assuming the prior beliefs about \( \theta \) can also be expressed as a normal distribution, the Bayesian approach would combine the likelihood with this prior distribution to obtain the posterior distribution \( p(\theta | y) \), which would also be normal in this conjugate prior distribution situation [67]. The Bayesian can then use this posterior
distribution to make direct probability statements about $\theta$, such as its mean and the 95% probability (or credibility) interval within which it lies. The frequentist cannot make such direct probability statements about $\theta$ as it only uses $p(y/\theta)$, which assumes $\theta$ is fixed, and so the frequentist has to make 'long-run' probability statements such as a 95% confidence interval, which means that if the situation was repeated an infinite amount of times 95% of the confidence intervals formed would contain the true value of $\theta$.

The main difference between frequentist and Bayesian statistics is that the Bayesian approach additionally incorporates prior information alongside the data. Frequentist statisticians object to the use of prior information because it is often subjective, representing the personal judgement of the analyst, and they therefore believe that Bayesian methods suffer from a lack of objectivity. Bayesian statisticians believe the use of prior information leads to more appropriate and interpretable results as all pertinent external information can be incorporated and direct, clinically relevant probability statements can be made from the posterior distribution. Furthermore, Bayesian statisticians would argue that the frequentist use of $p(y/\theta)$ to make inferences regarding $\theta$ is counter-intuitive, especially as $p(y/\theta)$ assumes $\theta$ is fixed (and therefore known) when actually it is unknown. In response to this, proponents of the frequentist approach would say that Bayesians can only consider $p(\theta|y)$ because they specify a possibly subjective prior for $\theta$.

1.8.2 Markov Chain Monte Carlo methods

When a normal prior distribution and normal likelihood are specified, I mentioned in Section 1.8.1 that this also produces a normal posterior distribution. However, in other more complex and multivariate situations the posterior distribution will often not be of a recognised form and the parameters must be estimated by numerical methods [69]. Indeed,
in multivariate situations the usual requirement is the \textit{marginal} posterior distribution for each of the \( n \) parameters of interest; for example, given \( n \) parameters, the marginal distribution of interest for \( \theta_1 \) would be:

\[
p(\theta_1|y) = \int p(\theta|y) \, d\theta_2 \ldots d\theta_n
\]  

(1.12)

The reason for desiring only marginal distributions is that many parameters in multivariate situations are essentially nuisance parameters, and are not of particular interest themselves. For example, if a Bayesian alternative to the frequentist random-effect meta-analysis model of equation (1.4) is taken, there are still two unknown parameters, \( \tau^2 \) and \( \beta \), but one is usually only interested in the pooled value \( \beta \) and would therefore primarily desire \( p(\beta|y) \), which is independent of the nuisance parameter \( \tau^2 \).

Before 1990, the computational problems associated with deriving the marginal posterior distributions hindered the use of Bayesian statistics in practice. However, since 1990, the growth of computer technology and the development of simulation methods, in particular Markov Chain Monte Carlo (MCMC) methods, has unleashed and revolutionised the subject. Of all MCMC methods, the Gibbs Sampler has proved particularly useful [69], and this is now described. Suppose one has an \( n \)-dimensional parameter vector \( \theta \) and posterior distribution \( p(\theta|y) \). Starting with an initial value of \( \theta \), say \( \theta^{(0)} \), the Gibbs Sampler proceeds by generating \( \theta^{(1)} \), \( \theta^{(2)} \), ... according to the following scheme from the set of full conditional distributions, such that at the \( m \)th iteration samples are drawn as follows:

Given \( \theta^{(m)} = (\theta^{(m)}_1, \theta^{(m)}_2, \ldots, \theta^{(m)}_n) \):

- Generate \( \theta^{(m+1)}_1 \) from \( p(\theta^{(m)}_1|\theta^{(m)}_2, \ldots, \theta^{(m)}_n, y) \)
- Generate \( \theta^{(m+1)}_n \) from \( p(\theta^{(m)}_n|\theta^{(m+1)}_1, \ldots, \theta^{(m+1)}_{n-1}, y) \)

This now provides \( \theta^{(m+1)} \).
This Gibbs Sampler sequence $\theta^{(1)}$, $\theta^{(2)}$, ... is a Markov chain [69], and the general theory of Markov chains says that, under broad conditions [69], the distribution of $\theta^{(m)}$ will converge to a unique limiting equilibrium distribution, independent of $\theta^{(0)}$. Furthermore, it can be shown that the way the Gibbs Sampler is constructed implies that its equilibrium distribution is the true posterior distribution $p(\theta|y)$ [69], such that when the Markov chain has converged, $\theta^{(m)}$ are samples from the marginal posterior distributions of the parameters of interest, for example $p(\theta_i|y)$. It is therefore imperative to make sure the chain has converged before making inferences from $\theta^{(m)}$, and this is why a burn-in period, say $\theta^{(1)}$ to $\theta^{(M)}$ is used. Choosing the length of burn-in, $M$, is subjective and although formal methods for deriving the length of burn-in for a specific model/dataset have been developed [70], they appear to be only rarely used in practice [67]. It is also important to assess the correlation between successive $\theta^{(m)}$'s, as autocorrelation suggests a lack of independent samples and that a longer burn-in and/or sample are potentially required. However, it should be noted that strong correlation and non-convergence might not necessarily be resolved simply by increasing the burn-in length.

The easiest way to implement Bayesian statistics is by using a statistical package called WinBUGS (Bayesian inference Using Gibbs Sampling) [71]. This powerful program has made the application of Bayesian methodology much more feasible and practical in real life situations [72].
1.8.3 A Bayesian random-effects meta-analysis model to assess LDH

A Bayesian alternative to the frequentist random-effects meta-analysis model of equation (1.4) is shown in equation (1.13), and this model can be easily fitted within WinBUGS (Appendix A1). Of course, for a Bayesian fixed-effects meta-analysis $\tau^2 = 0$ in equation (1.13).

\[
\begin{align*}
\text{Likelihood:} & \quad \hat{Y}_i \sim N(\theta_i, s_i^2) \\
& \quad \theta_i \sim N(\beta, \tau^2) \\
\text{‘Vague’ prior distributions:} & \quad \tau^{-2} \sim \text{Gamma} (0.001, 0.001) \\
& \quad \beta \sim N(0, 1000000)
\end{align*}
\tag{1.13}
\]

When no prior information exists, perhaps the largest problem for the Bayesian approach is how to specify ‘vague’ or ‘non-informative’ prior distributions. For the random-effects meta-analysis in equation (1.13) I have suggested two possible ‘vague’ prior distributions that could be used, but in practice one should always assess Bayesian meta-analysis results using a range of different ‘vague’ prior distributions so to ensure the most appropriate results and conclusions are obtained [67]. To illustrate this, I will now consider two situations for a Bayesian random-effects meta-analysis of the LDH data for DFS from the Ewing’s sarcoma systematic review (Section 1.7.4) [56].

Firstly, consider that no prior information was available and so the Bayesian analysis of equation (1.13) was applied using the ‘vague’ prior distributions as specified. Furthermore, two other similar analyses were applied to assess the robustness of the results when other ‘vague’ prior distributions were used for $\tau^2$ (Table 1.4). A 100000 burn-in was used in each case, but for each analysis there were question marks regarding convergence because the samples produced did not appear to have settled down (Figure 1.3). Furthermore, the posterior results (from 100000 additional samples) appeared to be highly influenced by the prior distribution specified for $\tau^2$. Such problems are likely to be indicative of the fact that...
there were only 6 studies available for the analyses, with the uncertainty of the loge(HR) estimate quite large in some of them (Table 1.2), and there is difficulty estimating the parameters. This analysis indicates the importance of critically assessing Bayesian meta-analysis results before interpreting them, and the particular difficulty in specifying prior distributions for variance components such as \( \tau^2 \). Although convergence is a concern, all the analyses produced similar posterior results to those from the frequentist random-effects meta-analysis, although the 95% interval for the pooled estimate is slightly wider due to the incorporation of the uncertainty in the posterior \( \tau^2 \) now being taken into account.

**Figure 1.3:** The 100000 samples that were taken from the MCMC chain following the 100000 burn-in for the LDH Bayesian random-effect meta-analysis of equation (1.13) for DFS; \( \hat{\beta} \) is denoted by 'pooled hazard ratio' and \( \hat{\tau}^2 \) is denoted by 'tausq'.

![Figure 1.3: The 100000 samples that were taken from the MCMC chain following the 100000 burn-in for the LDH Bayesian random-effect meta-analysis of equation (1.13) for DFS; \( \hat{\beta} \) is denoted by 'pooled hazard ratio' and \( \hat{\tau}^2 \) is denoted by 'tausq'.](image)

**Table 1.4:** LDH frequentist and Bayesian random-effects meta-analysis results for DFS. The mean value of \( \bar{\beta} \) and the median value of \( \bar{\tau}^2 \) are reported for the Bayesian analyses.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Prior for ( \tau^2 ) or ( \tau )</th>
<th>Prior for ( \beta )</th>
<th>Pooled HR, i.e. ( \exp(\bar{\beta}) )</th>
<th>95% interval for pooled HR</th>
<th>Between-study variance ( \bar{\tau}^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequentist</td>
<td>(-)</td>
<td>(-)</td>
<td>3.376</td>
<td>2.282, 4.993</td>
<td>0.110</td>
</tr>
<tr>
<td>Bayesian</td>
<td>( \tau^2 \sim Gamma(0.001,0.001) )</td>
<td>( \beta \sim N(0,1000000) )</td>
<td>3.589</td>
<td>2.012, 5.597</td>
<td>0.022</td>
</tr>
<tr>
<td>Bayesian</td>
<td>( \tau \sim Uniform(0,10) )</td>
<td>( \beta \sim N(0,1000000) )</td>
<td>3.544</td>
<td>1.618, 6.242</td>
<td>0.116</td>
</tr>
<tr>
<td>Bayesian</td>
<td>( \tau \sim N(0,1) )</td>
<td>( \beta \sim N(0,1000000) )</td>
<td>3.540</td>
<td>1.791, 6.041</td>
<td>0.098</td>
</tr>
<tr>
<td>Bayesian</td>
<td>( \tau \sim Uniform(0,0.2) )</td>
<td>( \beta \sim N(0,0.5) )</td>
<td>3.238</td>
<td>2.115, 4.800</td>
<td>0.010</td>
</tr>
</tbody>
</table>

N.B. the frequentist estimates are from RIGLS.
In the second hypothetical situation consider that, alongside the data, prior information has also been elicited from clinicians in the field who believe that the HR for LDH is most likely 1, i.e. on average they do not believe it can be used to distinguish low-risk and high-risk patients for DFS. However, they are not very confident about this estimate and their prior knowledge can be summarised by $\beta \sim \mathcal{N}(0, 0.5)$, i.e. the $\log_e(\text{HR})$ has mean zero and they believe it to lie between $-1$ and $1$ with a probability of 0.95. Furthermore, assume that they also believe that the underlying HR will be very similar across studies because from their experience tumour markers in Ewing’s sarcoma perform equally for all ages, treatments and other such sources of between-study heterogeneity. They consider a uniform$(0, 0.2)$ prior distribution for $\tau$ to be suitable, with the upper limit acknowledging their belief that there is zero probability that $\tau$ is above 0.2. This upper limit indicates that they believe the underlying $\log_e(\text{HR}) (\theta_i)$ in each study will approximately be between $\beta \pm 0.4$ (i.e. $\beta \pm 2 \times \tau$).

A Bayesian approach allows the incorporation of this prior clinical knowledge alongside the data and so one can obtain a posterior pooled estimate that acknowledge both these sources of evidence. Hence, a Bayesian random-effects meta-analysis was applied using these two prior distributions for $\tau$ and $\beta$. A 100000 burn-in was again used, and for this analysis the convergence appeared better than in previous analyses because the Markov chain of samples appeared to have settled down more, most likely due to the narrower posterior distribution for $\tau^2$, although there was still some reasonable variability due to the small number of studies involved (Figure 1.4). The posterior results are still similar to those from the frequentist model (Figure 1.5, Table 1.4), although the mean pooled estimate is slightly reduced due to the inclusion of the clinicians’ prior beliefs. Importantly there is a 95% probability that the pooled HR for DFS is between 2.12 and 4.80; hence, despite the clinicians’ prior view that LDH is of little prognostic worth, the posterior
estimate and 95% interval for the HR suggests that LDH is of potential prognostic value. The evidence from the data far outweighs the relatively vague prior clinical judgement about $\beta$ and so the posterior estimate of the HR is dominated by the data, leading to results very similar to the frequentist analysis.

Figure 1.4: The 100000 samples that were taken from the MCMC chain following the 100000 burn-in for the LDH Bayesian random-effects DFS meta-analysis incorporating informative prior distributions; $\tilde{\beta}$ is denoted by 'pooled hazard ratio' and $\tilde{\tau}^2$ is denoted by 'tausq'.

Figure 1.5: Posterior distributions for $\exp(\tilde{\beta})$ (denoted 'hr') and $\tilde{\tau}^2$ (denoted 'tausq') from the LDH Bayesian random-effects DFS meta-analysis incorporating informative prior distributions about these parameters; x-axis = parameter value, y-axis = density.

This analysis has of course been a hypothetical example, in that there was actually no prior elicitation from clinicians about the value of LDH. However, in truth many clinicians do not use LDH in practice and do not view it as a valuable tool [56]. This Bayesian analysis shows that, even when incorporating sceptical prior information about the value of LDH as a prognostic marker, the posterior meta-analysis results still suggest LDH could potentially be used to distinguish Ewing's sarcoma patients with different levels of risk. This illustrates the general importance of meta-analysis of prognostic marker studies and the specific advantages of a Bayesian approach when prior information is available about the
parameters of interest. Of course, there is still a concern that dissemination bias may be affecting these LDH meta-analysis results, and the clinical heterogeneity across studies makes it difficult to make specific clinical recommendations for using LDH in practice (see Section 1.7.4). However, the Bayesian analysis has strengthened the message that LDH may be important and that further research should be undertaken about if and how LDH can be most suitably utilised in the clinical management and treatment of Ewing’s sarcoma patients [1].

1.8.4 Benefits and limitations of a Bayesian approach to meta-analysis

I will conclude this introduction to a Bayesian meta-analysis by summarising the main advantages and disadvantages of the approach, with many of the points below based on Sutton et al. [43;73].

Advantages of a Bayesian approach to meta-analysis:

(i) Inferences can be made directly in terms of probability statements.

(ii) Prior knowledge and information from other sources can be incorporated into the meta-analysis alongside the studies identified by the systematic review. This enables all the evidence to be included in a meta-analysis and therefore more appropriate evidence-based results can be made.

(iii) Bayesian modelling takes into account the uncertainty in all the parameters to be estimated. For example, the uncertainty of \( \hat{\tau}^2 \) from a random-effects meta-analysis is automatically taken into account when estimating the pooled estimate and its standard error. This is not true in the frequentist approach of equation (1.4) where \( \hat{\tau}^2 \) is assumed known, even though it is actually only an estimate. The use of profile likelihoods can help overcome this problem in the frequentist framework [45].

(iv) Predictive statements can easily be made, conditional on the current knowledge state, using the posterior predictive distributions [67].
(v) It is often easier to make complex extensions to the more standard meta-analysis approaches using a Bayesian framework, where models can be more readily specified and estimated using Gibbs Sampling.

(vi) It facilitates extensions to evidence-based decision models, which can incorporate additional cost and benefit parameters [74].

Disadvantages of a Bayesian approach to meta-analysis:

(i) Prior distributions are hard to specify when no prior information is available.

(ii) Even where prior information is available, the elicitation of prior beliefs is non-trivial and often subjective, with few guidelines on what approach to take [75].

(iii) It may often be difficult to ascertain whether the Gibbs sampler has achieved convergence, and therefore whether the posterior results are appropriate to be used in practice.

(iv) Some complex models may be very time-consuming and relatively computationally intensive.

Despite these disadvantages, I consider Bayesian statistics a worthwhile part of the meta-analysis armoury, and in some situations it could prove beneficial over more standard frequentist methods. In this thesis, a Bayesian approach to meta-analysis of prognostic marker studies could be particularly useful when prior information needs to be incorporated in the evidence synthesis alongside the data available from the studies identified by the systematic review. This may be especially true within a hierarchical modelling framework when there a small number of studies, as external information may substantially help to estimate the model parameters (see Chapters 7 and 8).
1.9 Rationale and structure for the thesis

A systematic review with meta-analysis is clearly an ideal approach to evidence synthesis of prognostic marker studies and would facilitate evidence-based use of prognostic markers in healthcare. However, Altman and Lyman have suggested that a reliable and clinically useful systematic review and meta-analysis may not be feasible for prognostic marker studies, especially if one seeks to use extract and synthesise information from the published literature [76]. Indeed, although the approach has so far been rare, when evidence syntheses of prognostic markers have been performed they have often experienced problems [77]. For example, the Ewing’s sarcoma review previously introduced was limited by the difficulty in extracting summary statistics from individual primary studies and the heterogeneity across studies in cut-off level, amongst other clinical and statistical factors [56]. I have already indicated that these issues raise concerns about the validity and applicability of both the frequentist and Bayesian meta-analysis results presented in Sections 1.7.4 and 1.8.3. Altman has also recently documented further concerns that systematic reviews and meta-analysis are likely to be limited for prognostic marker studies [7;78].

The fears of Altman and others have so far been difficult to formally demonstrate because of the relatively few and often small-scale systematic reviews of prognostic markers currently published. However, formal demonstration is urgently required in order to make researchers aware of the issues and to clearly ascertain those problems limiting evidence-based results that need to be addressed in future prognostic marker research. Hence, rather than simply reviewing and discussing the problems suggested by the few previous publications on this subject [7;76;78], my research in this thesis will begin by formally demonstrating the problems facing an evidence synthesis of prognostic marker studies (see Chapter 2). This will enable me to discuss the previous publications in the context of the problems actually demonstrated, and it will thereby allow me to form an appropriate
research agenda to motivate the subsequent research in this thesis. I will now briefly
discuss the structure of each of the following chapters, and I will emphasise how the
research undertaken in this thesis seeks to facilitate evidence-based reviews and the
development of clinically relevant evidence-based prognostic marker results in the future.

Chapter 2 forms the basis of the thesis as it describes a large-scale empirical investigation
of the feasibility of performing a systematic review and meta-analysis of prognostic marker
studies in neuroblastoma. Many methodological problems that limit clinically relevant
meta-analysis are formally demonstrated, such as poor reporting in primary studies and
heterogeneity across studies in clinical and statistical factors. To help address these issues,
guidelines are developed to improve the quality and reporting of future primary studies,
and these should also allow more pertinent and clinically relevant evidence synthesis in the
long-term. In particular, the benefits of having IPD available from each study are discussed
in detail.

Chapter 3 begins by emphasising that guidelines and IPD may improve the current
situation but many methodological problems, although reduced, are likely to remain a
concern. One such problem is how to obtain evidence-based prognostic marker results for
two outcomes (e.g. OS and DFS) when one or both of these outcomes are often not
reported in the primary studies to be synthesised. The possibility of using bivariate meta-
analysis models is therefore discussed in order to utilise the, often strong, correlation
between these two outcomes and limit the problem of missing data. Previous applications
of bivariate meta-analysis are then reviewed, and some important yet currently unanswered
questions are identified; in particular, is a bivariate meta-analysis beneficial over a
standard univariate meta-analysis (i.e. those models described in Section 1.7) for both
complete-case and missing data situations, and is it practically possible to apply bivariate
models when the within-study correlation between outcomes is not available from each primary study to be synthesised?

Chapter 4 begins to address such questions by deriving and understanding the analytic pooled solutions obtained when fitting a bivariate random-effects model using RIGLS. These are then compared to the equivalent analytic solutions from a univariate random-effects model, and in particular the ability of the bivariate approach to obtain a smaller standard error of the pooled estimates is demonstrated.

Chapter 5 then introduces an extensive simulation study to further assess the benefits of bivariate random-effects meta-analysis for a variety of different settings for complete-case data. This chapter highlights important issues not identified analytically in Chapter 4, such as the difficulty in estimating the between-study correlation and why, when it is estimated as 1 or -1, this can produce biased estimates.

Chapter 6 then extends the simulations to missing data situations, and ascertains that, when the between-study correlation is estimated well, the benefits of the bivariate approach are substantially larger when there is missing data than when there is complete-case data. The chapter also demonstrates why it is necessary for missing data to be 'missing (completely) at random' for a bivariate random-effects model to be suitable. The plausibility of this missing data mechanism is also specifically considered for the missing OS and DFS summary statistics from the prognostic marker review in neuroblastoma.

Chapter 7 then introduces a Bayesian approach to bivariate random-effects meta-analysis, and demonstrates the benefit of prior information toward overcoming the problems associated with the between-study correlation. It also highlights the need to be cautious
about Bayesian bivariate meta-analysis results when only ‘vague’ prior distributions are specified, as they can sometimes be influential as regards the posterior pooled estimates.

Chapter 8 then focuses on the major problem of unknown within-study correlation between OS and DFS summary statistics, a problem that may limit application of bivariate meta-analysis to prognostic marker reviews in practice. To overcome this, an alternative model for bivariate random-effects meta-analysis is then introduced which does not require the within-study correlations to be known. This model is then shown by simulation to produce statistically appropriate meta-analysis results with potential benefits again over a standard univariate meta-analysis, especially in terms of the precision and coverage of the pooled estimates. The model is then applied to MYCN and other prognostic markers from the neuroblastoma review of Chapter 2, with extensions and applications also made beyond the prognostic marker field.

Chapter 9 considers how one may assess the possible impact of dissemination bias on the results from a bivariate meta-analysis, and in particular whether the bivariate framework can be useful when all of the missing data may be ‘not missing at random’. The MYCN dataset is again used to illustrate possible extensions to standard funnel plot assessments, and the particular benefit of the bivariate framework for aiding the implementation of selection models is discussed.

Chapter 10 summarises the key findings from the thesis and discusses their importance to future evidence syntheses of prognostic marker studies. The findings are placed in context of the remaining methodological difficulties that this thesis has not addressed, and then suggestions are made for further research priorities to help ensure the positive move toward evidence-based prognostic markers continues in the future.
Chapter 2

A FEASIBILITY STUDY OF A SYSTEMATIC REVIEW AND META-ANALYSIS OF PROGNOSTIC MARKER STUDIES

Chapter overview

In this chapter I will perform a large-scale systematic review and, if appropriate, meta-analysis of prognostic markers studied in neuroblastoma. I will use this as an empirical investigation of the feasibility of evidence synthesis of prognostic marker studies, and to formally demonstrate the problems suggested by Altman for making evidence-based recommendations of prognostic markers [76;78]. Following this, I will seek to generalise the problems identified in the neuroblastoma review to other areas of oncology, and indeed other disease settings. Finally, and perhaps most importantly, I will then develop specific guidelines for overcoming some of the problems encountered in order to help facilitate clinically relevant evidence syntheses of prognostic marker studies in the future.

2.1 Rationale and methods for a systematic review of markers in neuroblastoma

2.1.1 Neuroblastoma

Neuroblastoma is a neuroblastic tumour of the primordial neural crest and is the most common extracranial solid tumour of childhood [1], with about 80% of children presenting at less than 4 years old (median age is 22 months). The incidence in the U.K. and the U.S.A. is approximately 1 in 7000 live births, and there is slight sex predominance in most series with a male-to-female ratio of 1.2:1. The disease accounts for 15% of all childhood cancer deaths, indicating the poor prognosis of many of the tumours [79-81]. Children with stage 1, 2 or 4s disease, or presenting in the first year of life have a good prognosis. In contrast, children (1 year of age and over) with stage 3 and 4 disease have 3-year survival rates of 50% and 15% respectively. Most children present over the age of 1 year with metastatic (stage 4) disease; this group has an OS of 10-20% [80;81].
A number of genetic and biological features have been investigated in recent years in an effort to improve the understanding of the behaviour of neuroblastoma and to identify tumour markers that would improve survival rates by facilitating the screening, diagnosis, prognosis or monitoring of patients. In particular, a large number of prognostic studies have identified a wide range of tumour markers associated with OS or DFS, including \( MYCN \) copy number, ploidy, and deletion or loss of heterozygosity of chromosome 1p (Ch1p) and gain of chromosome 17q (Ch17q) (see Section 1.3.3). However, it has proved difficult to identify which prognostic tumour markers are the most useful, reflecting the complex nature of the tumour and the lack of large prospective studies in which clinical outcomes have been assessed.

The study of prognostic tumour markers for neuroblastoma forms an active research area within which a large body of evidence exists, and guidance is needed to clarify the most appropriate markers for clinical use and to help the prioritisation for further prognostic marker research. This makes it an suitable area for an empirical investigation of evidence synthesis of prognostic marker studies, and as such the problems identified in this investigation are highly likely to generalise to other settings.

### 2.1.2 Search strategy

A description of the systematic review strategy is now given. The review seeks to evaluate the prognostic value of genetic and biological tumour markers, which are measurable parameters by which a transformed cell can be differentiated from its corresponding progenitor cell, and they can be detected in abnormal amounts in the blood, urine, body tissue or indeed the tumour itself (see Section 1.3.3) [82]. Of course other factors may also be of prognostic value in neuroblastoma, such as age and histological characteristics of tumours (e.g. the presence of differentiated ganglia in neuroblastoma), but these were not assessed unless the extracted evidence for the tumour markers compared such factors directly; e.g. OS results for tumour maker \( MYCN \) are often adjusted for age.
The systematic review followed the guidelines contained in the NHS Centre for Reviews and Dissemination guidelines and its underlying philosophy was to maintain breadth, synthesise the evidence qualitatively and then, only where appropriate, use quantitative methods, making procedures explicit and transparent throughout [83]. The three on-line bibliographic databases Medline, Embase and Cancerlit were chosen as a basis for identifying the relevant literature from 1966 to February 2000. An iterative procedure was used to develop an optimal search strategy, which culminated in the use of 3 important sets of keywords in the strategy (Table 2.1). The keywords in {Neuroblastoma} related to the family of this disease, whereas those in {Tumour Marker} included the named markers thought a priori to be potentially important. The set {Clinical Area} included more specific terms for the clinical use of markers in children. A paper was included if a word from {Neuroblastoma} AND a word from {Tumour Marker} AND a word from {Clinical Area} were included anywhere in the paper.

I performed the assessment of all the papers identified, however two clinicians with more experience of the neuroblastoma literature also checked the abstracts of all those papers where classification was ‘uncertain’, and approximately 10% of the papers in each of the ‘relevant’ and ‘not relevant’ categories. The first clinician was a paediatric oncology consultant, with a special interest in neuroblastoma, whilst the second clinician was a translational scientific research fellow, with a special interest in small round cell cancers of childhood. Both these investigators held regular meetings with me about the review process I was using and how I was assessing the literature, and we all had previous experience of identifying relevant tumour marker literature from a previous systematic review and subsequent meta-analysis [1;56]. Copies of all the papers classified as ‘relevant’, together with all papers whose relevance remained unclear after additional assessment of abstracts by the clinicians, were obtained and then read thoroughly to make a final decision as to their inclusion.
Table 2.1: Sets of keywords used in the literature search of Medline, Embase and Cancerlit

<table>
<thead>
<tr>
<th>(Neuroblastoma)</th>
<th>(Clinical Area)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neuroblastoma</td>
<td>Patient(s)</td>
</tr>
<tr>
<td>Ganglioneuroblastoma</td>
<td>Child</td>
</tr>
<tr>
<td>Ganglioneuroma</td>
<td>Children</td>
</tr>
<tr>
<td></td>
<td>Prognosis</td>
</tr>
<tr>
<td></td>
<td>Diagnosis*</td>
</tr>
<tr>
<td></td>
<td>Monitoring*</td>
</tr>
<tr>
<td></td>
<td>Follow-up</td>
</tr>
<tr>
<td></td>
<td>Prognostic</td>
</tr>
<tr>
<td></td>
<td>Diagnostic</td>
</tr>
<tr>
<td></td>
<td>Pediatric</td>
</tr>
<tr>
<td></td>
<td>Paediatric</td>
</tr>
<tr>
<td>Tumour marker(s)</td>
<td>DOPA</td>
</tr>
<tr>
<td>Tumor marker(s)</td>
<td>Neuron-specific enolase</td>
</tr>
<tr>
<td>Marker(s)</td>
<td>NSE</td>
</tr>
<tr>
<td>N-MYC</td>
<td>Ferritin</td>
</tr>
<tr>
<td>NMYC</td>
<td>Lactate dehydrogenase</td>
</tr>
<tr>
<td>MYCN</td>
<td>LDH</td>
</tr>
<tr>
<td>Tyrosine Hydroxylase</td>
<td>Ganglioside(s)</td>
</tr>
<tr>
<td>TH</td>
<td>Monosialganglioside</td>
</tr>
<tr>
<td>Dopa-decarboxylase</td>
<td>Disialoganglioside</td>
</tr>
<tr>
<td>DDC</td>
<td>C-neu</td>
</tr>
<tr>
<td>Phenylethanolamine-N-methyl transferase</td>
<td>C-myc</td>
</tr>
<tr>
<td>PNMT</td>
<td>Neuropeptide(s)</td>
</tr>
<tr>
<td>PGP9.5</td>
<td>Somatostatin receptors</td>
</tr>
<tr>
<td>Dopamine-beta-hydroxylase</td>
<td>Telomerase</td>
</tr>
<tr>
<td>DBH</td>
<td>CD44</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>Mitotic index</td>
</tr>
<tr>
<td>Drug resistance</td>
<td>RT-PCR</td>
</tr>
<tr>
<td>MRP</td>
<td>Dopamine</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>NB84</td>
</tr>
<tr>
<td>3,4 – dihydroxy phenyl alanine</td>
<td>Noradrenaline</td>
</tr>
<tr>
<td>1p deletion</td>
<td>Adrenaline</td>
</tr>
<tr>
<td>DNA diplody</td>
<td>Vanillylmandelic acid</td>
</tr>
<tr>
<td>17q</td>
<td>VMA</td>
</tr>
<tr>
<td>14q</td>
<td>Epinephrine</td>
</tr>
</tbody>
</table>

* The terms ‘diagnosis’, ‘monitoring’ and ‘screening’ were included because this systematic review of prognostic markers was taken from a larger scale systematic review which assessed the use of tumour markers in paediatric oncology for diagnosis, monitoring and screening, as well as prognosis [1].

2.1.3 Inclusion criteria

To be included in the systematic review a paper had to provide a quantitative result or give tabulated IPD evaluating the use of a tumour marker in neuroblastoma. The paper had to be based on a primary research study of humans relevant to the clinical area of prognosis. The criteria for classifying a ‘prognosis paper’ was that the paper had to present data in the form of summary statistics or IPD for tumour marker levels at a measured point in time with relation to the outcome of patients at the end of a specific follow-up period. There

R.D. Riley, Ph.D. Thesis

Chapter 2
was no restriction on age of patients in the study, although approximately 90% of papers included just 0-18 year olds.

2.1.4 Exclusion criteria

Papers that reported only laboratory work, methodology for identifying new markers or results from animal studies were excluded. Review articles and non-English language papers were also excluded, the latter on the grounds of time and resources available.

2.1.5 Information extracted

From the included papers, information was extracted on the outcome reported (overall (OS) or disease-free (DFS) survival), the tumour markers assessed, year of publication, the cut-off levels for each marker if applicable and, if so, the total number of patients and deaths within each high and low subgroup. The age range and stages of neuroblastoma disease represented by the patients in each study were also recorded, as these were known a priori to be important prognostic clinical features [32]. The definition of OS and DFS were given in Section 1.7. However, although the OS of patients is a standardised outcome across studies as the single event of interest is death, the definition of DFS may actually be slightly different across studies because individual studies may vary in how they diagnose and classify a ‘recurrence of disease’. In the review, wherever DFS was reported I assumed, unless otherwise stated, that this referred to the return of cancer following primary treatment, from which the patient is deemed ‘disease-free’ (see Section 2.9.1 for further discussion on this issue).

2.1.6 Meta-analysis of the prognostic marker studies

Due to time-restraints, only meta-analysis of those markers on which 8 or more papers provided data was considered. The loge(hazard ratio) and its variance (denoted loge(HR) and var(loge(HR)) from here onwards) were the essential information required from each
study, as they provide an important comparative estimate of the risk of death/disease recurrence between two groups of patients (see Section 1.7). Furthermore, there are several indirect estimation methods available when these statistics are not directly reported [84], and the \( \log_e(HR) \) has an approximate Normal distribution for large samples, making it particularly amenable to meta-analysis techniques. Where the \( \log_e(HR) \) and its variance were available from a study, I made the assumption of non-informative censoring and proportional hazards unless it was otherwise indicated by the primary study itself (this issue will be discussed further in Section 2.9.1). It was common for a paper to report more than one prognostic result by relating one or more markers to OS and/or DFS, and also by providing unadjusted and/or adjusted results (e.g. adjusted for age, stage of disease). Estimates of the \( \log_e(HR) \) and \( \text{var}(\log_e(HR)) \) comparing two groups defined by a single marker level were sought from \textit{all} the OS and DFS reports. Two groups were common because, although some markers took only binary values (e.g. Ch1p – deletion or no deletion), it was also usual for primary studies to dichotomise continuous variables using a cut-off level in order to categorise patients into high and low risk groups.

Ideally, for making the most important clinical conclusions and the fairest comparisons across studies, \textit{adjusted} \( \log_e(HR) \) results would have been preferred across studies (e.g. \( \log_e(HR) \) results for \textit{MYCN} adjusted for age and stage of disease). However, prior knowledge indicated that adjusted results were likely to be highly inconsistent in the factors for which adjustment was made (see Section 2.3.2) [76]. Hence, as a starting point, unadjusted estimates were sought and an adjusted estimate was only sought in the absence of an unadjusted result. However, even where unadjusted results were obtained, the information about possible available adjusted results was also recorded so that if ultimately a meta-analysis of the adjusted results was possible, I could then return to obtain any other adjusted results required. These decisions were made partly due to time-constraints but
also because a previous small-scale systematic review of prognostic markers had
documented problems extracting results even for simple unadjusted HRs [56].

A five-step sequential process (Figure 2.1) using ten different direct and indirect methods
(Table 2.2), based on the approach of Parmar et al. [84], was used in attempt to obtain
\( \log_e(\text{HR}) \) and \( \text{var}(\log_e(\text{HR})) \). Studies that had a sample size less than 25 were not included
in Steps 2-5, again due to time-constraints. The following points about the sequential
process should also be noted:

- If the \( p \)-value was needed but reported as ‘\( p < X \)’ then \( p = X \) was assumed in order to
  make use of the result and produce conservative estimates. However, if ‘\( p > X \)’ or
  ‘not significant’ was reported then no \( p \)-value was assumed.

- If a paper gave two or more results comparing the same outcome and patients but
  used a different cut-off level in each case, then each of these was counted as a
different report. Subsequent meta-analyses would only include one of these results.

- If a report only presented results that compared 3 or more groups of patients (defined
  by 2 or more cut-off levels) these were not used to estimate the \( \log_e(\text{HR}) \) and its
  variance comparing 2 groups. This would have required further estimation methods
  and introduced unwanted heterogeneity into the results.

- During Steps 1 and 3, a result from a Cox regression analysis was preferred to an
  equivalent result from a log-rank/Wilcoxon test because the former was more likely
to directly provide the \( \log_e(\text{HR}) \) and its variance required. Similarly, a \( \chi^2 \)-statistic
  was preferred to the \( p \)-value if given a choice because the \( p \)-value presented was
  often inexact (e.g. \( p < 0.05 \)).

- For Step 2, IPD was defined as such only if it gave baseline marker levels, the time to
death or recurrence of disease or end of follow-up, and also the final disease status at
that time.
• If IPD was presented but estimates using Step 2 were still not possible the rest of the paper was screened to assess whether Steps 3, 4 or 5 could be used.

• For Step 5, survival curves without censoring points on it were not considered in Step 5 because of the further uncertainties this introduced, such as the unknown time that patients were lost to follow-up. Step 5 was not applicable if Step 3 had failed, because if a survival curve with censoring points had been presented this would have been used to calculate the missing statistics required in Step 3 (e.g. the number of deaths).

• The order of preference of Methods 3-10 is that order shown in Table 2.2.

Table 2.2: Description of the methods used to obtain estimates of the loge(hazard ratio) and its variance. The summary statistics required for each method are shown together with the number of times each method was successfully used in Steps 1-5 of the extraction process. Methods 3-10 were used in order of preference shown.

<table>
<thead>
<tr>
<th>Method</th>
<th>Summary statistics or data required</th>
<th>Step 1</th>
<th>Step 2</th>
<th>Step 3</th>
<th>Step 4</th>
<th>Step 5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>HR or loge(HR) &amp; V</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>Individual patient data</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>loge(HR) &amp; CI</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>HR &amp; CI</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>5</td>
<td>loge(HR) &amp; p-value</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>HR &amp; p-value</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>7</td>
<td>HR, group numbers &amp; total events</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>$\chi^2$-statistic, group numbers &amp; total events</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>9</td>
<td>p-value, group numbers &amp; total events</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>88</td>
<td>88</td>
</tr>
<tr>
<td>10</td>
<td>Survival curve</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

Total 124 41 27 8 4 204

HR = hazard ratio; V = variance of the loge(HR); CI = confidence interval
Figure 2.1: The methods and results at each stage of the sequential process used to obtain a single direct or indirect estimate of the $\log_e(\text{hazard ratio})$ and its variance for each of the reports where one of the 13 tumour markers was related to overall or disease-free survival by summary statistics or individual patient data across the literature. Five steps were used, with unadjusted estimates sought primarily in each unless only an adjusted result was available or otherwise stated.

**STEP 1**
Extract direct estimates of $\log_e(\text{hazard ratio})$ & its variance given, or apply Methods 3-9 [see Table 2] using summary information explicitly provided

**STEP 2**
Use available individual patient data to calculate an unadjusted estimate from a Cox regression analysis

**STEP 3**
Apply Methods 5, 6, 8 or 9 to use a given p-value or $\chi^2$-statistic with other summary information, at least one part of which also needs estimating from figures or tables in the paper in order to proceed

**STEP 4**
Use Step 1 or 3 to obtain an adjusted estimate for those cases where an unadjusted estimate was not possible from Steps 1-3 but an equivalent adjusted result was also presented in the paper

**STEP 5**
Use a survival curve to calculate an indirect unadjusted estimate

Total no. successes = 204, out of 575 possible (35.5%)
2.2 Results from the literature search and data extraction

2.2.1 Literature Search Results

3415 papers were identified from the literature search. The two clinicians agreed that 85.7% of my 'relevant' papers were indeed relevant or uncertain (42 out of 49 checked). They also agreed that 193 of 222 (86.9%) papers I deemed 'not relevant' were indeed not relevant, and classified the 29 others as 8 papers 'relevant' and 21 papers 'uncertain'. After obtaining and reading the entire articles, 15 of these 21 'uncertain' papers were ultimately classified as 'not-relevant'. Thus, according to the two clinicians, 208 of my 222 'not-relevant' papers checked had been correctly classified (93.7%). After further assessment when obtaining and then reading all the 'relevant' papers, a total of 260 papers were classified as 'relevant' to prognosis and these studied a total of 130 different tumour markers for risk stratification of patients. For references, a list of all the different markers, and further details of the search results (e.g. for details of the number of papers from Medline and Embase etc.) see Riley et al. [1].

2.2.2 Data extraction of prognostic marker results

The 13 most commonly studied prognostic markers were each selected for an in-depth study to establish their individual value as a prognostic tool (Table 2.3). Expression of CD44 gene was studied in 8 papers and the other twelve markers were studied in 10 or more prognosis papers (Table 2.3). This involved 211 (81.2%) of all the prognosis papers. Within these there were 575 different reports where levels of any of these 13 tumour markers were related to OS or DFS by summary statistics or IPD, and from each of these reports an estimate of the $\log_e(\text{HR})$ and $\text{var}(\log_e(\text{HR}))$ were sought. Only 204 (35.5%) estimates of both the $\log_e(\text{HR})$ and $\text{var}(\log_e(\text{HR}))$ could be calculated from Steps 1-5 using Methods 1-10 (Figure 2.1, Table 2.2). In particular, the $\log_e(\text{HR})$ and $\text{var}(\log_e(\text{HR}))$ were both directly provided on only 3 occasions in the 575 reports (0.005%) (Table 2.2 - Method 1), and all were from a single paper by Berthold et al. [85].
Fortunately, IPD was frequently presented within this literature and from this 41 direct estimates were made (Table 2.2 - Step 2, Method 2). The remaining 160 successful estimates were obtained using Methods 3-10 (Table 2.2), the most frequently required of which used a p-value or $\chi^2$-statistic in combination with group numbers and total number of events, i.e. deaths/recurrences of disease (102 times) (Table 2.2 - Methods 8 and 9).

2.3 Problems limiting meta-analysis

2.3.1 Poor reporting of primary studies

Primary studies of prognostic tumour markers are clearly essential, however the general standard of reporting primary studies was inadequate (Figure 2.2). In particular, it was disappointing to only obtain 35.5% of the estimates required despite the intensive, time-consuming extraction procedure (Section 2.1.6, Figure 2.1), and this clearly hinders the suitability of meta-analysis because I could not incorporate the majority of results reported in the literature. The multitude of missing summary statistics makes one concerned that the set of available summary statistics extracted from the literature do not actually reflect the overall evidence-base. Thus, there is a strong potential for any meta-analysis results to be biased if some of the missing summary statistics were not missing at random, perhaps because of publication bias or other related dissemination biases (see Section 1.6, and also Chapter 9 later).

Figure 2.2: Examples of how prognostic results are reported in the neuroblastoma literature, and why Methods 1 – 10 of Table 2.2 were required to obtain the $\log_e(\text{HR})$ and var($\log_e(\text{HR})$)

(a) A rare occasion where the $\log_e(\text{HR})$ and var($\log_e(\text{HR})$) were presented (Method 1 could be used for this situation):

<table>
<thead>
<tr>
<th>Factor</th>
<th>$\beta$/SE($\beta$)</th>
<th>$\exp(\beta)$</th>
<th>Unfavourable</th>
</tr>
</thead>
<tbody>
<tr>
<td>MYCN</td>
<td>2.53</td>
<td>4.26</td>
<td>amplified</td>
</tr>
<tr>
<td>Age at diagnosis</td>
<td>2.06</td>
<td>5.09</td>
<td>&gt;1 year</td>
</tr>
</tbody>
</table>

R.D. Riley, Ph.D. Thesis  Chapter 2  48
(b) Example of a paper reporting the HR, its confidence interval and a (inexact) p-value (Methods 4 or 6 could be used in this situation):

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>CATEGORIES COMPARED*</th>
<th>HAZARD RATIO (95% CONFIDENCE INTERVAL)</th>
<th>P VALUE†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical factors</td>
<td>Stage: III or IV vs. I, II, or IVS</td>
<td>5.6 (2.3-13.4)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Age: &gt;1 vs. &lt;1 yr</td>
<td>3.7 (1.7-8.0)</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>Ferritin: &gt;142 vs. ≤142 µg/liter</td>
<td>6.4 (3.0 - 13.7)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>LDH: &gt;1500 vs. ≤1500 U/liter</td>
<td>4.6 (2.1-9.9)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Genetic factors</td>
<td>N-myc: &gt;1 copy vs. 1 copy</td>
<td>6.8 (3.5-13.4)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Chromosome 1p: Loss vs. no loss</td>
<td>6.7 (3.4-13.3)</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

(c) Common example of a paper presenting only a p-value and the number patients and events in each group defined by the marker level or status (Method 9 could be used in this situation):

<table>
<thead>
<tr>
<th>Prognostic Variable</th>
<th>No. of Patients</th>
<th>Deaths</th>
<th>Expected</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thoracic site</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>227</td>
<td>39</td>
<td>120</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>No</td>
<td>1,108</td>
<td>523</td>
<td>442</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 1 yr</td>
<td>490</td>
<td>76</td>
<td>247</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>&gt; 1 yr</td>
<td>845</td>
<td>486</td>
<td>315</td>
<td></td>
</tr>
<tr>
<td>Stage</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>211</td>
<td>7</td>
<td>119</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>B</td>
<td>118</td>
<td>15</td>
<td>63</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>246</td>
<td>81</td>
<td>109</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>675</td>
<td>465</td>
<td>230</td>
<td></td>
</tr>
<tr>
<td>DS</td>
<td>83</td>
<td>14</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td>DNA index</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 1</td>
<td>228</td>
<td>129</td>
<td>72</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>&gt; 1</td>
<td>426</td>
<td>120</td>
<td>177</td>
<td></td>
</tr>
<tr>
<td>N-myc</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nonamplified</td>
<td>396</td>
<td>94</td>
<td>147</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>Amplified</td>
<td>96</td>
<td>73</td>
<td>21</td>
<td></td>
</tr>
</tbody>
</table>

(d) Example of a paper presenting a survival curve to show the survival of two groups defined by the marker level or status (Method 10 could be used in this situation):

(e) Example of a paper presenting the IPD (Method 2 could be used in this situation):
Figure 2.3: A description of the key reporting problems that prevented estimation of the loge(hazard ratio) and its variance in 371 (64.5%) of the reports.

<table>
<thead>
<tr>
<th>Key Problems in the Reporting of Prognostic Markers Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>I studied 13 tumour markers in 211 papers, and identified 575 reports (involving summary statistics or IPD) which assessed their prognostic value. On trying to extract the loge(hazard ratio) and its variance from these reports I found five main problems:</td>
</tr>
<tr>
<td>(1) No appropriate statistical analysis performed or reported.</td>
</tr>
<tr>
<td>In 133 (23.1%) reports a paper reported prognostic data (e.g. the number of patients who had an event in each group) but not any results from a Cox regression analysis or log-rank/Wilcoxon test, often because no such analyses had been performed. Hence, Methods 1-10 could not be used.</td>
</tr>
<tr>
<td>It is clearly important that where one of the purposes of the study is to assess the prognostic value of markers that appropriate statistical analyses should be performed to calculate a comparative group estimate, e.g. hazard ratio, with some measure of precision, e.g. confidence interval.</td>
</tr>
<tr>
<td>(2) Hazard ratio not calculated or not reported.</td>
</tr>
<tr>
<td>In only 57 of the 575 reports (9.9%) were direct estimates of the hazard ratio or loge(hazard ratio) provided. For these, a variance (or a standard error) was given 3 times and a confidence interval 34 times. In the other 20 at most a p-value was given.</td>
</tr>
<tr>
<td>In 222 reports a Cox regression analysis or log-rank/Wilcoxon test had been performed and results given, but without a hazard ratio being stated. Instead either a p-value/x²-statistic from the analysis (n = 210) or only a survival curve (n = 12) was presented.</td>
</tr>
<tr>
<td>This illustrates the tendency of authors/journals to base the importance of a result on a p-value rather than a comparative group estimate with some measure of precision.</td>
</tr>
<tr>
<td>(3) Inexact p-values provided.</td>
</tr>
<tr>
<td>Overall, 273 of the 575 reports presented a p-value from a Cox regression analysis or log-rank/Wilcoxon test. In 126 of these, the p-value was stated as 'p &lt; X' or 'p = significant' and in 13 reports the p-value was stated as 'p &gt; X' or 'p = not significant'.</td>
</tr>
<tr>
<td>This again shows inappropriate emphasis placed on the p-value for a statistically significant result.</td>
</tr>
<tr>
<td>(4) Group numbers and group events not given.</td>
</tr>
<tr>
<td>There were 210 reports where a p-value/x²-statistic from a Cox regression analysis or log-rank/Wilcoxon test was presented but without a hazard ratio or loge(hazard ratio). From the 194 of these that had a sample size &gt; 25, only 104 indirect estimates were obtained because the group numbers and/or group events were not reported and could not be estimated from figures or tables.</td>
</tr>
<tr>
<td>The number of patients and events in groups defined by marker levels are often smaller than the overall numbers because of missing or incomplete patient data. Hence, it is important to report numbers for the groups themselves.</td>
</tr>
<tr>
<td>(5) Marker studies too small.</td>
</tr>
<tr>
<td>In Steps 2-5, estimation methods were only considered if the sample size was greater than 25. The fact that only 196 of 318 reports, which were otherwise suitable for these Steps, included more than 25 patients is a concern.</td>
</tr>
<tr>
<td>When necessary and possible, research groups need to collaborate to achieve larger sample sizes and thus increase statistical power.</td>
</tr>
</tbody>
</table>

R.D. Riley, Ph.D. Thesis  Chapter 2  50
Importantly, the reporting problems across studies are not because I specifically chose to extract the loge(HR), as similar problems were also found when seeking other summary statistics (see Section 2.5.1) and much of the missing information is straightforward, such as the number of patients in each risk group (Figure 2.3). Indeed, amongst the 371 reports that did not enable estimates to be made there were 5 common reporting problems, most of which can be simply addressed (Figure 2.3). Encouragingly, there was some evidence that the reporting of prognostic markers has improved over the last 10 years because all the papers that did provide a HR or loge(HR) were published after 1990. However, these papers still only represented approximately 17% of the total literature identified over this period (i.e. only 26 out of 157 papers published after 1990 directly reported a HR).

### 2.3.2 Heterogeneity of clinical and statistical factors

The synthesis of my estimates was also restricted by the large variability in both clinical and statistical factors. For each estimate of the loge(HR) and var(loge(HR)) obtained, the cut-off level used to dichotomise the continuous markers, stage of disease, age of patients and outcome (OS or DFS) were recorded, and also whether the estimate was unadjusted or adjusted and, if so, what adjustment factors were used. There was great diversity in these features (Table 2.3). For example, for the marker MYCN there were 94 estimates of the loge(HR) and var(loge(HR)) obtained but these involved 9 different cut-off levels, 9 different stage groups, 4 different age groups, 77 unadjusted/17 adjusted estimates and 2 different outcomes (Table 2.4). Furthermore, of the 17 estimates that were adjusted for other prognostic markers or clinical features (using a Cox regression model) only 2 were adjusted for exactly the same set of factors, and these were from the same article [79]. The heterogeneity of adjustment factors is the reason why I did not seek further adjusted results from the literature where an unadjusted result had already been obtained (see Section 2.1.6).
Table 2.3: Names of the 13 markers grouped by tumour marker class, with the number of prognosis papers identified for each, the number of reports when it was related to either overall or disease-free survival by summary statistics or individual patient data, and the number of successful estimates made of the log_e(hazard ratio) and variance; evidence of heterogeneity is shown for outcome, cut-off levels, age, stage and adjustment factors.

<table>
<thead>
<tr>
<th>Marker Class</th>
<th>Marker Name</th>
<th>Papers</th>
<th>OS and DFS Reports</th>
<th>Total Successful Estimates (ψ)</th>
<th>OS / DFS Successes</th>
<th>Different Cut-off Groups</th>
<th>Different Stage Groups</th>
<th>Different Age Groups</th>
<th>U/A</th>
<th>Different sets of adjustment factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA or chromosome</td>
<td>MYCN</td>
<td>151</td>
<td>194</td>
<td>94</td>
<td>48 / 46</td>
<td>9</td>
<td>9</td>
<td>4</td>
<td>77 / 17</td>
<td>16</td>
</tr>
<tr>
<td>DNA index</td>
<td>44</td>
<td>62</td>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>18 / 1</td>
<td>1</td>
</tr>
<tr>
<td>Chromosome 1p</td>
<td>40</td>
<td>49</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>18 / 2</td>
<td>2</td>
</tr>
<tr>
<td>Urinary catecholamines</td>
<td>VMA</td>
<td>36</td>
<td>40</td>
<td>4</td>
<td>3 / 1</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>4 / 0</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>HVA</td>
<td>26</td>
<td>29</td>
<td>2</td>
<td>2 / 0</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2 / 0</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>VMA:HVA</td>
<td>20</td>
<td>28</td>
<td>5</td>
<td>2 / 3</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>5 / 0</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Dopamine</td>
<td>10</td>
<td>11</td>
<td>2</td>
<td>1 / 1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2 / 0</td>
<td>-</td>
</tr>
<tr>
<td>Biological markers</td>
<td>CD44</td>
<td>8</td>
<td>8</td>
<td>3</td>
<td>0 / 3</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3 / 0</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>TrkA</td>
<td>16</td>
<td>21</td>
<td>11</td>
<td>4 / 7</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>9 / 2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>NSE</td>
<td>28</td>
<td>39</td>
<td>9</td>
<td>4 / 5</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>8 / 1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>LDH</td>
<td>26</td>
<td>30</td>
<td>12</td>
<td>5 / 7</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>8 / 4</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Ferritin</td>
<td>33</td>
<td>41</td>
<td>7</td>
<td>3 / 4</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>6 / 1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>MDR</td>
<td>16</td>
<td>30</td>
<td>16</td>
<td>9 / 7</td>
<td>8</td>
<td>3</td>
<td>3</td>
<td>13 / 3</td>
<td>2</td>
</tr>
</tbody>
</table>

VMA = vanillylmandelic acid; HVA = homovanillic acid; NSE = neuron-specific enolase; MDR = multi-drug resistance protein; LDH = lactate dehydrogenase.

There were 211 papers overall which reported overall survival (OS) and/or disease-free survival (DFS) results or individual patient data (IPD) for one or more of the markers.

* age groups were 'all ages', '< 1 year', '> 1 year' and 'unknown'; stage of disease groups were 'unknown' and combinations of stage 1, 2, 3, 4, 4s; cut-off includes a group for when it was 'unknown' (for an example see Table 2.4).
Table 2.4: Heterogeneity in the 94 estimates of the loge(hazard ratio) and its variance for MYCN

<table>
<thead>
<tr>
<th>Outcome</th>
<th>no.</th>
<th>Cut-off Point</th>
<th>no.</th>
</tr>
</thead>
<tbody>
<tr>
<td>DFS</td>
<td>46</td>
<td>1 copy</td>
<td>23</td>
</tr>
<tr>
<td>OS</td>
<td>48</td>
<td>2 copies</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 copies</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 copies</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5 copies</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10 copies</td>
<td>18</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Result Type</th>
<th>no.</th>
</tr>
</thead>
<tbody>
<tr>
<td>unadjusted</td>
<td>77</td>
</tr>
<tr>
<td>adjusted</td>
<td>17</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stage groups</th>
<th>no.</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>68</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>1, 2, 3</td>
<td>3</td>
</tr>
<tr>
<td>1, 2, 3, 4</td>
<td>5</td>
</tr>
<tr>
<td>2, 3, 4, 4S</td>
<td>2</td>
</tr>
<tr>
<td>3, 4</td>
<td>3</td>
</tr>
<tr>
<td>unknown</td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age groups</th>
<th>no.</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>78</td>
</tr>
<tr>
<td>&lt; 1 year</td>
<td>2</td>
</tr>
<tr>
<td>&gt; 1 year</td>
<td>5</td>
</tr>
<tr>
<td>unknown</td>
<td>9</td>
</tr>
</tbody>
</table>

OS = overall survival; DFS = disease-free survival

This inconsistent and variable reporting was reflected equally in the estimates obtained for the other 12 markers (Table 2.3). Type of treatment of patients and method of measuring the markers were not recorded but both would have added further clinical and methodological heterogeneity to that observed. The problem of heterogeneity makes coherent evidence synthesis difficult because each study is assessing the marker for slightly different subgroups of patients, different treatments, different cut-off levels and other pertinent factors. Hence, it is difficult to disentangle these factors and obtain evidence synthesis results that are relevant for clinicians in practice.

2.3.3 Publication bias and other dissemination bias issues

The common problem of publication bias alongside other dissemination biases may also have affected the data extraction. In particular, some results that do not generate formal statistically significant or clinically valuable findings may not have been published, because of a reluctance of journals to report or of researchers to present negative findings. Such problems severely limit the conclusions that can be drawn from meta-analyses because not all the available evidence can be included, and therefore the pooled results are
likely to be biased. I investigated the estimates obtained for MYCN and indeed there did appear to be evidence of publication bias, with a number of studies with smaller HRs considered to be missing (Figure 2.4). This problem is likely to be closely related to the problem of small sample sizes in some primary studies (Figure 2.3, Key Problem 5). The issue of publication and other dissemination biases for the prognostic marker datasets is considered in more extensive detail in Chapter 9. However, it is important to note that the search strategy and extraction procedure I used for the systematic review may itself have introduced some bias (e.g. by excluding non-English language articles) and this will be discussed further in Section 2.5.1.

**Figure 2.4:** Funnel plot of the OS $\log_e(HR)$ for MYCN with pseudo 95% confidence limits [41]. The assumption is that a funnel plot should form a funnel shape if there is no dissemination bias present, as estimates from smaller studies will be more widely spread about the mean effect due to larger standard errors. However, this plot for OS is not indicative of a very funnel-like shape, and asymmetry is apparent with a gap in the bottom right of the plot (see Section 9.2.1 for further details and discussion on this).

2.4 Should I proceed with meta-analysis?

2.4.1 Limited evidence-based results are possible

The poor reporting, potential for publication bias and, in particular, the large heterogeneity across studies meant it was practically impossible to perform reliable meta-analyses that would determine the clinical importance of each marker studied. Even the analysis of
subgroups of estimates was not considered realistic because it was virtually impossible to obtain subgroups that reflected patients with similar features. For example, for marker MYCN there were 48 OS estimates obtained, of which 41 were unadjusted, and 30 related to ‘all’ stages and ‘all’ ages. Furthermore, only 8 of these 30 estimates related to the most commonly used cut-off level of ‘1 copy number’, and there is then the additional problem of heterogeneity for treatment used and method of measuring the marker, not to mention the potential impact of publication or other dissemination biases. The subgroup numbers were even smaller for the other, less-studied markers; for example, lactate dehydrogenase had only 2 unadjusted OS estimates relating to the most common cut-off level (1500 U/l) and patients of ‘all’ ages and ‘all’ stages. For all the markers, it was also very difficult to appropriately use a meta-regression to explain the between-study heterogeneity because many of the heterogeneous factors were patient level characteristics (e.g. stage of disease, age, type of treatment received). As IPD was not available for the majority of studies, there would be very low statistical power to assess such heterogeneity across studies (see Section 1.7.4 and also Lambert et al. [54]).

The only possible benefit of meta-analysis using the estimates that I extracted is to highlight the results of previous studies and help prioritise which markers should be researched in the future. However, it is clear that limited clinical policy decisions can be made from this evidence-based review.

2.4.2 Meta-analysis results

Whilst acknowledging the problems of heterogeneity and poor reporting, it remained important to utilise the data extracted for each marker and determine which were potentially the most important markers to help prioritise future research. This was particularly important given the lack of clarity regarding prognostic markers in neuroblastoma from the current literature [1;86]. Meta-analysis was therefore performed
for each of the 13 markers separately for OS and DFS (Table 2.5). For the meta-analyses, if a paper provided 2 or more estimates within an outcome for the same marker and same patients by using different cut-offs then only one of these estimates was used in the meta-analysis, that based on the greatest number of patients and/or which cut-off was the most comparable with other studies (for more specific details of the few studies this affected see Riley et al [1]). Furthermore, as the majority of studies included all types of patients (e.g. all ages, all stages), the pooled OS and DFS estimates presented in Table 2.5 should therefore be considered to relate to an ‘average’ individual with neuroblastoma.

Given the large heterogeneity, univariate random-effects meta-analysis (using the frequentist approach of equation (1.4)) was performed separately for each marker and the results presented in Table 2.5 have been classified into three marker groups - DNA/chromosome abnormalities, biological markers, and urinary catecholamines. A forest plot for the MYCN results is shown in Figure 2.5, and for the other markers in Appendix A2. For the vast majority of markers, there is a statistically significant difference (in terms of the HR) between the groups defined by the markers. For example, there was strong statistically significant evidence that amplification of the MYCN gene was associated with a worse OS and DFS. The risk of death was 5.48 times greater for patients with MYCN amplification compared to those that did not have amplification (HR = 5.48, 95% CI [4.30, 6.97]), and similarly for either risk of disease recurrence or death (HR = 4.28, 95% CI [3.34, 5.49]). All the papers that could be included in the meta-analyses were published from 1985 onwards, with the majority after 1989. For example, of the 70 articles used in the MYCN OS and DFS meta-analyses, none were published before 1985, only 8 were published from 1985 to 1989, and 62 were published from 1990 onwards. Hence, the prognostic studies I included have been reported following the improved method for staging and treatment of neuroblastoma that has improved survival for children with this disease since the late 1980s [87-89].
Table 2.5: Meta-analysis results for the 13 prognostic markers, grouped by tumour marker class, for overall survival (OS) and disease-free survival (DFS) together with the number (no.) of prognosis papers identified overall and the number of estimates of the \( \log_e(\text{hazard ratio}) \) and variance obtained for each outcome.

<table>
<thead>
<tr>
<th>Marker Type</th>
<th>Tumour Marker (relationship with prognosis)</th>
<th>No. Prognosis Papers</th>
<th>Outcome</th>
<th>No. Estimates Obtained</th>
<th>Pooled Hazard Ratio*</th>
<th>95% Confidence Interval</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA or chromosome abnormalities</td>
<td>MYCN (amplification poor outcome)</td>
<td>151</td>
<td>DFS</td>
<td>46</td>
<td>4.28</td>
<td>3.34 to 5.49</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>DNA index (diploidy poor outcome)</td>
<td>44</td>
<td>DFS</td>
<td>8</td>
<td>0.33</td>
<td>0.20 to 0.56</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>Chromosome 1p (deletion poor outcome)</td>
<td>40</td>
<td>DFS</td>
<td>9</td>
<td>3.93</td>
<td>2.31 to 6.68</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Urinary catecholamines</td>
<td>VMA (elevated poor outcome)</td>
<td>36</td>
<td>DFS</td>
<td>1</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>HVA (elevated poor outcome)</td>
<td>26</td>
<td>DFS</td>
<td>0</td>
<td>1.14</td>
<td>0.65 to 1.98</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>VMA: HVA (small ratio (e.g. &lt;1) poor outcome)</td>
<td>20</td>
<td>DFS</td>
<td>3</td>
<td>0.35</td>
<td>0.17 to 0.72</td>
<td>0.0043</td>
</tr>
<tr>
<td></td>
<td>Dopamine (elevated poor outcome)</td>
<td>10</td>
<td>DFS</td>
<td>1</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Biological markers</td>
<td>CD44 (high expression good outcome)</td>
<td>8</td>
<td>DFS</td>
<td>3</td>
<td>0.06</td>
<td>0.02 to 0.21</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>TrkA (high expression good outcome)</td>
<td>16</td>
<td>DFS</td>
<td>7</td>
<td>0.26</td>
<td>0.16 to 0.42</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>NSE (high serum levels poor outcome)</td>
<td>28</td>
<td>DFS</td>
<td>5</td>
<td>5.56</td>
<td>2.11 to 14.7</td>
<td>0.0005</td>
</tr>
<tr>
<td></td>
<td>LDH (high serum levels poor outcome)</td>
<td>26</td>
<td>DFS</td>
<td>7</td>
<td>3.20</td>
<td>2.06 to 4.98</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>Ferritin (high serum levels poor outcome)</td>
<td>33</td>
<td>DFS</td>
<td>4</td>
<td>4.26</td>
<td>2.42 to 7.53</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>MDR (high expression poor outcome)</td>
<td>16</td>
<td>DFS</td>
<td>7</td>
<td>6.37</td>
<td>3.71 to 10.9</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

VMA = vanillylmandelic acid; HVA = homovanillic acid; NSE = neuron-specific enolase; MDR = multi-drug resistance protein; LDH = lactate dehydrogenase.

NA = meta-analysis not possible because no estimates or only one estimate was available.

* A hazard ratio > 1 indicates a greater risk of death/disease recurrence for patients with high/positive marker levels or those with presence of the marker, compared to those patients with low/negative levels or without the presence of the marker respectively. The pooled results should be interpreted with caution given the large heterogeneity of clinical/statistical factors across studies.
Figure 2.5: Forest plots for (a) the MYCN disease-free survival (DFS) meta-analysis and (b) the MYCN overall survival (OS) meta-analysis

(a) DFS

<table>
<thead>
<tr>
<th>paper</th>
<th>n</th>
<th>u/a</th>
<th>age</th>
<th>stage</th>
<th>cut-off</th>
</tr>
</thead>
<tbody>
<tr>
<td>44</td>
<td>90</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>1</td>
</tr>
<tr>
<td>68</td>
<td>95</td>
<td>u</td>
<td>unknown</td>
<td>all</td>
<td>1</td>
</tr>
<tr>
<td>71</td>
<td>50</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>1</td>
</tr>
<tr>
<td>106</td>
<td>59</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>1</td>
</tr>
<tr>
<td>122</td>
<td>14</td>
<td>u</td>
<td>unknown</td>
<td>all</td>
<td>expression 0/1</td>
</tr>
<tr>
<td>146</td>
<td>12</td>
<td>u</td>
<td>unknown</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>147</td>
<td>24</td>
<td>u</td>
<td>1.5 yr</td>
<td>all</td>
<td>1</td>
</tr>
<tr>
<td>149</td>
<td>92</td>
<td>u</td>
<td>1 yr</td>
<td>all</td>
<td>1</td>
</tr>
<tr>
<td>173</td>
<td>10</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>1</td>
</tr>
<tr>
<td>179</td>
<td>29</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>200</td>
<td>61</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>213</td>
<td>20</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>1</td>
</tr>
<tr>
<td>214</td>
<td>74</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>1</td>
</tr>
<tr>
<td>215</td>
<td>72</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>1</td>
</tr>
<tr>
<td>219</td>
<td>26</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>1</td>
</tr>
<tr>
<td>220</td>
<td>70</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>227</td>
<td>48</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>228</td>
<td>47</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>230</td>
<td>28</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>1</td>
</tr>
<tr>
<td>235</td>
<td>23</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>238</td>
<td>60</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>243</td>
<td>31</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>244</td>
<td>28</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>246</td>
<td>25</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>250</td>
<td>21</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>1</td>
</tr>
<tr>
<td>257</td>
<td>31</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>258</td>
<td>32</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>263</td>
<td>32</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>267</td>
<td>38</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>269</td>
<td>36</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>272</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>273</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>274</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>275</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>276</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>277</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>278</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>279</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>280</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>281</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>282</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>283</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>284</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>285</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>286</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>287</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>288</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>289</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>290</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>291</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>292</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>293</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>294</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>295</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>296</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>297</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>298</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>299</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>300</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>301</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>302</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>303</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>304</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>305</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>306</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>307</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>308</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>309</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>310</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>311</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>312</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>313</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>314</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>315</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>316</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>317</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>318</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>319</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
</tbody>
</table>

Pooled result (95% CI) 4.28 (3.34, 5.49)

Hazard Ratio (log-scale) 0.5 1.0 2.0 10.0 100.0

R.D. Riley, Ph.D. Thesis Chapter 2 58
Figure 2.5 continued ...

(b) OS

<table>
<thead>
<tr>
<th>paper</th>
<th>n</th>
<th>u/a</th>
<th>age</th>
<th>stage</th>
<th>cut-off</th>
</tr>
</thead>
<tbody>
<tr>
<td>67</td>
<td>33</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>10</td>
</tr>
<tr>
<td>44</td>
<td>23</td>
<td>u/a</td>
<td>all</td>
<td>all</td>
<td>1</td>
</tr>
<tr>
<td>58</td>
<td>48</td>
<td>u</td>
<td>all</td>
<td>unknown</td>
<td>1</td>
</tr>
<tr>
<td>89</td>
<td>48</td>
<td>u</td>
<td>unknown</td>
<td>all</td>
<td>neg, pos</td>
</tr>
<tr>
<td>89</td>
<td>48</td>
<td>u</td>
<td>unknown</td>
<td>all</td>
<td>neg, pos</td>
</tr>
<tr>
<td>99</td>
<td>28</td>
<td>u/a</td>
<td>all</td>
<td>all</td>
<td>1</td>
</tr>
<tr>
<td>105</td>
<td>36</td>
<td>u/a</td>
<td>all</td>
<td>all</td>
<td>1</td>
</tr>
<tr>
<td>107</td>
<td>35</td>
<td>u/a</td>
<td>all</td>
<td>unknown</td>
<td>1</td>
</tr>
<tr>
<td>111</td>
<td>69</td>
<td>u/a</td>
<td>all</td>
<td>unknown</td>
<td>1</td>
</tr>
<tr>
<td>119</td>
<td>12</td>
<td>u/a</td>
<td>all</td>
<td>unknown</td>
<td>1</td>
</tr>
<tr>
<td>146</td>
<td>22</td>
<td>u/a</td>
<td>all</td>
<td>unknown</td>
<td>1</td>
</tr>
<tr>
<td>160</td>
<td>48</td>
<td>u/a</td>
<td>all</td>
<td>unknown</td>
<td>1</td>
</tr>
<tr>
<td>175</td>
<td>69</td>
<td>u/a</td>
<td>all</td>
<td>unknown</td>
<td>1</td>
</tr>
<tr>
<td>176</td>
<td>99</td>
<td>u/a</td>
<td>all</td>
<td>unknown</td>
<td>1</td>
</tr>
<tr>
<td>182</td>
<td>32</td>
<td>u/a</td>
<td>all</td>
<td>unknown</td>
<td>1</td>
</tr>
<tr>
<td>197</td>
<td>23</td>
<td>u/a</td>
<td>all</td>
<td>unknown</td>
<td>1</td>
</tr>
<tr>
<td>214</td>
<td>28</td>
<td>u/a</td>
<td>all</td>
<td>unknown</td>
<td>1</td>
</tr>
<tr>
<td>216</td>
<td>91</td>
<td>u/a</td>
<td>all</td>
<td>unknown</td>
<td>1</td>
</tr>
<tr>
<td>217</td>
<td>10</td>
<td>u/a</td>
<td>all</td>
<td>unknown</td>
<td>1</td>
</tr>
<tr>
<td>247</td>
<td>26</td>
<td>u/a</td>
<td>all</td>
<td>2-4, 4+</td>
<td>10</td>
</tr>
<tr>
<td>254</td>
<td>32</td>
<td>u/a</td>
<td>all</td>
<td>2-4, 4+</td>
<td>10</td>
</tr>
<tr>
<td>260</td>
<td>36</td>
<td>u/a</td>
<td>all</td>
<td>2-4, 4+</td>
<td>10</td>
</tr>
<tr>
<td>277</td>
<td>32</td>
<td>u/a</td>
<td>all</td>
<td>2-4, 4+</td>
<td>10</td>
</tr>
<tr>
<td>285</td>
<td>32</td>
<td>u/a</td>
<td>all</td>
<td>2-4, 4+</td>
<td>10</td>
</tr>
<tr>
<td>287</td>
<td>23</td>
<td>u/a</td>
<td>all</td>
<td>2-4, 4+</td>
<td>10</td>
</tr>
<tr>
<td>291</td>
<td>23</td>
<td>u/a</td>
<td>all</td>
<td>2-4, 4+</td>
<td>10</td>
</tr>
<tr>
<td>296</td>
<td>32</td>
<td>u/a</td>
<td>all</td>
<td>2-4, 4+</td>
<td>10</td>
</tr>
<tr>
<td>312</td>
<td>20</td>
<td>u/a</td>
<td>all</td>
<td>2-4, 4+</td>
<td>10</td>
</tr>
<tr>
<td>314</td>
<td>26</td>
<td>u/a</td>
<td>all</td>
<td>2-4, 4+</td>
<td>10</td>
</tr>
<tr>
<td>319</td>
<td>89</td>
<td>u/a</td>
<td>all</td>
<td>2-4, 4+</td>
<td>10</td>
</tr>
<tr>
<td>342</td>
<td>48</td>
<td>u/a</td>
<td>all</td>
<td>2-4, 4+</td>
<td>10</td>
</tr>
<tr>
<td>352</td>
<td>32</td>
<td>u/a</td>
<td>all</td>
<td>2-4, 4+</td>
<td>10</td>
</tr>
<tr>
<td>354</td>
<td>58</td>
<td>u/a</td>
<td>all</td>
<td>2-4, 4+</td>
<td>10</td>
</tr>
<tr>
<td>356</td>
<td>64</td>
<td>u/a</td>
<td>all</td>
<td>2-4, 4+</td>
<td>10</td>
</tr>
<tr>
<td>358</td>
<td>48</td>
<td>u/a</td>
<td>all</td>
<td>2-4, 4+</td>
<td>10</td>
</tr>
<tr>
<td>370</td>
<td>24</td>
<td>u/a</td>
<td>all</td>
<td>2-4, 4+</td>
<td>10</td>
</tr>
<tr>
<td>372</td>
<td>58</td>
<td>u/a</td>
<td>all</td>
<td>2-4, 4+</td>
<td>10</td>
</tr>
<tr>
<td>391</td>
<td>60</td>
<td>u/a</td>
<td>all</td>
<td>2-4, 4+</td>
<td>10</td>
</tr>
<tr>
<td>401</td>
<td>49</td>
<td>u/a</td>
<td>all</td>
<td>2-4, 4+</td>
<td>10</td>
</tr>
<tr>
<td>412</td>
<td>10</td>
<td>u/a</td>
<td>all</td>
<td>2-4, 4+</td>
<td>10</td>
</tr>
<tr>
<td>416</td>
<td>12</td>
<td>u/a</td>
<td>all</td>
<td>2-4, 4+</td>
<td>10</td>
</tr>
<tr>
<td>431</td>
<td>18</td>
<td>u/a</td>
<td>all</td>
<td>2-4, 4+</td>
<td>10</td>
</tr>
<tr>
<td>441</td>
<td>18</td>
<td>u/a</td>
<td>all</td>
<td>2-4, 4+</td>
<td>10</td>
</tr>
<tr>
<td>448</td>
<td>18</td>
<td>u/a</td>
<td>all</td>
<td>2-4, 4+</td>
<td>10</td>
</tr>
</tbody>
</table>

Pooled result (95% CI) 5.48 (4.30, 6.97)
2.4.3 Clinical interpretation of the meta-analysis results

This is the first systematic review of tumour markers for neuroblastoma that has been reported and it forms a knowledge base for future research. MYCN, Ch1p, DNA index, VMA:HVA ratio, CD44, Trk-A, NSE, LDH, ferritin and multi-drug resistance were all identified as potentially important prognostic tools (see footnote at bottom of Table 2.5 for full unabbreviated names of these markers) [86]. However, in terms of clinical application, the pooled results must be treated with caution given the reporting problems and large heterogeneity of clinical/statistical factors across studies. For example, as publication bias is likely to be affecting the results it is likely that the pooled estimates should be somewhat smaller than those obtained. This issue is considered in further detail in Chapter 9. The pooled results are meant to guide which markers should be prioritised for further research, but the markers cannot be placed in any specific priority order, again because the reporting problems and large heterogeneity of clinical/statistical factors are potentially affecting the pooled results.

Recent studies have also indicated that Ch17q gains have important prognostic significance [27-30;90-92]. Unfortunately, many of these studies were published after the start of my review, and consequently Ch17q was not amongst the prognostic markers I selected for an in-depth evaluation. However, in light of this current knowledge, I have subsequently extracted, wherever possible, HR results from each of the 8 prognosis papers my review identified for this marker. Meta-analysis of these suggests that patients with gain of Ch17q have a statistically significantly worse DFS (from 3 studies: HR = 4.16, 95% CI [2.56, 6.77]) and OS (from 3 studies: HR = 4.30, 95% CI [2.70, 6.86]) compared to those who did not. However, these results are again subject to the problems of poor reporting and heterogeneity.
It was not possible to compare subgroups of patients or compare the individual prognostic markers, or indeed assess the benefits of using any of the markers in combination. This was again because of the poor reporting (see Section 2.3.1), and there is very low statistical power to make assessments of patient level characteristics when summary statistics, rather than IPD, are available [54]. Further research is therefore required and ideally large multicentre studies should be initiated to assess the benefits of using Ch17q, MYCN and the other potentially important prognostic markers, both individually and in combination, to improve strategies for the stratification of children with neuroblastoma for therapy.

Furthermore, prospectively planned pooled analyses (rather than the retrospective pooled analyses I have performed) using IPD may be a better way forward for future evidence syntheses in this field (see Section 2.9.7), an approach that is currently a particularly important concept for epidemiological research [93]. The multiplicity and complexity of tumour markers in neuroblastoma underlines the need for studies to be coordinated though cooperation of cancer research groups, utilising multiple laboratories and standardising methods of analysis and reporting. Large, multi-centre collaborative studies would require agreement as to which markers to measure, and my work provides an initial base that highlights the ones with the greatest importance so far (Appendix A3) [94].

2.5 Discussion and generalisations from the empirical investigation

2.5.1 Appraisal of the systematic review and data extraction

An in-depth critical assessment of this systematic review is given elsewhere [1], but there are some key issues I want to discuss here. During the systematic review I considered 3415 papers overall and identified 260 studies with results assessing tumour markers in relation to prognosis. This review will have identified the majority of the English-language literature but inevitably some papers will have been excluded unintentionally. However, it seems plausible that the reporting in such papers, and equally non-English papers, would
be equally poor and heterogeneous. My initial search strategy included the names of those markers known *a priori* to be potentially important but, in light of the large number of markers identified during the review, this list was certainly not exhaustive, although I did include more general terms such as 'marker' that will have limited the number of markers missed.

Initially my strategy had been to check the references quoted in all the 'relevant' papers as part of a 'reference explosion', in order to identify any other papers missing from the three databases I searched. However, this was not feasible given the large literature base and time-constraints, nor was it possible to systematically check for duplicates of patients across papers, although this would have been very difficult anyway due to the difficulty in differentiating sets of patients across studies [1]. I did not include non-English language papers because of the difficulties in translation, and this may have introduced bias if statistically or clinically significant studies were more likely to be (re)written for publication in an English language journal [95;96]. Other biases may have also been unintentionally introduced in my extensive search strategy adopted in Section 2.1.2, but my decisions were based on trying to utilise as much of the available published literature as possible and to limit the poor reporting problems. For example, I may have added to the publication bias issue by assuming $p = \alpha$ where $p < \alpha$ was given, but not assuming $p = \alpha$ where $p > \alpha$ or 'not-significant' was presented, and hence excluding the more negative results. This decision was based on the former being a conservative estimate of the HR, and the latter vulnerable to large inaccuracies (e.g. if $p = 0.05$ was assumed where $p > 0.05$ was presented). Fortunately, there were only 13 reports across all the 13 prognostic markers studied for which the $p$-value was required to make estimates of the $\log_e(HR)$ and $p > \alpha$ or 'not-significant' had been given. Another concern is that the extraction of summary statistics, required for the estimation of the $\log_e(HR)$ and $\text{var}(\log_e(HR))$, were not
double-checked, and thus this could mean some unintentional mistakes have been made. It should be noted that none of the aforementioned flaws in my search strategy and analysis decisions are likely to detract from the main methodological issues (such as poor reporting, heterogeneity of results) that I identified.

I used the indirect methods suggested by Parmar et al. to increase the number of occasions an estimate of the \( \log_e(HR) \) and its variance could be obtained [84]. However, the estimates they provide are only approximate and simply make the best possible use of the results presented. Questions still exist about how best to combine indirect estimates with direct estimates. For this reason, I did not use other indirect methods available beside those suggested by Parmar et al. [97-99]. For example, given further assumptions, I could have used estimates of the proportion surviving to 2, 3, 5 or 10 years to obtain estimates of the \( \log_e(HR) \) and \( \text{var}(\log_e(HR)) \) [98]; however, the papers identified were equally inadequate at presenting these survival statistics. For example, in the 26 prognosis papers for the serum marker lactate dehydrogenase (LDH) only 12 gave actuarial estimates of the proportion surviving, and only 6 of these also gave a confidence interval or standard error. They were also heterogeneous - 5 estimates were for OS, 6 were for DFS and 1 was unspecified; estimates were made at 2, 3, 4, or 5 years. Further, very few reported numbers at risk explicitly at these time points, as required for reliable estimation [84;100].

2.5.2 Generalisations of the problems to other prognostic markers

Although these reporting problems were observed for tumour markers within the neuroblastoma literature, they have also limited reviews in other paediatric cancers [56], and it seems plausible that the reporting will be equally poor for all types of prognostic markers, for other areas of oncology, and indeed other disease areas. Altman discusses the potential problems involved in systematic reviews of prognostic markers, in particular that
of poor and heterogeneous reporting of primary studies [7]. Cut-off levels are frequently used to dichotomise continuous markers and define groups, whilst different outcomes, adjustment factors and groups of patients are common features across prognostic studies. Inadequate reporting and presentation of survival data has also been shown to be a concern in the cancer literature [101].

Reliable and clinically useful meta-analyses of observational and non-randomised studies, such as the majority of prognostic marker studies, are generally difficult to perform [102]. Other recent systematic reviews of prognostic markers have encountered similar problems to the ones I identified. Parker et al. performed a systematic review in prostate cancer to establish whether age is a prognostic marker, but the incomplete and heterogeneous nature of the reports prohibited any quantitative overview [103]. Similarly, a systematic review of prognostic laboratory variables in patients with unresected colorectal liver metastases was limited by the heterogeneity and poor quality of individual studies [104]. Zandbergen et al. performed a systematic review of biochemical markers of brain damage for identifying poor outcome in anoxic-ischaemic coma, but conclusions were limited by small sample sizes and different cut-offs and/or laboratory techniques [57]. The problem of cut-offs and unreliable quantitative results also severely limited a sensible interpretation from the meta-analysis performed by Ferrandina et al. to examine the relationship between cathepsin D (a proteolytic enzyme) and DFS in node-negative breast cancer, despite the considerable effort made in obtaining the individual study estimates [77].

Meta-analyses of prognostic markers have been facilitated when IPD were available [105-108], in particular to determine a consistent cut-off level [109]. For example, Look et al. performed a pooled analysis using IPD to assess prognostic markers uPA and PAI-1 in breast cancer, an approach that involved 18 studies across Europe and over 8000 patients
For those investigators currently interested in performing a quantitative review of prognostic markers, I recommend they consider asking authors for IPD and/or the extra information they require, such as the log_e(HR) and its variance, as this approach is likely to be the most productive. The benefits and issues associated with IPD and its use in meta-analysis are considered in further detail throughout Section 2.9.

### 2.5.3 The need to improve the quality and reporting of primary studies

'Systematic reviews of some of the hundreds of thousands of reports of trials published since 1948 are beginning to make painfully clear that, in most of these studies, inadequate steps were taken to control biases, many questions and outcomes of interest to patients were ignored, and insufficient numbers of participants were studied to yield reliable estimates of treatment effects. In brief, a massive amount of research effort, the goodwill of hundreds of thousands of patients and millions of pounds have been wasted.'

Chalmers, 1998 [110]

The fears of Altman and others about performing systematic reviews of prognostic marker studies have been confirmed by this empirical investigation in neuroblastoma, and it has also emphasised the disconcerting message given by Chalmers above [76;78;86]. In addition to some of the problems I demonstrated in the neuroblastoma review, Altman also discusses some further issues for evidence synthesis of prognostic markers studies such as poor study quality, poor study design and biased choice of cut-off levels in the original primary studies [7]. Indeed, Altman says that 'the poor quality of the published literature is a strong argument in favour of systematic reviews but simultaneously also an argument against formal meta-analysis', and that 'the main outcome from the systematic review may well be the demonstration that there is little good quality published information' [7].
Systematic reviews and meta-analyses do not always produce meaningful evidence-based results for clinical practice. However, the systematic review process is not necessarily unproductive in this situation, as it may uncover underlying problems across individual primary studies and therefore help to ascertain how future primary research can improve [37]. From my systematic review in neuroblastoma, the main finding is clearly that there is a need to improve both the standard of primary studies and the quality of how prognostic marker results are reported, so that future evidence syntheses in this area have a better opportunity to obtain important evidence-based results for clinical practice. In addition to the problems of poor study quality and poor reporting, Altman also warns that the current drive toward multiple and separate uncoordinated studies will further delay the process of clarifying the role of prognostic markers [7], with many markers remaining under investigation for numerous years after initial studies without any resolution of the uncertainty. Given this motivation to improve current standards, I will now consider suggestions and guidelines for how primary prognostic marker research should be undertaken and reported. There are three major, and inherently related, problems for primary prognostic marker studies that I will consider in turn: (i) poor quality study design, (ii) lack of suitability of the results for clinical practice, and (iii) poor quality reporting.

2.6 Suggestions for improving the quality of prognostic marker studies

As I did not directly assess the quality of prognostic marker studies in the neuroblastoma review, my discussions that follow are based predominately on previous publications which identify two specific issues that ultimately cause a problem for evidence synthesis [33;76].

2.6.1 Study quality issue 1: the purpose of prognostic marker studies

There are a number of purposes of prognostic marker studies, including those that seek to improve understanding of a disease process, those that seek to improve the design and
analysis of trials, and also those which try to develop prognostic models or strategies to be implemented in practice (Figure 2.6) [76]. However, the vast majority of prognostic marker studies have the purpose to establish the association between a single putative marker of interest to some outcome (e.g. OS or DFS), and many such studies were included in the neuroblastoma review. However, there are also some studies that do not have this primary purpose yet still report summary statistics or IPD relating prognostic markers to outcome as secondary results. Many of these were also included in the neuroblastoma systematic review.

Figure 2.6: Different purposes for prognostic marker studies [76]

- Improve understanding of the disease process.
- Improve the design and analysis of clinical trials (e.g. risk stratification).
- Assist in comparing outcome between treatment groups in non-randomised studies, by allowing adjustment for case mix.
- Define risk groups based on prognosis.
- Predict disease outcome more accurately or parsimoniously.
- Guide clinical decision making, including treatment selection, and patient counselling

Some meta-analysts may deem that, other things being equal, only those studies that were specifically designed for the purposes of evaluating one or more prognostic markers should be included in the meta-analysis (i.e. those where prognostic marker results were the primary endpoint). This issue relates to currently unanswered questions facing meta-analysts about which studies they should include in a meta-analysis [111]. For example, should they only include high quality studies and ignore those with less quality? Should they only include studies with large numbers of patients and exclude those with smaller numbers? Indeed, what factors determine a 'high quality study' and what constitutes a 'large number of patients'? These are difficult questions to answer and are the subject of much ongoing research [111-115]. Unfortunately, the assessment of prognostic marker
study quality is particularly hard because it is very difficult to distinguish those studies with different purposes, as the study aims are rarely specified in the published article and so, for example, it is hard to see whether the prognostic results were a primary or secondary objective. Indeed a study may have started off with prognostic results as a secondary aim but then, perhaps after witnessing the results obtained, may have switched the primary focus of the study to the prognostic markers. In terms of evidence synthesis, it is therefore very difficult to identify and combine those studies whose original objective were similar, and this potentially limits the credibility of the evidence-based results possible. This difficulty was the reason why I did not assess the original purpose of the primary studies in the neuroblastoma review.

2.6.2 Study quality issue 2: the type of prognostic marker studies

There are essentially four different types of prognostic marker studies, which were initially described by Simon and Altman but were then updated by Altman and Lyman (Figure 2.7) [33;76]. The first three types have been termed ‘Phase I’, ‘Phase II’ and ‘Phase III’, and they all have the aim of establishing the relationships of markers with prognosis or other prognostic characteristics of patients. Phase I studies are early exploratory analyses to identify potential prognostic markers and hypotheses that need exploring. Phase II studies follow on from these and come in two parts - those assessing the relationship between markers and prognosis, and those relating these markers to particular patients and therapy. Phase III studies are large studies which come in two forms – those confirmatory studies to validate the results of Phase II studies, and those to identify which subsets of patients are likely to benefit from therapy. The fourth, slightly different, type of prognostic marker study (denoted ‘Type IV’ here onwards) seeks to develop and validify a prognostic model to maximise the ability of combinations of markers to predict outcome for groups and individual patients.
Figure 2.7: Types of prognostic marker studies [76]

1. **Phase I**: exploratory studies (hypothesis generating) which seek an association between a prognostic marker and characteristics of disease thought to have prognostic importance.

2. **Phase II**: exploratory studies attempting to use values of a prognostic marker to:
   a) discriminate between patients at high and low risk of disease progression or death; or to
   b) indicate which subsets of patients are likely to benefit from therapy.

3. **Phase III**: confirmatory studies of a priori hypotheses attempting to use values of a prognostic marker to
   a) discriminate between patients at high and low risk of disease progression or death; or to
   b) indicate which subsets of patients are likely to benefit from therapy

4. Studies to develop a prognostic model combining many prognostic variables in an attempt to maximise the ability to predict outcome for groups and individual patients

Again the key problem for evidence synthesis is that it is very difficult to differentiate between these four types of studies, especially Phase I, II and III. For the most appropriate evidence-based results for clinical practice, one may decide to prioritise the pooling of Phase III studies, as these should be of the highest quality and therefore may contain the largest number of patients because they are confirmatory studies. However, most studies do not state in their published article what type of prognostic study they are (e.g. exploratory – Phase I, confirmatory – Phase III). It is highly likely that the majority of prognostic studies in the published literature are actually Phase I or II because, as seen in the neuroblastoma review, they often contain small numbers of patients and report a large number of results for a variety of possible markers, and therefore do not seem to be focused on making confirmatory conclusions. The idea of large-scale, high-quality Phase III studies to assess a small number of markers, that have been previously identified in Phase I and II studies, has evidently not been taken on board. This discussion emphasises further why the meta-analysis results from my neuroblastoma review should be treated...
with caution. Indeed, as most of the studies in the review will have been Phase I or Phase II studies, the meta-analysis results themselves should be treated as exploratory results and need to be confirmed further by Phase III studies, which backs up my original conclusion that the pooled results should only be used to help the prioritisation of further research.

In terms of putting evidence synthesis results into practice, the pooling of Type IV studies could arguably be the most important as it would allow evidence-based results about combinations of prognostic markers, which may be better predictors in reality than the individual markers themselves. However, even if such studies could be identified, it would again be very difficult to pool together such studies using the published results, because they are likely to have assessed different sets of prognostic markers, and in addition they may have only reported prognostic models that contained the subset of markers that were deemed important or significant. The way to limit this problem would be to obtain the full IPD from each study, so that one can develop a single prognostic model that utilised all the original data from across all studies and that considered all the (non-significant and significant) markers from each study (see Section 2.9).

### 2.6.3 Summary of improving study quality in prognostic marker studies

The need to improve study quality is clearly important for all of Phase I, II, III or Type IV prognostic marker studies, and Altman and Lyman present some very important guidelines for improving and clarifying the design and purpose of such studies [76]. Furthermore, there is a great need to move away from the thousands of Phase I and Phase II studies, which mostly contain small numbers of patients, toward a greater number of large-scale Phase III and Type IV studies. Finally, those performing primary prognostic marker studies need a greater awareness of the importance of evidence synthesis and also need to be willing to facilitate this approach. The apt guidelines suggested by Altman and Lyman are
a huge step forward to this, and to complement these I will develop additional guidelines for how to report the results of prognostic marker studies in Section 2.9.

2.7 Clinical requirements from a prognostic marker study

The second major problem for primary prognostic studies is the lack of direct implications for clinical practice that can be drawn from them. For evidence-based results to be useful, the individual primary studies themselves need to be trying to answer the questions that clinicians need addressing. If individual studies do not target such questions then it makes it even harder for an evidence synthesis of such studies to provide any clinical answers and recommendations.

The clinician has to make real-life decisions, such as treatment strategy and policies, and provide predictions of outcome to help counsel their patients. Hence, the results from prognostic studies should be enhancing these processes but in reality they often only serve as a guide to clinicians because they do not naturally tell the clinician what to do, whom to treat and whom not. This may be a problem for most Phase I to III(a) studies, and perhaps only Type IV studies truly meet the clinicians need, as they specifically develop and validate prognostic models or strategies, although even these are not always related to specific treatment strategies. When Phase I to III(a) studies show the importance of a single or set of prognostic markers they may help the clinician to identify those with poor and good prognoses, but they do not tell him/her who to treat first or indeed how [116]. In fact, the argument to treat those with the poorest prognosis first may be as correct in some situations as the argument to treat those with the best prognosis in other situations [116]. Wyatt and Altman discuss this and other problems of prognostic results and models to explain why only a few become established in current practice [117].
To truly benefit the clinician, results of prognostic studies must take into account the treatment or therapeutic strategy involved because this is the only way to learn how and whom to treat. Many prognostic marker studies, however, do not report treatment received when assessing the importance of a single or combination of markers [78]. In fact, what a clinician really wants to see are the results of treatment strategies which actually use prognostic markers, for it is this information that will help individual doctors to decide the best strategy for their own patients [116]. However, most studies of prognostic markers either just report whether there is an association with markers and outcome, or report a prognostic model without any validation of its benefit if put into practice. This brings up again the important differences between studies set up to simply identify prognostic markers and their relationship with outcome (Phase II(a) and Phase III(a) studies), and those set up to specifically consider a therapeutic strategy based on markers for clinicians to use (Phase II(b) and Phase III(b) studies) (Figure 2.7). The difference between studies in (a) and (b) is very important, because clinicians should primarily be using results from (b) as these relate to specific therapeutic strategies. Unfortunately the vast majority of prognostic marker studies are type (a). However, due to unclear reporting it is often hard to differentiate published studies of type (a) to those of type (b), and so clinicians may still try to apply results from studies of type (a) in practice, which will potentially be dangerous because their results have not been evaluated against any specific treatment strategies.

Given these points, it is highly recommended that when results from Phase II(a) or Phase III(a) studies are reported it is clearly emphasised that they provide only limited guidance on their own for the clinician.

The key point is that clinicians want to use an established prognostic marker that has been studied in relation to specific patients using specific treatment regimens. Phase II and Phase III studies of type (b) both report such results. However, this unfortunately
introduces the additional problem for the clinician of distinguishing between Phase II(b) and Phase III(b) studies, and this can be very difficult as discussed in Section 2.6.2. Unfortunately, in current practice, Phase III studies are rare because of a lack of study design, a lack of prespecified hypotheses and a reluctance to perform confirmatory work. The majority of prognostic marker studies are of the Phase II type with authors exploring and reporting associations of multiple markers with prognosis, patients and therapies. This in itself is fine so long as these Phase II studies are interpreted as ones that merely screen for important individual and combinations of prognostic markers and the results are not accepted without further confirmation from Phase III studies. However, many Phase II studies are treated and interpreted as Phase III studies, which is potentially dangerous because spurious associations are more likely given the inadequate study design and multiple testing of markers [33; 118]. It is therefore again highly recommended that authors of published primary studies should acknowledge whether their study was Phase II or Phase III in the published article. This would also help those performing an evidence synthesis of such studies and enable meta-analysis results to be kept in the proper perspective. For example, meta-analysis results from only Phase II studies should, like the primary studies themselves, only be viewed as exploratory findings to take forward in further research.

The clinician also needs to be aware of secondary reporting of prognostic marker results in other studies, where relationship of markers with prognosis were not the prior primary objective of the study and are often only presented as an ‘after thought’. Unless these results are reported with cautionary advice, clinicians may interpret them as a basis for making therapeutic decisions. This is potentially dangerous again, as there is the real threat of a mistaken result (e.g. because of multiple testing) because the study was not designed for such interpretation or conclusion. Windeler says that everything should be done to
avoid a clinical interpretation of prognostic results that were actually designed for other purposes (e.g. to develop an understanding of the disease) than specifically aiding treatment policies [116].

Clinicians and researchers both have things to learn about prognostic marker studies. Clinicians need to be aware that there are a number of different types of prognostic marker studies which are initiated for different purposes. Prognostic marker studies vary in design and aim, with many just exploratory and only relatively few actually capable of informing a clinician of how best to treat their patients. On the other hand, researchers involved in prognostic marker studies need to improve study design and quality, be aware of the requirements of clinicians, and in particular emphasise if and how the results should be used in practice. Simon and Altman present four requirements of a prognostic marker for acceptance in clinical practice [33], and these include the need to relate the marker(s) to therapeutic interventions and use conclusions based on independently confirmed Phase III studies (Figure 2.8). Hence, researchers should make clear distinctions between prognostic marker studies whose results are intended to be used in practice (e.g. Phase III(b) studies) and those studies whose results are actually only a guide or a final-step toward a clinical strategy (e.g. Phase II or Phase III(a) studies).

**Figure 2.8:** Requirements of a prognostic marker for acceptance in clinical practice [33]

- Determination is reproducible and widely available with quality control
- Substantial predictive value beyond recognised prognostic systems is demonstrated
- The predictions have therapeutic implications that are readily interpretable by the clinician and of benefit to the patient
- Conclusions are based on independent confirmed Phase III studies
Prognostic marker research needs to be collaborative and multi-disciplinary. In particular, alongside statisticians and other suitable researchers, clinicians should also be involved right from the beginning of a prognostic marker research project so that the most appropriate results can be obtained at the end of the study, with suitable inferences then made with respect to clinical practice. Availability of such clinically meaningful results from high quality studies would thereby also facilitate better evidence-based reviews, which would then also be able to answer questions of real clinical interest. Evidence synthesis cannot overcome poorly designed or poorly targeted primary studies, so improving this area is a vital first step toward evidence-based use of prognostic markers in healthcare. Researchers who are considering initiating a primary prognostic marker study should read the invaluable guidelines of Altman and Lyman, which extended those by Simon and Altman, and suggest how to appropriately design and target primary prognostic marker studies [33;76].

2.8 Why do reporting problems exist in prognostic marker studies?

The third major problem of current primary prognostic marker studies is poor reporting, and this was clearly evident in the primary studies included in the neuroblastoma review through inadequately reported summary statistics, heterogeneity of statistical/clinical factors, and the threat of dissemination bias. Even if primary studies are high quality Phase III(b) studies, if they are not adequately reported they will still be of limited use for the clinician and those performing evidence synthesis. Although guidelines have been formed about the design and targeting of prognostic marker studies, there has been very little advice on how and what results should be reported when the study is published. I now aim to address this, and to aid this process I will firstly discuss what I consider are five of the main causes of poor reporting in published prognostic marker studies.
2.8.1 Poorly designed and poorly targeted studies

The problems extensively discussed in Sections 2.6 and 2.7 are inevitably related to the problem of poor reporting at the end of a study. Most prognostic marker studies are inadequately planned or designed to answer the clinical questions that they ultimately try to answer. This causes heterogeneity in characteristics across studies, and leads to often selective reporting of only statistically significant markers and/or poor reporting of those not statistically significant. Observational studies should try to emulate the careful design standards used in clinical trials in order to produce the most reliable results [76], but this is not the case for the majority of prognostic marker studies and it causes problems for evidence synthesis.

2.8.2 Lack of collaborative research

One common problem in prognostic marker studies is small sample sizes. This is a particular problem for rare diseases, such as neuroblastoma or other childhood cancers, and leads to individual studies having low statistical power [1]. It also leads to a large number of slightly different studies of the same prognostic markers as each research group reports the results for their own small sample. This in turn leads to heterogeneity and inconsistency in, for example, cut-offs, adjustment factors and characteristics of patients included. To increase sample sizes, there is a need for large, multi-centre studies to be initiated and this requires research groups to collaborate. Furthermore, such collaboration would also facilitate an increased consistency across studies in cut-off levels, adjustment factors, and which markers were studied (see Section 2.9.7).

2.8.3 Biased reporting of results

For prognostic marker studies, there is likely to be the common problem of publication bias, where results that do not generate formal statistically significant or clinically valuable
findings may not be published. Other related dissemination bias problems may exist. For example, when non-significant or negative results are reported, they are often not given in as much detail as significant or positive results, thus making the extraction of desired summary statistics for meta-analysis more difficult (e.g. \( p > 0.05 \) is often presented). This may be a deliberate or unintentional phenomenon, often perhaps due to a reluctance of journals or researchers to report negative findings in detail. Other forms of dissemination bias may include outcome reporting bias, subgroup reporting bias and language bias, amongst other factors (see Chapter 9) [119].

One bias more specific to prognostic marker studies is in the choice of cut-off level used to dichotomise the marker and thus define a low-risk and a high-risk group of patients. The choice of cut-off level is often specifically chosen to optimise the difference between the groups and produce a result with the maximum statistical or clinical significance possible [120]. This approach leads to large heterogeneity in reporting and can confuse clinicians as to how they should use the marker for distinguishing groups of patients for different types of treatment and care. For example, at least 13 different cut-offs were used for MYCN and at least 10 different cut-offs for lactate dehydrogenase in the neuroblastoma review. The issue of how to overcome biased cut-off levels is discussed further in Section 2.9.2.

2.8.4 Lack of appreciation of evidence-based research

With evidence-based practice still coming to the forefront of medical research, most researchers (including statisticians) do not consider their studies and results in the wider context of a systematic review or meta-analysis, either when designing their studies or writing for publication. Hence, heterogeneity and inconsistencies across studies (e.g. cut-offs, adjustment factors, outcomes, summary statistics) are inevitable, as is poor or incomplete reporting of some results, particularly those not statistically significant. As
evidence-based research grows, and its importance and benefits become even more established, one would hope that it becomes more common practice for researchers to consider the larger evidence-based picture within which their individual studies will be needed. Encouraging examples of such a movement can be seen in an increased number of pooled analyses using IPD [106], the prospective registration of studies [121;122], and an increased awareness that ethics committees need to ensure trials and their follow-up results are published [123].

2.8.5 Lack of statistical knowledge

The use of statistics within medical research has grown considerably since about 1990 but many researchers, including those conducting prognostic marker studies, are still either unaware of its importance or find it a difficult subject to master. A shortage of statisticians also means that statistical advice and collaboration is often very difficult. There have been an increasing number of papers and books to facilitate the understanding of the basic statistical concepts and hopefully these will bring further fruition [43;124]. Where statistical knowledge is limited, the availability of statistical and reporting guidelines can help researchers to analyse and report results appropriately in primary studies [101;125;126], which in the wider context facilitates their inclusion within an evidence synthesis and meta-analysis. To this end, I will now consider the development of guidelines for the reporting of primary prognostic marker studies.
2.9 Towards guidelines for improved reporting of primary prognostic marker studies to facilitate future evidence-based reviews

Altman and Lyman provide some general recommendations about the analysis and reporting of prognostic marker studies, such as the need to consider bias due to missing data and to adjust results of a new marker for existing markers of recognised and accepted clinical importance (Figure 2.9) [76]. These recommendations lay the foundation for my new guidelines that will now follow, which largely concentrate on the reporting and presentation of specific summary statistics or pieces of information in the published article.

To this end my new guidelines will primarily complement point 9 in Figure 2.9, although they will go into more explicit detail, akin to Altman’s other general guidelines on the ‘presentation of results’ for survival studies (Figure 2.10) [101]. The key difference between my new guidelines and those previously published is that, in addition to seeking to improve the interpretability of the individual primary study itself, I also seek to facilitate evidence synthesis across a number of prognostic marker studies.

My new reporting guidelines are shown in Figure 2.11 [86], and they complement (and do not replace) those presented in Figure 2.9 and Figure 2.10. If the prognostic marker studies in my review had been presented as in Figure 2.11 it would have been far easier to extract information from across studies and I would not have had to make approximate, indirect HR estimates. Collaboration of research groups is required to promote such practice and achieve both the consistency and standards required. These guidelines should be applied to all markers studied, not just those that are statistically significant. In defence of some of the poor reporting observed in the studies of my systematic review, some authors may argue that the analysis/presentation of prognostic data was only a secondary part of their study. However, it is clearly important to analyse and report prognostic data to the guidelines proposed whenever or however it is studied, be it in a Phase I, II or III study.
Figure 2.9: Altman and Lyman recommendations for the analysis and reporting of primary prognostic marker studies [76]

1. Base analysis, including any hypothesis testing, on the primary and major secondary outcomes prespecified prior to the study.

2. Consider possible bias due to missing data.

3. Consider the issue of multiple comparisons when evaluating many prognostic markers or cut-off levels and adjust tests of significance accordingly.

4. Beware of the problems associated with the interpretation of stepwise multiple regression models, including model instability and likely exaggeration of the coefficient estimates and their associated p-values.

5. Adjust the effect of new prognostic markers for existing prognostic markers of recognised and accepted importance.

6. Outcome differences between subgroups should be assessed by testing the interaction between the prognostic marker and the variable defining the subgroups rather than any separate analyses within the subgroups.

7. Interpret with caution apparent outcome or prognostic marker differences between subgroups (many such differences arise from multiple testing or small sample size within subgroups).

8. Analysis of subgroups defined only during or after completion of the study should be acknowledged as exploratory.

9. In reporting the results of a prognostic marker study:
   a) Clearly state the study design:
      - Exploratory/confirmatory, prospective/retrospective, treatment (e.g. randomised or standardised), blinding, main outcomes, etc.
   b) Report the number of patients excluded because of missing data.
   c) Specify study duration including criteria for study termination (if relevant).
   d) Report methods of measurement of prognostic markers, if possible with information about reproducibility.
   e) Define clearly all study endpoints.
   f) Summarize outcomes as quantitative estimates and confidence intervals.
   g) Emphasise the outcome differences observed for all patients more than those found among subgroups.
   h) Discuss any weaknesses of the study, especially related to any subgroup analyses and multiple comparisons.
Figure 2.10: Altman et al. recommendations for the presentation of results from survival analyses in general [101].

- Give a summary of overall survival: preferably median and/or percent surviving $n$ years.
- If study is a randomised clinical trial, give separate summaries of survival for each treatment group.
- When reporting results of any test, give the test statistic, the degrees of freedom (when applicable) and the exact $p$-value.
- When presenting results of a log-rank test also report the numbers of observed and expected events in each group (desirable).
- When comparing survival in two or more groups, give an estimate of the survival in each group, e.g. median survival time, survival probabilities for a particular time point, hazard ratio.
- When presenting the results of a Cox regression analysis report the estimated coefficients (or estimated hazard ratios), measures of their precision (i.e. standard errors or confidence intervals) and/or the associated $p$-values.
- Do not use crude rates to summarise the data.

**Graphs:**
- Use meaningful time intervals.
- Use a step function to join Kaplan-Meier survival estimates.
- Mark the survival time of censored observations (desirable).
- If survival curves are reported in the same plot use different line types (desirable).
- Give number of patients at risk at selected time points (desirable).
- Mark confidence intervals or standard errors for some of the selected time points (desirable).
Figure 2.11: Riley et al. guidelines on how to report primary prognostic marker studies in order to improve current reporting standards and allow clinically useful evidence-based reviews to be made (Appendix A4) [86]. These complement those shown in Figure 2.9 and Figure 2.10.

**Guidelines for Reporting Prognostic Marker Studies**

**Objective:** To improve reporting of prognostic marker results and facilitate access to individual patient data for evidence-based reviews.

Results of all the marker analyses should be presented – both significant and non-significant results - and the following is strongly recommended:

**Essential** to present:

1) The **hazard ratio** and its **confidence interval**, or the loge(hazard ratio) and its variance. Continuous markers should be modelled as a continuous variable using appropriate methods. If there is a justifiable reason for using a cut-off level for a continuous marker it should be specified at the start of the study and clearly reported; furthermore, sensitivity analyses should also be presented to assess the robustness of the results to a variety of other cut-off levels.

2) The **number of patients** and **number of events in total**. For binary markers (and continuous markers if a cut-off level is used) also report the numbers within each group.

3) Both **unadjusted** and **adjusted** results for each marker. For adjusted results, clearly state what variables have been adjusted for. Ideally, a consistency in the set of adjustment factors used across studies should be sought through collaborative groups working toward prospectively planned pooled analyses. In particular, one should always adjust for those prognostic factors of recognised importance for current clinical practice.

4) **Individual patient data** in the paper or on the Internet, or make available with details clearly indicated within the paper. Data on markers that were not analysed should be included. Subject to any restrictions imposed by data protection laws and guidelines, include:
   - Exact initial marker level and how marker was measured
   - Time of disease recurrence (if appropriate)
   - Follow-up time
   - Final disease status
   - Levels of other existing prognostic markers of recognised and accepted importance for current clinical practice
   - All patient subgroup information available, especially age, stage of disease and type of treatment received
   - Details of inclusion/exclusion criteria would also be beneficial.

**Highly desirable** to present:

5) **Exact p-values.** Reporting of results as 'significant' or 'not significant' is insufficient. Very small p-values can be given as p < X (e.g. p < 0.0001), but in this case the exact $\chi^2$-statistic is also needed.

6) **Survival curves** showing the difference in survival over time between the groups, with clear step and censoring points; also the initial numbers in each group, and the number of events and remaining numbers at various time-points during follow-up are needed.

7) **% survival at n years** with a confidence interval using Kaplan-Meier or other methods that allow for censoring, together with the number of patients at risk at that time in each group.
My guidelines provide a basis for how to report and present both summary data and IPD. It is important that time to event is incorporated within prognostic marker analyses and thus I prefer the HR to other measures of relative risk such as the odds ratio, which relates to a fixed time-point and ignores censoring. However, in addition researchers are also advised to present the more familiar actuarial survival rate at \( n \) years, as some clinicians may find this more interpretable than the HR. When either a HR or actuarial survival is presented, its 95% confidence interval and also an exact \( p \)-value should be provided. If it is not reasonable to report the exact \( p \)-value because it is very small, say less than 0.001, then ‘\( p < 0.001 \)’ should be presented together with the exact test statistic. It is important to also explicitly define the prognostic groups being compared by reporting the specific cut-off level that distinguishes the groups. Also one should report the group sizes and the number of events in each patient group, something that was commonly lacking in the studies identified in the neuroblastoma review. Furthermore, the presentation of survival curves also needs to be improved, with each step and censoring point clearly marked on the graph, as well as the initial numbers in each group, and the number of events and remaining numbers at various time-points during follow-up [101]. Figure 1.1 demonstrates how a survival curve should be presented.

### 2.9.1 IPD and evidence synthesis of prognostic marker studies

Improved reporting of summary statistics is clearly very important, particularly as meta-analysis of survival studies have been previously beneficial when aggregated data was available [127]. However, the availability of IPD is perhaps the most viable way forward in order to produce valid and clinically useful evidence-based reviews of prognostic markers, and indeed IPD is the generally preferred option for an evidence synthesis [128;129]. For the systematic review in neuroblastoma, IPD was often available in the published articles assessed because many studies had a small sample size. However, in areas other than paediatric oncology and for more common diseases, it is likely to be less common for IPD...
to be in the published literature as the necessary publication space will often not be available. Subject to any restrictions imposed by data protection laws and guidelines, presentation or availability of full IPD using my guidelines would help overcome variability in cut-off level, type of estimate (unadjusted or adjusted), outcome assessed (OS or DFS) and adjustment factors; the study of markers in subgroups of patients (e.g. different ages, treatments) would also be easier, and it would help standardise the definition of outcomes such as DFS, rather than making assumptions as to their meaning as I did in the neuroblastoma review (see Section 2.1.5). I emphasise 'help' here because IPD will often not solve all these problems; for example, if time of disease recurrence had not been recorded in the IPD it would not be possible to assess DFS, and if continuous marker levels are presented as dichotomised then one could not then choose a consistent cut-off level across studies (see Section 2.9.2). In order to achieve a high standard, I have outlined many of the necessary details required when IPD is reported (Figure 2.11). The reporting of IPD in this way would importantly also facilitate the identification of different publications whose results relate to the same or overlapping set of patients, and it would also allow an evaluation of combinations of markers, which may produce more specific and accurate prognostic assessments than the individual markers themselves [130;131].

IPD would also enable direct estimates of the HR when data were available but not used, analysed or presented properly in the primary study. Forty-one (20%) of the 204 estimates I obtained in my systematic review were direct estimates calculated from IPD that would not have otherwise been possible. IPD would also allow model assumptions, for example proportional hazards, to be checked as necessary, and enable the baseline survival function to be estimated [132], something that is not possible if only the summary statistics are available. For my review I assumed proportional hazards was plausible across the whole follow-up period wherever a HR was available, unless otherwise indicated by the primary study (see Section 2.1.6). However, this assumption may not always be valid and, rather
than assessing the entire follow-up period, it may often be a more apt approach to assess separate time-periods throughout the duration of the study [44]. Indeed, the assumption of proportional hazards may be more plausible within each time-period. Importantly, IPD would facilitate such an assessment, and would thereby enable the benefit of a prognostic marker to be assessed over time and in relation to time-dependent characteristics such as treatment received and stage of disease. Thus, potentially more specific (e.g. in terms of treatment, time, stage of disease) prognostic marker results could be obtained from an evidence synthesis using IPD, which could be potentially invaluable for those clinicians deciding if and how to use the prognostic marker in practice [132-134].

IPD would also allow summary statistics other than the HR to be calculated if desired, and would reduce the problem of extracting estimates when inexact p-values were presented. Furthermore, if the IPD contained levels of all the prognostic markers measured (even those producing non-significant results) then the problem of dissemination bias would also be reduced, although it is unlikely that IPD would be available from every study or for every outcome, and so one would still be concerned about missing and biased reporting of data within the IPD itself. Hence, alongside improvements in primary studies, also required is the development of suitable meta-analysis methods which take into account such problems as missing outcomes, missing data and the threat of dissemination bias in order to further facilitate the most appropriate (e.g. the least biased) evidence-based results from future evidence-based reviews (see Chapters 3 to 9).

It is clearly important to include predominately direct estimates in any quantitative synthesis, and the potential for substantial differences in meta-analysis of survival data when using a number of indirect literature-based results instead of direct results from IPD has recently been shown in the head and neck cancer literature [100]. Indeed, Stewart and Parmar recommend that, whenever possible, meta-analysis using IPD is preferred because
it produces the least biased answers and the most suitable way of addressing questions that have not been or could not be resolved by individual clinical trials [128]. Availability of IPD is therefore clearly desirable, and if it is not appropriate or feasible to provide IPD within a paper itself then there is the opportunity to publish it on the Internet [135]. Of course, even making IPD available on the web is not without its problems, with the non-permanency of individual web-pages, and so perhaps a central repository to collate and manage IPD is needed within each disease area. The United Kingdom Children’s Cancer Study Group (UKCCSG) have already initiated this type of approach within paediatric oncology [136]. Authors may also wish to state in their paper that the IPD is available upon request (with contact details indicated) for those requiring it for evidence-based reviews. Of course, alongside any data protection issues, authors also needs to consider if there are any ethical issues associated with making their IPD freely available. To limit this potential problem in the future, those researchers currently involved in new primary studies should try to obtain their patients’ consent for making the IPD available at the end of the study.

2.9.2 IPD and the problem of heterogeneous cut-off levels

The use of different cut-offs makes synthesis of results particularly difficult. Of added concern is the possibility that the choice of cut-off level in a report may be specifically chosen to optimise the difference between the groups and produce a result with the maximum statistical or clinical significance possible (see Section 2.8.3) [76;120]. If there is good clinical reason to use a cut-off level then it should be specified at the start of a study and clearly reported within the results (Figure 2.11). However, Altman suggests that continuous markers should not be dichotomised because, amongst other reasons, this approach discards potentially important quantitative information and considerably reduces the power to detect a real association between the marker and outcome [7]. Hence, researchers should be encouraged to analyse and report results (e.g. HR estimates) of
continuous markers on their original continuous scale (Figure 2.11). If a cut-off level is used, it is recommended that the study report results not just for a single cut-off level but also for a wide variety of cut-off levels, so that the robustness of results to the choice of cut-off level can be assessed (Figure 2.11). Importantly, availability of IPD including exact marker levels would allow data to be re-analysed where cut-off levels were not consistent, and also where continuous marker results were desired but results using a cut-off level were given (or vice versa). Indeed, the most suitable analysis of continuous prognostic markers may require non-linear modelling techniques, as highlighted by Sauerbrei et al. [137]; researchers should consult those experienced with such techniques in this situation.

2.9.3 IPD and the problem of heterogeneous adjustment factors

It is clear that once important prognostic markers have been identified they need to be evaluated against, and also used in combination with, other known clinically useful prognostic factors, such as clinical characteristics (e.g. age, stage of disease) or other marker levels, and also in relation to specific treatment and therapeutic strategies (see Section 2.7). Prognostic marker results for specific treatment regimen that are adjusted for other known prognostic factors will have the greatest implications for clinical practice, and subsequently meta-analyses of adjusted results are the necessity. However, if authors are inconsistent in the sets of adjustment factors they use it becomes very difficult and impractical to pool results across studies and make a proper evaluation of markers over and above other factors. For the 17 adjusted MYCN estimates from my systematic review, there were 16 different sets of adjustment factors each containing one or more of age, stage of disease, Shimada index, lactate dehydrogenase and eight other prognostic markers.

Individual study estimates of risk (e.g. the HR) can be influenced by which adjustment factors are used [138], and so there may be an additional reporting bias concern if researchers specifically only report those adjusted estimates with the most statistically significant result.
Research groups should be encouraged to collaborate and identify the most commonly used prognostic tools in current practice, so that adjusted results of new prognostic marker studies can use consistent sets of adjustment factors. These identified prognostic tools should also be presented within the available IPD alongside the new markers being studied (Figure 2.11). This would allow adjusted results to be calculated independently across studies, using consistent sets of adjustment factors. Prognostic indexes could also be calculated across studies and evaluated in a meta-analysis if desired. Furthermore, Lambert et al. have shown that IPD is generally required when investigating patient level characteristics as effect modifiers in a meta-analysis, as otherwise the statistical power to detect any relationships is very low [54].

Another step in the right direction toward important adjusted prognostic marker results is in the initiation of tumour banks, which store patient tumour samples in a central repository and allows retrospective analyses [139-141]. The UKCCSG have initiated such an approach for storing tumour samples of childhood cancers [136]. This set-up allows new potentially important prognostic markers to be retrospectively analysed in comparison to the set of markers previously found to be important. Hence, if IPD were also stored for each study, a patient’s tumour sample could be reassessed and the information for the new marker placed alongside the other markers details already available in the IPD.

2.9.4 IPD and the evidence synthesis of sub-populations

IPD would also facilitate evidence synthesis of relevant subgroups across studies, and allow more specific clinical questions to be addressed (Chapter 2.7). For example, one would be able to use the IPD to identify those patients less than 1 year old in each study and then synthesise the evidence about a prognostic marker for just this subgroup.

Alternatively, one may want to assess a prognostic marker for a specific treatment and
again one could specifically analyse data for just those patients receiving this treatment across studies. The meta-analysis of subgroups is severely limited where only summary statistics are available because different studies report results for different subgroup analyses or simply analyse all subgroups (e.g. all ages, all treatments) together.

2.9.5 **Limitations of an IPD approach to evidence synthesis**

Although IPD is clearly more advantageous for evidence synthesis than summary statistics, it is important to emphasise that the availability of IPD will not overcome the problems of poor study design or poor study quality. The IPD itself is only worthwhile if it is from a study that is properly designed (e.g. to remove all potential biases) and properly targeted to study the outcomes and markers of key interest for clinical practice. Furthermore, appropriate implications for clinicians cannot be made if all the IPD are from Phase I or Phase II studies. Hence, the reporting guidelines I provide in Figure 2.11 can only make a positive difference if they are also accompanied by an improvement in study quality and design (see Sections 2.6 and 2.7). Furthermore, if the study design is not made explicit within the published article, it may be difficult to assess study quality and thus whether the IPD should or should not be used in a meta-analysis. Meta-analysts may therefore also find they need to obtain not only the IPD but also further information about how and why the study was originally performed. This may not always be straightforward.

There are additional issues for conducting IPD reviews [144], especially cost and time, and these have to be weighed against the substantial problems I encountered when trying to perform meta-analysis using summary statistics in the neuroblastoma review. For example, will the IPD available simply relate to those studies which were already the best designed, most appropriately targeted and well reported? For such decisions, cost-benefit models would be useful to help ascertain the cost-effectiveness of obtaining IPD instead of or in
addition to summary statistics, and this area would make worthy further research in the prognostic marker field, and indeed other research areas [53;74]. Of course, even when prioritising the IPD approach, the meta-analyst will in practice end-up with a mixture of estimates obtained from IPD and estimates obtained from summary statistics; hence, novel meta-analysis methods which take these different sources into account are also needed [142]. This is especially challenging for evidence synthesis of prognostic studies as opposed to RCTs, as there is a real need to make adjustments for specific patient and treatment covariates, and indeed other markers.

When first seeking IPD from authors, providing them with a clear documentation of why the IPD is essential may improve the chances of the IPD being provided. This approach has helped me to obtain IPD from the UKCCSG for another evidence synthesis project, but this may not always be the case. Of course, even if IPD were provided one must be prepared to spend time managing and collating all the IPD into a usable spreadsheet, ready for subsequent analysis. This may also involve further contact with authors about data headings, missing information, subgroup definitions and other features that need clarifying. Those planning an IPD approach need to consider these and other such issues, many of which have been discussed in detail elsewhere [128;143;144].

There is also an additional issue for those considering funding an evidence synthesis using IPD: would the money be better spent funding one new, large-scale, well-designed, and appropriately targeted primary study. For example, one particularly worrying aspect of the neuroblastoma review was that one new and suitable primary study may have answered more questions than the evidence extracted from across the majority of hundreds of studies published between 1966 and 2000. Indeed there is little benefit of obtaining IPD from previous studies if they are all of poor quality. This issue opens up the need for cost-benefit
models to help assess the benefit of funding a new primary study as an alternative to
funding an evidence synthesis of previous studies (see Sections 10.4.5 and 10.5 for further
discussion on this issue). One new Phase III(b) study may well be more expensive but as it
could be done prospectively one could ensure it met the required standards and it could be
appropriately targeted (e.g. in relation to specific treatment strategies) to answer real
clinical questions to aid practice and ultimately patient care.

2.9.6 Implementation of the guidelines and recommendations in practice

Guidelines for the design, purpose and reporting of prognostic marker studies are clearly
very important. However, there is little evidence to suggest that solely publishing such
guidelines will cause the changes in practice required. Awareness also needs to be raised
through presentation and dissemination at national and international meetings and
conferences; for example, I have recently presented the reporting guidelines of Figure 2.11
at local and national UKCCSG meetings. However, perhaps the most pivotal role in
ensuring the guidelines are adhered to is held by the Editors of and reviewers for clinical
journals. Editors can be considered as the 'gate-keepers' to publication, and they can
therefore enforce standards that an article must meet if it is to be considered for
publication. Indeed, some editors ensure certain standards are met at an early stage through
the prospective registration of trials with their journal (e.g. The Lancet). However, it may
be more difficult to take this approach with prognostic marker studies as there is usually no
intervention as such and many aspects have already been performed before the study is
initiated (e.g. tumour samples already taken and stored in a tumour bank).

Encouragingly, a number of working groups for the reporting of prognostic marker studies
have recently been initiated and they involve many leading researchers within the field,
including clinicians, statisticians and those involved with leading clinical journals. In
particular a working group involving Doug Altman and Willi Sauerbrei are currently
forming guidelines for the design and reporting of prognostic studies, which they view as
complementing existing guidelines such as the CONSORT and STARD statements
[126;145], whilst building on previously published work in the prognostic marker field
[33;76;78;86]. This group have used my reporting guidelines in Figure 2.11 to help inform
their discussions and direction, even though their focus is more on the reporting of how the
study was undertaken (e.g. presentation of cut-off, method of measurement etc.) rather than
what summary statistics or IPD should be provided in the published article. This
collaborative drive provides encouragement that steady improvements in the design and
reporting of primary prognostic marker studies will be made in the long-term.

2.9.7 Prospectively planned pooled analyses – the ultimate solution?
The need for IPD combined with the need for new studies that are properly designed and
suitably targeted suggests that evidence synthesis of prognostic marker studies may be best
achieved through prospectively planned pooled analyses [93]. I am currently unaware of
any evidence synthesis that has used this approach in the prognostic marker field. Even
those evidence-based reviews that have used IPD have all been retrospectively planned,
including the work of Look et al. [105;106], but this is perhaps not surprising given the
move toward evidence-based research is only just beginning in this field. A prospective
approach to evidence synthesis would allow multiple studies to focus on answering the
same important clinical questions whilst achieving consistency in design, markers
assessed, marker measurement, treatments received, patients included, outcomes recorded
and other pertinent information. It is therefore the most exciting way forward for obtaining
clinically relevant evidence-based results about prognostic markers, and the collaboration
of research groups to achieve this goal if perhaps the most pressing recommendation I can
make from the research undertaken in this chapter.
2.10 Summary and rationale for subsequent chapters

In this chapter I have demonstrated and discussed many of the problems associated with an evidence synthesis of prognostic marker studies by using an empirical investigation in neuroblastoma. Improvements are clearly needed at the individual study level if more suitable evidence-based reviews can be undertaken in the future, and to this end I have introduced and discussed guidelines toward an improvement in the quality of future primary studies (Appendix A4) [86]. However, even if IPD were available for every study, there is likely to remain a number of methodological issues that those performing meta-analysis need to be aware of. For example, when a meta-analysis for both OS and DFS is desired, often only one of these outcomes will be available and this may be a consequence of publication bias or other related dissemination biases. Hence, the development of suitable meta-analysis methods that take into account such problems are required to complement the guidelines developed and facilitate the most reliable (e.g. least biased) evidence-based prognostic marker results to appropriately inform clinical decision-making and policy.
Chapter 3

BIVARIATE RANDOM-EFFECTS META-ANALYSIS

AND ITS POTENTIAL ROLE IN THE EVIDENCE

SYNTHESIS OF PROGNOSTIC MARKER STUDIES

Chapter overview

In this chapter I will firstly consider why bivariate random-effects meta-analysis could help facilitate the synthesis of prognostic marker studies when two outcomes are of interest but there is the problem of missing data. A general framework for bivariate meta-analysis is then considered, firstly when IPD is available and then secondly when only summary statistics are available. I will then review previous applications of bivariate meta-analysis in the published literature, and this will identify a number of important questions that need to be addressed, setting up the rationale for the research in subsequent chapters of the thesis.

3.1 Rationale for developing bivariate meta-analysis methods in this thesis

Altman provides a list of the main issues of performing a systematic review of prognostic marker studies (Figure 3.1) [7], and the large majority have either been demonstrated or discussed in Chapter 2. Guidelines for the design and reporting of prognostic markers studies have the potential to greatly reduce these problems and facilitate clinically relevant evidence-based results of prognostic markers. However, in reality it is going to be many years before such guidelines make substantial improvements [146], and even then, given the extent of the problems observed in the neuroblastoma review, it is unlikely that all the methodological problems for evidence synthesis will disappear completely. Indeed, it is highly likely that reporting standards in prognostic marker studies may never be at a level
that allows simple meta-analysis to be performed without any possible caveats. For example, even when outcomes are reported more consistently, it is inevitable that some outcomes will remain unavailable for a proportion of studies. In reality, clinicians need to have evidence-based guidance about prognostic markers immediately, and it is not providing equity for individual patients if they have to wait until reporting standards are at such a high level to allow a perfect evidence synthesis. Hence, to impact upon current clinical practice, both now and in the foreseeable future, there is a need to develop meta-analysis methods that best utilise the data available to produce the least biased and most clinically relevant results possible.

**Figure 3.1:** Summary of the main problems with systematic reviews of prognostic marker studies from published studies [7]

- Difficulty of identifying all studies
- Negative (non-significant) results may not be reported (publication bias)
- Inadequate reporting of methods
- Variation in study design
- Most studies are retrospective
- Variation in inclusion criteria
- Lack of recognised criteria for quality assessment
- Different assays/measurement techniques
- Variation in methods of analysis
- Differing methods of handling of continuous variables (some data-dependent)
- Different statistical methods of adjustment
- Adjustment for different sets of variables
- Inadequate reporting of quantitative information on outcome
- Variation in presentation of results (e.g. survival at different time points)

The question I therefore want to consider in the remainder of this thesis is, how can one best perform meta-analysis of the results from across primary prognostic marker studies when confronted with some or all of the problems identified by the neuroblastoma review? The methodological problems restricting meta-analysis are clearly related to those points in Figure 3.1, and I have summarised some of the main ones in Figure 3.2. For this thesis, it is
obviously not possible to consider all the meta-analysis problems, and from this point onwards I am going to focus solely on the problem of performing meta-analysis where the two, potentially correlated, outcomes of OS and DFS are of interest but are not always available from every primary prognostic study to be synthesised. This relates to the first two points in Figure 3.2.

**Figure 3.2:** Summary of the main problems preventing clinically useful meta-analysis of summary statistics extracted from published prognostic marker studies.

- Missing and poorly reported outcomes and summary statistics across studies, possibly due to within-study selective reporting
- The threat of publication bias, and other related dissemination biases
- Direct estimates (e.g. of the hazard ratio) for some studies, only indirect estimates for others
- Individual patient data for some studies, only summary statistics for others
- Heterogeneous choice and use of adjustment factors across studies
- Heterogeneous and biased choice of cut-off levels across studies
- Other heterogeneity across studies in type of statistical analysis used, method of measuring the markers, treatment the patients received, and stage of disease
- Possibly overlapping sets of patients across studies
- Difficulty in assessing whether patients characteristics are effect modifiers across studies [54]
- Difficulty in assessing the benefits of a prognostic marker over time and in relation to time-dependent covariates such as treatment received and stage of disease

Missing survival outcomes and their missing summary statistics were substantial methodological problems for evidence synthesis during the neuroblastoma review (see Section 2.3.1), and the cause of the problem may unfortunately be dissemination bias (see Section 2.3.3). Such missing OS and DFS information is likely to be a relevant problem for evidence synthesis of prognostic markers even if IPD were available from every study, because there is no guarantee that both outcomes will have been recorded for the individual patient and dissemination bias may still affect what is available in the IPD itself. The study of how to perform meta-analysis of two potentially correlated summary statistics (i.e. the HR for OS and the HR for DFS) is therefore pertinent to evidence synthesis of...
prognostic markers both today and in the foreseeable future. It is also a research area applicable to other evidence synthesis situations outside the prognostic marker field, which therefore makes it a particularly sensible meta-analysis issue to consider. For example, Nam et al. performed a meta-analysis where the two correlated summary statistics of interest from each study were the log-odds ratio for developing asthma and the log-odds ratio of developing lower respiratory disease, comparing children exposed and unexposed to passive smoking, and there were a considerable number of missing log-odds ratio estimates across studies [147].

The so-called ‘bivariate meta-analysis’ methods that I will consider in the remainder of the thesis focus predominately on using summary statistics rather than IPD. This decision was primarily based on the practical issue that I had limited IPD available from the neuroblastoma review. However, for cost and other reasons (Section 2.9.5), a meta-analysis using summary statistics may be the most viable option for many future evidence synthesizes of prognostic marker studies. Hence, it is important to develop the most suitable meta-analysis methods for this situation, even if the availability and synthesis of IPD will be the ultimate goal in the long-term.

3.2 Missing and correlated OS and DFS estimates for marker MYCN

Having guidance about both OS and DFS is very important as different treatment strategies may be used before and after disease recurrence based on prognostic marker levels. However, in the neuroblastoma review only a small proportion of the primary studies presented both an OS and a DFS result, and this caused an incomplete set of HR estimates across studies for each outcome. It is likely though that OS and DFS estimates of the HR are highly correlated because a patient’s time of a recurrence of disease is likely to be associated with their time of death, especially in neuroblastoma where recurrence of disease if often quickly followed by death [22]. Furthermore, there is an inherent structural
relationship between OS and DFS, as some patient's DFS time may actually be their OS
time because DFS is defined as the time to either recurrence of disease or death. Hence,
having an estimate of the HR for OS may also be informative about the expected value of
the HR for DFS where DFS is not available, and vice-versa.

As an example, consider the OS and DFS HR estimates for MYCN obtained from the
systematic review in neuroblastoma. In this review, due to time restrictions, I only
attempted to extract HR estimates from those individual studies for which the sample size
was \( \geq 25 \) patients (see Section 2.1.6). As MYCN will be my main motivating dataset for
the remainder of this thesis, I now also extracted, where possible, HR estimates from those
studies with < 25 patients. This process gave, in addition to the 94 estimates originally
extracted during the neuroblastoma review, 13 more extracted HRs; hence, in total 107 HR
estimates for MYCN were available. However, as I only felt it appropriate to use one
estimate per outcome per study in the meta-analyses (see Section 2.4.2 for justification on
this), only 98 of the 107 estimates obtained overall for MYCN were considered and, of
these, 42 estimates were for DFS and 56 were for OS (Table 3.1). Of the 81 studies these
98 estimates represented, only 17 studies provided both an OS and a DFS HR estimate,
whilst 39 studies provided only OS and 25 studies provided only DFS (Table 3.1). In the
17 studies that provided both, there is clearly a strong linear relationship between OS and
DFS estimates, which suggests they have a highly positive correlation (Figure 3.3).

A bivariate meta-analysis model may be a suitable way to assess and synthesise these DFS
and OS HR estimates for MYCN, because such a model can incorporate the correlation
between two summary statistics [148]. Indeed, by utilising this correlation the model may
also help limit the problem of unavailable summary statistics across studies; for example,
in those studies where a DFS HR was unavailable, the model could potentially 'borrow
strength' for DFS from the correlated OS HR that was available.
Figure 3.3: Relationship between overall (OS) and disease-free (DFS) survival hazard ratio (HR) estimates from the 17 studies providing both outcomes; a standard linear regression line is shown.

Table 3.1: The 42 disease-free survival (DFS) and 56 overall survival (OS) estimates of the loge(hazard ratio) [loge(HR)] and its standard error (s.e.) for marker MYCN which were extracted from the neuroblastoma review.

<table>
<thead>
<tr>
<th>STUDY ID</th>
<th>DFS loge(HR) (s.e.)</th>
<th>OS loge(HR) (s.e.)</th>
<th>DFS loge(HR) (s.e.)</th>
<th>OS loge(HR) (s.e.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.11 (0.67)</td>
<td>-0.14 (0.81)</td>
<td>18</td>
<td>0.25 (0.29)</td>
</tr>
<tr>
<td>2</td>
<td>0.30 (0.26)</td>
<td>0.43 (0.81)</td>
<td>19</td>
<td>0.29 (0.59)</td>
</tr>
<tr>
<td>3</td>
<td>0.41 (0.82)</td>
<td>0.67 (0.29)</td>
<td>20</td>
<td>0.52 (0.41)</td>
</tr>
<tr>
<td>4</td>
<td>0.47 (0.53)</td>
<td>0.70 (0.55)</td>
<td>21</td>
<td>0.55 (0.38)</td>
</tr>
<tr>
<td>5</td>
<td>0.76 (0.49)</td>
<td>0.71 (0.63)</td>
<td>22</td>
<td>0.84 (0.26)</td>
</tr>
<tr>
<td>6</td>
<td>1.06 (0.54)</td>
<td>1.32 (0.51)</td>
<td>23</td>
<td>0.93 (0.32)</td>
</tr>
<tr>
<td>7</td>
<td>1.46 (0.41)</td>
<td>1.38 (0.37)</td>
<td>24</td>
<td>1.18 (0.57)</td>
</tr>
<tr>
<td>8</td>
<td>1.64 (0.64)</td>
<td>1.51 (0.48)</td>
<td>25</td>
<td>1.34 (0.51)</td>
</tr>
<tr>
<td>9</td>
<td>1.64 (0.64)</td>
<td>1.54 (0.52)</td>
<td>26</td>
<td>1.43 (0.37)</td>
</tr>
<tr>
<td>10</td>
<td>1.64 (0.51)</td>
<td>1.82 (0.71)</td>
<td>27</td>
<td>1.44 (1.17)</td>
</tr>
<tr>
<td>11</td>
<td>1.70 (0.39)</td>
<td>1.83 (0.47)</td>
<td>28</td>
<td>1.45 (0.57)</td>
</tr>
<tr>
<td>12</td>
<td>1.83 (0.66)</td>
<td>2.08 (0.67)</td>
<td>29</td>
<td>1.52 (0.35)</td>
</tr>
<tr>
<td>13</td>
<td>1.90 (0.46)</td>
<td>2.59 (1.04)</td>
<td>30</td>
<td>1.60 (0.49)</td>
</tr>
<tr>
<td>14</td>
<td>1.90 (0.88)</td>
<td>2.75 (1.10)</td>
<td>31</td>
<td>1.62 (0.42)</td>
</tr>
<tr>
<td>15</td>
<td>2.19 (0.42)</td>
<td>2.90 (1.10)</td>
<td>32</td>
<td>1.77 (0.46)</td>
</tr>
<tr>
<td>16</td>
<td>2.26 (1.08)</td>
<td>2.99 (0.51)</td>
<td>33</td>
<td>1.90 (0.58)</td>
</tr>
<tr>
<td>17</td>
<td>5.70 (1.73)</td>
<td>5.70 (1.73)</td>
<td>34</td>
<td>1.92 (0.34)</td>
</tr>
<tr>
<td>35</td>
<td>2.04 (0.62)</td>
<td></td>
<td>36</td>
<td>2.19 (0.35)</td>
</tr>
<tr>
<td>37</td>
<td>2.37 (1.00)</td>
<td></td>
<td>38</td>
<td>2.39 (0.73)</td>
</tr>
<tr>
<td>39</td>
<td>2.50 (0.76)</td>
<td></td>
<td>40</td>
<td>2.56 (0.55)</td>
</tr>
<tr>
<td>41</td>
<td>2.98 (0.58)</td>
<td></td>
<td>42</td>
<td>3.29 (0.50)</td>
</tr>
</tbody>
</table>

In addition, (i) 70 studies allowed neither OS nor DFS to be extracted, and (ii) unknown unpublished studies could also exist.
NA = not available
There are, however, some concerns about the missing summary statistics across MYCN studies. Firstly, is their unavailability related to dissemination bias? For example, are studies that only provide one outcome (e.g. study 18 in Table 3.1) deliberately not reporting the other outcome because it was not statistically significant at the 5% level? Furthermore, is the relationship between the two outcome HR estimates the same in those 17 studies providing both outcomes as in those other studies providing only one of the two outcomes? Also, what are the likely missing summary statistics in those other 70 known studies for which neither an OS nor a DFS HR were available, and are there some unknown studies affected by publication bias, where perhaps the study has not been published because neither of the outcomes gave statistically significant results? I raise these issues here because they will continually resurface throughout the remainder of the thesis, and they are related to whether one assumes missing data was 'missing completely at random' (MCAR), 'missing at random' (MAR) or 'not missing at random' (NMAR) (see Section 6.2).

The formal bivariate meta-analysis models I will introduce before Chapter 9 all assume missing data is either MCAR or MAR, and the plausibility of this for the MYCN and other prognostic marker datasets is considered in Section 6.3.1. However, specific methods for assessing meta-analysis results for dissemination bias, which most likely causes data to be NMAR, will be considered in Chapter 9, where there will be a particular emphasis on how a bivariate meta-analysis framework may facilitate such assessments. Furthermore, in Chapter 10 all the research performed on bivariate meta-analysis in this thesis will be placed in context for the other meta-analysis issues that remain unaddressed by the bivariate methods considered, such as the problem of heterogeneous cut-off levels and adjustment factors (see Figure 3.2).
3.3 Multiple summary statistics and multivariate meta-analysis models

For the remainder of this chapter I will consider the background to formal bivariate meta-analysis models and review how and why they have been previously applied. I will begin in this section by introducing the reasons why multiple summary statistics will often be required from each of the primary studies identified by a systematic review.

There is sometimes only one outcome of interest from each study in a systematic review, and a single univariate meta-analysis can be used to synthesise the evidence for this outcome across studies. However, many systematic reviews are interested in producing meta-analysis results for each of a number of outcomes across studies, as seen in the neuroblastoma review where OS and DFS pooled HRs were the objective [94]. Instead, or as well as multiple outcomes, the meta-analyst may also be interested in multiple treatment groups; for example, some studies assign independent groups of subjects to one of \( j \) treatments or a control group, and therefore \((j + 1)\) treatment group estimates (one for each of treatment and control groups) or \( j \) treatment effect estimates (one comparing each treatment group to the control group) are possible from each study; thus \((j + 1)\) or \( j \) separate meta-analysis results may be desired respectively [149]. Of course, the meta-analyst may also be interested in multiple outcomes for each of the multiple treatment groups, and so the scope for desiring more than one summary statistic from each study in the systematic review is large. I have classified the main situations where multiple meta-analyses may be of interest from a systematic review in Figure 3.4, and in these situations meta-analysts using the published literature will desire multiple summary statistics from each primary study identified. These situations could be extended further; for example, DuMouchel fits a ‘repeated measures meta-analysis’ model for multiple outcomes for multiple treatments for multiple cohorts at multiple time-points [150].
The need to produce multiple meta-analysis results from the same studies but for different outcomes and/or treatment groups is likely to involve the synthesis of multiple summary statistics that are correlated [148]. In order to appropriately utilise this correlation one can use a multivariate meta-analysis model, where the multiple summary statistics can be jointly modelled and their pooled values estimated within the same model.

However, there have been surprisingly few applications of multivariate meta-analysis in practice, and meta-analysts generally apply a separate univariate meta-analysis to each of the different summary statistics of interest (i.e. to each of the outcomes and treatment groups separately). This is most likely due to the increased complexity of the multivariate approach, but it may also be due to tradition and because the benefits of multivariate meta-analysis have not been clearly demonstrated.

**Figure 3.4:** Main situations where meta-analysts may desire more than one summary statistic from each primary study identified by the systematic review

<table>
<thead>
<tr>
<th>Situations where Multiple Summary Statistics are of Interest</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>(1) Multiple outcomes</strong> – One summary statistic for each outcome is desired. This is where a study reports a summary statistic (e.g. hazard ratio) for ( j ) different outcomes (e.g. OS and DFS, ( j = 2 )) but for the same patients.</td>
</tr>
<tr>
<td><strong>(2) Multiple treatment groups</strong></td>
</tr>
<tr>
<td>(i) One summary statistic for each treatment group is desired. This is where a study reports a summary statistic (e.g. log-odds) for the same outcome for each of the ( j ) treatment groups of interest.</td>
</tr>
<tr>
<td>(ii) One summary statistic for each treatment group compared to a common control group is desired. This is where a study reports a summary statistic (e.g. log-odds ratio) comparing each of the ( j ) treatment groups to a control group for the same outcome.</td>
</tr>
<tr>
<td><strong>(3) Multiple treatment groups and multiple outcomes</strong> – a combination of (1) and (2), where one summary statistic for each outcome for each treatment group is desired. For example, a study may report a summary statistic (e.g. hazard ratio), which compares a treatment group to a control group, for more than one outcome (e.g. OS and DFS) and for more than one treatment group (e.g. drug A, drug B, and drug C).</td>
</tr>
</tbody>
</table>
3.4 A hierarchy for an IPD bivariate random-effects meta-analysis

To properly understand a multivariate meta-analysis using summary statistics (rather than IPD), I think it is useful to begin by considering how one would proceed if IPD were available, especially as the hierarchical structure for a meta-analysis using summary statistics is contained within the hierarchy of an IPD meta-analysis. This will provide better understanding regarding the specific information the literature-based meta-analyst needs to extract in order to perform a multivariate meta-analysis using summary statistics. Firstly, I will consider how one would use IPD to fit a univariate random-effects meta-analysis (where only one pooled estimate is desired) and then I will extend this to a bivariate random-effects meta-analysis (where two pooled estimates are desired).

3.4.1 The hierarchy for an IPD univariate random-effects meta-analysis

Consider that IPD rather than summary statistics is available, and that a univariate random-effects meta-analysis (URMA) is desired. I consider an IPD URMA framework to have three possible levels, one relating to each of patient (k), within-study, and between-study. Level 1 represents the individual patient data (Xi) for each study from which an estimate (Yi (e.g. loge(HR)) with standard error si) of the true underlying summary statistic (θi) can be calculated for each study (i). Level 2 represents the summary statistics for the within-study data (Yi and si), expressing the variability of the Yi about the true underlying summary statistic (θi) for each study; this is the level where standard literature-based (non-IPD) meta-analysis would begin. Finally, Level 3 represents the between-study data, expressing the variability (τ2) of all the underlying summary statistics (θi) from each study about one overall pooled or average value (β). An estimate of β is the primary objective from most meta-analyses.
Consider the example of Berkey et al. [151], where the authors have a dataset of 5 different trials that study the difference in a surgical and non-surgical procedure for treating periodontal disease, with probing depth one outcome of interest. Each patient received the surgical and non-surgical procedure, each to a different area of his or her mouth. Hence, at Level 1 in this example will be the difference in probing depth \( (X_{ik}) \) between the two procedures for each patient in each of the 5 trials, and the **mean difference in probing depth** is the one summary statistic \( (\theta_i) \) of interest from each study. If IPD was available for every study, one could make an estimate \( (\hat{Y}_i) \) with some variability \( (s_i^2) \) of the true underlying value of this summary statistic \( (\theta_i) \) for each of the 5 studies. There will be 5 \( \hat{Y}_i \)s and 5 \( s_i^2 \)s at Level 2 in this example, and one can use these to average across studies and obtain an estimate at Level 3 of the true pooled or average summary statistic \( (\beta) \) across studies.

Algebraically, assuming normality at each level, for the Berkey data one can express the IPD URMA model for \( i = 1 \) to \( n \) studies and \( k = 1 \) to \( m \) patients as:

\[
\text{LEVEL 1} \quad X_{ik} \sim N(\theta_i, \nu^2) \\
\text{LEVEL 2} \quad \hat{Y}_i \sim N(\theta_i, s_i^2) \\
\text{LEVEL 3} \quad \theta_i \sim N(\beta, \tau^2)
\]

All these parameters have been defined previously except \( \nu^2 \), which represents the variation of the patient data \( (X_{ik}) \) about the underlying mean value \( (\theta_i) \) within each study.

Equation (3.1) is a multi-level model [152]. Level 2 is where the literature-based meta-analysis would begin, as one does not have the patient-level information for Level 1 and so the meta-analyst has to come straight in at Level 2. As \( \hat{Y}_i \) is an estimate of \( \theta_i \) from the IPD...
in Level 1 it would be the desired summary statistic required from each study for a non-IPD meta-analysis to proceed. For example, for each published study in the Berkey data, the literature-based meta-analyst would hope that, for the \( m_i \) patients in each study, the following REML estimates were presented in the publication:

\[
\hat{Y}_i = \frac{1}{m_i} \sum_{k=1}^{m_i} X_{ik} \quad \text{and} \quad s_i^2 = \frac{1}{m_i - 1} \sum_{k=1}^{m_i} (X_{ik} - \hat{Y}_i)^2
\]

The IPD meta-analyst does not have this concern because he or she could use the IPD to directly calculate the required summary statistic and its standard error in each study. Furthermore, as an even easier option, if IPD were available for every study then one could avoid calculating the summary statistics at Level 2 altogether and just fit a model that directly pools the IPD and goes straight from Level 1 to Level 3 in equation (3.1), i.e. the IPD model would be \( X_{ik} \sim N(\theta_i, \nu) \) and \( \theta_i \sim N(\beta, \tau^2) \). This model makes Level 2 of equation (3.1) redundant because the summary statistics for Level 2 are inherently calculated within the IPD model at Level 1 and then automatically fed into the estimation at Level 3 [153-155].

In this section I have used the Berkey data as a simple illustrative example to introduce the IPD approach to meta-analysis. However, it is important to note here that level 1 in equation (3.1) would not be appropriate for survival data, because survival times are not likely to be normally distributed and a more complex survival model (e.g. a Cox proportional hazard model [44]) which takes into account censored observations would be required to estimate the underlying hazard ratio (i.e. the summary statistic of interest) in each study. It is unnecessarily complicated to consider this in detail here, particularly as meta-analyses will not be made using IPD directly in this thesis, but further details on using IPD to fit survival models for time to event outcomes are considered elsewhere [44].
3.4.2 The hierarchy for an IPD bivariate random-effects meta-analysis

I will now extend the URMA IPD framework to a bivariate random-effects meta-analysis (BRMA) framework, where two correlated summary statistics are now of interest. Consider the Berkey example again. Alongside difference in probing depth, the authors also consider the difference in attachment level between the treatment groups; hence there were two summary statistics of interest from each study, i.e. the mean difference between treatment groups for each of the two outcomes. There are two possible sources of correlation between these two summary statistics. Firstly, there may be a within-study correlation between the two summary statistics in each study because each patient received both the surgical and non-surgical procedure (to different areas of the mouth). Secondly there may also be a between-study correlation between the two summary statistics. For example, the mean difference in probing depth may be smaller in studies involving mainly older patients than in those involving mainly younger patients, and similarly the mean difference in attachment level may also be different in these studies in a related way.

I consider the hierarchical framework for an IPD BRMA, where two correlated summary statistics are of interest from \( n \) studies, to have three possible levels, one relating to each of patient \((k)\), within-study, and between-study. As for the URMA framework, Level 1 represents the individual patient-level data \((X_{ijk})\) for each study from which an estimate \((\hat{Y}_{ij}, \text{ with standard error } s_{ij})\) of each of the \( j \) \((j = 1 \text{ or } 2)\) true underlying summary statistics \((\theta_{ij})\) can be calculated for each study \((i)\). The within-study covariance \((\lambda_i)\) between \(\hat{Y}_{i1}\) and \(\hat{Y}_{i2}\) can also be estimated to assess the correlation between the two summary statistics within each study. Level 2 again represents the summary statistics for the within-study level data \((\hat{Y}_j, s_j\) and now also \(\lambda_i\) \) expressing the variability \((s^2_{ij})\) of the \(\hat{Y}_j\)'s about the
two true underlying summary statistics ($\theta_{ij}$ and $\theta_{i2}$) in each study. This is again the level where standard literature-based (non-IPD) BRMA would begin. Finally, Level 3 again represents the *between-study level data*, expressing the between-study variability ($\tau^2_j$) and now also the between-study covariance ($\tau_{12}$) of the two underlying summary statistics ($\theta_{ij}$) from each study about the two overall pooled or average values ($\beta_1$ and $\beta_2$).

Assuming normality at each level, algebraically one can write the complete IPD BRMA model for the Berkey data as:

\[
\begin{align*}
\text{LEVEL1} \quad X_{ijk} & \sim N(\theta_{ij}, \mathbf{V}) \\
\text{LEVEL2} \quad \mathbf{Y}_i & \sim N(\mathbf{\theta}_i, \mathbf{\delta}_i) \\
\text{LEVEL3} \quad \theta_{ij} & \sim N(\beta_{ij}, \mathbf{\Omega}_2)
\end{align*}
\]

\[
\mathbf{V} = \begin{pmatrix} \nu_1^2 & \varphi \\ \varphi & \nu_2^2 \end{pmatrix}
\]

\[
\mathbf{\delta}_i = \begin{pmatrix} s_{i1}^2 & \lambda_i \\ \lambda_i & s_{i2}^2 \end{pmatrix}
\]

\[
\mathbf{\Omega}_2 = \begin{pmatrix} \tau_1^2 & \tau_{12} \\ \tau_{12} & \tau_2^2 \end{pmatrix}
\]

(3.2)

All these parameters have been defined previously except $\nu_i^2$, which represents the variation of the patient data ($X_{ijk}$) about the underlying mean value ($\theta_{ij}$) within each study for each outcome $j$, and also $\varphi$, which represents the covariance of the patient data between the two outcomes in a study. Both the IPD and literature-based meta-analysis models revert back to a separate URMA for each outcome when $\tau_{12} = 0$ and $\lambda_i = \varphi = 0$.

If IPD were available for every study, the need to directly calculate the summary statistics at Level 2 would again become unnecessary (see Section 3.4.1). However, a literature-based meta-analysis would start at Level 2 and the meta-analyst would then require $\mathbf{Y}_i$, $s_{ij}^2$ and now also $\lambda_i$ (the within-study covariance between summary statistics) to be available from all the identified studies. For example, for each published study in the Berkey data,
the literature-based meta-analyst would hope that, for the \( m_i \) patients in each study, the following REML estimates were presented in the publication:

\[
\hat{Y}_g = \frac{1}{m_i} \sum_{k=1}^{m_i} X_{yk} \quad \text{var}(\hat{Y}_g) = s^2_g = \frac{1}{m_i - 1} \sum_{k=1}^{m_i} (X_{yk} - \hat{Y}_g)^2 \\
\text{cov}(\hat{Y}_{1i}, \hat{Y}_{12}) = \lambda_i = \frac{1}{m_i - 1} \sum_{k=1}^{m_i} (X_{1ik} - \hat{Y}_{1i})(X_{12k} - \hat{Y}_{12})
\]

Of course, \( \hat{Y}_g \) and \( s^2_g \) are not always available from a published article, and this was a particular problem for the \( \log_e(\text{HR}) \) and its variance in the neuroblastoma literature (see Chapter 2). However, the additional problem for literature-based multivariate meta-analysis is that an estimate of the within-study covariance (\( \lambda_i \)) (or equivalently an estimate of the within-study correlation (\( \rho_{wi} = \lambda_i / s_{1i}s_{12} \))) between the two summary statistics of interest is rarely, if ever, reported within a published article; the calculation of these values are non-standard and often irrelevant for the main messages of each individual study. Indeed, it is a concern that this issue may severely limit the application of bivariate meta-analysis to the prognostic marker datasets from the neuroblastoma review, as the within-study correlation between the neuroblastoma OS and DFS HRs was never reported in the literature, and it is generally non-standard to calculate \( \rho_{wi} \) or \( \lambda_i \) directly even if IPD were available.

As stated for the univariate IPD model of equation (3.1) (see Section 3.4.1), level 1 in equation (3.2) would not be appropriate for survival data and a more complex model structure would be required. A frailty model would perhaps be the best approach, as this allows the survival times to two different but dependent events to be assessed in the same model [156;157]. Unfortunately, for this thesis there was limited IPD available to fit frailty and other complex survival models. The only IPD available from the neuroblastoma review was that provided in published articles, and this meant that the sample size (and

---

*R.D. Riley, Ph.D. Thesis*  
*Chapter 3*  
*Page 108*
number of events) in such data was often small and usually involved less than 30 patients. Hence, given the limited IPD available, frailty and other models for using IPD to estimate $\rho_{wi}$ between OS and DFS HRs will not be considered in this thesis. However, the issue of unavailable within-study correlation estimates for bivariate meta-analysis will be considered in Section 3.6.2 and then at length during the whole of Chapter 8. In particular, where $\rho_{wi}$ is unavailable the meta-analyst wishing to undertake a bivariate meta-analysis will appear to have three main options: (i) seek the IPD from each study to calculate $\rho_{wi}$ themselves, (ii) make some assumptions regarding the value of $\rho_{wi}$ in each study (see Chapter 8.1), or (iii) simply perform two independent URMAs and ignore the within- and between-study correlations.

3.5 Applications of an IPD bivariate meta-analysis in practice

In both the URMA and BRMA IPD frameworks in equations (3.1) and (3.2), one does not necessarily have to assume normality for $X_{ijk}$ at Level 1. For example, for each patient in the Berkey data one may be interested in the binary outcome of whether a patient obtains a 'better outcome of probing depth' ($X_{ijk} = 1$) or a 'worse or no different outcome of probing depth' ($X_{ijk} = 0$) from the surgical procedure over the non-surgical procedure. In this situation the appropriate distribution for $X_{ijk}$ would be a Bernoulli distribution.

Application to binary data has been the primary focus of the small number of published articles about IPD multivariate meta-analysis, which are generally concerned with multiple treatment groups and estimating the pooled log-odds for each group. The reason for the predominant application to binary data is that even the meta-analyst using literature-based methods can actually obtain the IPD because it is commonly available in the published article itself, in the form of (mostly $2 \times 2$) contingency tables, which give the number of
events and patients in each treatment group. This is all the information required to be able to fit an IPD (rather than summary statistics) multivariate model for binary data, and so, for example, a BRMA using IPD is readily applicable. Van Houwelingen et al. published one of the first articles about a IPD approach to BRMA using $n$ studies which each report a $2 \times 2$ contingency table denoting the number of patients in each of two treatment groups (e.g. new drug and placebo) who did or did not have an event (e.g. death) [158]. The authors take an exact likelihood approach, using the bivariate binomial distribution (formed by the $n$ patients having an event ($X_{ijk} = 1$) or not ($X_{ijk} = 0$)). For each treatment group, the IPD feeds into the likelihood the proportion having ‘disease’ in each study by

$$\hat{\pi}_j = \frac{\sum_{k=1}^{m} X_{ijk}}{m_{ij}}$$

, where $m_{ij}$ is the number of patients in the $i$th study for group $j$. Given this data, $\beta_j$, the pooled log-odds in treatment group $j$, can be estimated. The authors consider the main advantages of the bivariate framework to be the ability to obtain separate pooled estimates (i.e. the pooled log-odds) for each group as well as one pooled estimate comparing the groups (i.e. the pooled log-odds ratio). Therefore, one can also assess the relationship between the two treatments, for example by plotting the log-odds treatment against log-odds control. The model also allows other parameters of interest to be estimated, such as the mean risk difference between treatment groups (probability of outcome in treatment group minus probability of outcome in control group). McIntosh and Arends et al. both essentially use the Van Houwelingen model to investigate the underlying risk as a source of heterogeneity in treatment effects across trials [159;160], with the former preferring to use a Bayesian framework in order to account for all parameter uncertainty (see Chapter 7). Turner et al. also provide an IPD multi-level framework that corresponds to the Van Houwelingen approach assuming the log-odds in each group are normally distributed.
To estimate the parameters penalized quasi-likelihood and second-order Taylor approximations are used and these produced comparable results to the exact likelihood method of estimation by Van Houwelingen et al.

Univariate IPD meta-analysis has been considered not only for binary outcomes but also for those on ordinal and continuous scales [76;154;155]. However, equivalent IPD multivariate meta-analyses have not been applied in the literature as far as I am aware, most likely due to far more detailed IPD being required than the IPD needed for a binary data analysis. For example, the IPD for the Berkey analysis needs to contain the change in probing depth and the change in attachment level for both the surgical treatment and the non-surgical treatment for all the patients in every study. It is unclear whether Berkey et al. had such detailed IPD available, but in general it is unlikely to be presented in the published article, especially when compared to a $2 \times 2$ table, which will take up considerably less space. Similarly, with reference to evidence synthesis of prognostic markers, a BRMA using IPD to obtain pooled OS and DFS HRs will rarely be possible because each study would need to provide IPD that included quite a large amount of information (see Figure 2.11). Even if the reporting guidelines presented in Figure 2.11 make a substantial impact, it is highly unlikely that the necessary IPD would be available from every study, and so if BRMA is to be applied, a summary statistics approach is perhaps the most likely route, although of course this too may be difficult because of unavailable within-study correlation values (see Section 3.4.2).

### 3.6 Bivariate random-effects meta-analysis using summary statistics: a general framework and literature review

From this point forward I will concentrate on non-IPD multivariate meta-analysis, where one only has summary statistics available from each study rather than the raw patient data.
I am also going to focus primarily on bivariate meta-analysis because this relates directly to the situation where the HR for OS and DFS are desired from each study, which of course was the case in the neuroblastoma review of Chapter 2. For the same reasons, I will in general only refer to using the model where the two summary statistics relate to two outcomes (e.g. HR for OS and HR for DFS) rather than two treatment groups (e.g. log-odds for treatment A and log-odds for treatment B) (Figure 3.4), although the model framework is no different for either. Finally, ‘URMA’ or ‘BRMA’ denotes literature-based (non-IPD) univariate or bivariate random-effects meta-analysis respectively from this point forward, i.e. unless IPD is specifically stated, references to univariate and multivariate meta-analysis assume a standard summary statistics approach.

### 3.6.1 A general bivariate random-effects meta-analysis model

I will now outline a general model (denoted ‘Model A’) for BRMA, which essentially corresponds to a meta-analysis starting at Level 2 of equation (3.2). From a systematic review of published and unpublished studies, assume that \(n\) studies are identified and there are two related summary statistics of interest, one for each of two related outcomes. Assume that from each study \(i\) (\(i = 1\) to \(n\)) an estimate \((\hat{Y}_{ij}\) with variance \(s_{ij}^2\)) of the summary statistic \((\theta_{ij}\) of interest is available for each outcome \(j\) (\(j = 1\) to \(2\)), and also that the within-study covariance \((\lambda_i\) between \(\hat{Y}_{i1}\) and \(\hat{Y}_{i2}\) is available. I will assume that \(s_{ij}^2\) is known, which is a common assumption for meta-analysis models [53], and also that \(\lambda_i\) is known (see Section 3.6.2). Each study estimate \((\hat{Y}_{ij}\) is assumed normally distributed about the true underlying value \((\theta_{ij}\) ), which in turn is assumed normally distributed about an underlying overall pooled value \((\beta_j\) with between-study variance \(\tau_j^2\). The BRMA model can therefore be written as follows:
Model A

\[ \tilde{Y}_i \sim N(\theta_i, \delta_i) \]
\[ \theta_i \sim N(\beta_j, \Omega_j) \]

where

\[ i = 1 \text{ to } n \text{ studies} \quad j = \begin{cases} 1 & \text{for outcome 1} \\ 2 & \text{for outcome 2} \end{cases} \]

Model A is a general framework for BRMA using summary statistics, and concurs with the previous literature on this subject [147; 148; 161]. The primary interest from this model for meta-analysts will usually be \( \beta_1, \beta_2 \) and also possibly \( \beta_1 - \beta_2 \). Model A can also be re-expressed as a linear model (see Chapter 4). A fixed-effects version of Model A can be obtained by setting \( \tau_{12} = \tau_{21} = 0 \), i.e. there is no between-study heterogeneity. Model A differs from two independent URMAs by the inclusion of the known \( \lambda_i \)'s and also the between-study covariance \( (\tau_{12}) \), the latter of which is to be estimated and can also be written in terms of the between-study correlation, \( \rho_B \), by \( \tau_{12} = \rho_B \tau_1 \tau_2 \). The model reverts back to two independent URMAs for each outcome when \( \tau_{12} = 0 \) and \( \lambda_i = 0 \).

3.6.2 The problem of unknown within-study correlation values

As discussed in Chapter 3.4.2, one major problem for implementation of Model A is that the within-study correlation \( (\rho_{wi}) \) between the two summary statistics is required for each study. This is in addition to the summary statistic \( (\tilde{Y}_i) \) itself and its variance \( (s_y^2) \) for each outcome. For the example of Berkey et al. introduced in Chapter 3.4, \( \rho_{wi} \) is the within-study correlation between the mean difference in probing depth and the mean difference in attachment level; fortunately, \( \rho_{wi} \) are available from all five studies alongside \( \tilde{Y}_i \) and \( s_{y_i}^2 \) [151]. However, this situation is unlikely and it is more common that \( \rho_{wi} \) will be

R.D. Riley, Ph.D. Thesis  Chapter 3  113
unavailable, at least for some of the studies. For example, OS and DFS HRs from the same study are structurally dependent and based on overlapping sets of patients, and so it makes sense for these two summary statistics to have a non-zero (and most likely positive) within-study correlation (see Section 3.2). However, $\rho_{wi}$ was not available from any of the OS and DFS HRs from the prognostic marker studies in the neuroblastoma review.

The problem of unavailable $\rho_{wi}$'s has been highlighted many times elsewhere [147; 149; 162]. What to do when $\rho_{wi}$ is unavailable is on-going research and I will consider this issue directly in Chapter 8, which will be particularly relevant for the prognostic marker datasets. However, I want to emphasise here that there are some situations where one may plausibly assume $\rho_{wi}$ to be zero. For example, if one were interested in multiple treatment groups, rather than multiple outcomes, the summary statistics for each treatment group are likely to be uncorrelated as the patients are often different in each group (e.g. due to randomisation) and therefore the groups are likely to be independent [148; 161]. Hence, Model A has potential application in some situations even where $\rho_{wi}$ is unavailable.

### 3.6.3 A literature review of multivariate meta-analysis in practice

To understand Model A, and other related multivariate approaches, I have critically assessed the published methodological and applied literature for multivariate meta-analysis using summary statistics. In particular, I have tried to ascertain the current evidence regarding the key applications, benefits and limitations of the approach, and my findings are now discussed.
The published methodological papers about multivariate meta-analysis have focused on the three main areas:

(i) Fitting the multivariate meta-analysis model in both frequentist and Bayesian frameworks [147; 148; 153; 158; 161; 163]

(ii) Deriving parameters and inferences from the model; for example the difference between the pooled estimates \((\hat{\beta}_1 - \hat{\beta}_2)\) in Model A, or the relationship of this pooled difference to baseline risk [148].

(iii) Providing examples of situations the model can be applied to [161; 164].

Applied examples of multivariate meta-analysis are mostly found in statistical journals, and the three main areas of application are as in Figure 3.4, i.e. to multiple treatment groups [148], multiple outcomes [151; 165], and a combination of multiple treatments and multiple outcomes [161]. These applications provide an insight into the main benefits and limitations of the multivariate approach, which I will now discuss in relation to some of the main publications. Raudenbush et al. wrote one of the first papers on fixed-effects meta-analysis of multiple summary statistics (using a linear model representation of Model A without between-study heterogeneities), and introduced a generalised least squares framework to compare univariate and multivariate solutions [162]. The authors make a few cautionary remarks about their approach, including the need to ensure model assumptions such as normality and fixed-effects (rather than random-effects) are plausible. After using a similar framework, Gleser and Olkin discuss that separate univariate meta-analyses are most likely used in preference to a single multivariate approach because of their simplicity, but highlight that a disadvantage of this is loss in efficiency of parameters estimates as the correlation between responses is not being used [149]. However, they do not consider these issues in any detail.
Dear also considers a multivariate fixed-effects model but applies it to survival data at multiple follow-up points, where the summary statistic estimates (i.e. the $\hat{Y}_{ij}$s) required from each study are the proportion of patients alive at various times of follow-up [166]. The author does not have estimates of the within-study correlation ($\rho_m$) between these estimates, and so suggests an iterative method for estimating the within-study correlations between the various time-points when one only has the survival proportions available from the individual studies. The estimated correlations are then incorporated into the multivariate meta-analysis model. The author considers that the model’s main advantage is the ability to model all outcomes in a unified framework and fit multiple meta-regression models to the data, using comparisons between the models to test hypotheses about the effects of treatment on the different outcomes.

Berkey et al. use data from 44 RCTs which evaluated the effectiveness of injectable gold, auranofin and placebo on the three outcomes tender joint count, grip strength, and erythrocyte sedimentation rate [164]. They fit a fixed-effects regression model with nine responses (i.e. similar to Model A expressed in a linear model but without any between-study heterogeneity and with nine rather than two responses), with the improvement in each of the three outcomes for each of the three different treatment groups forming the response variable ($\hat{Y}_{ij}, j = 1 \text{ to } 9$). All the meta-analyses indicated that gold was significantly better than auranofin on all three treatment outcomes, even though the individual trials reported no statistically significant differences. The authors state that the major benefit of the model proposed is its ability to incorporate trials that report all outcomes and also those that report only one or two, due to the very nature of the multi-level model framework used (missing data will be considered in detail in Chapter 6). This idea of ‘borrowing strength’ from all available outcome estimates led to all pooled estimates having smaller standard error in the multivariate model than in the three separate
URMAs. Interestingly, as the majority of within-study correlations were available, the authors also assess the robustness of their conclusions to different imputed values of the within-study correlation.

Van Houwelingen et al. were the first to introduce Model A as the general mixed model framework for BRMA [148], and also extend the model to meta-regression, i.e. where additional covariates are included to explain the between-study heterogeneities. The authors focus primarily on multiple treatment groups (rather than multiple outcomes) and discuss the main advantages of the model to be the ability to calculate pooled log-odds ($\hat{\beta}_j$) in each group, which provides the potential to examine the difference between treatment and baseline (placebo) risk, and subsequently calculate a pooled log-odds ratio $\left(\hat{\beta}_1 - \hat{\beta}_2\right)$. They also highlight how the approach can help assess the shape of the bivariate relation between the true log-odds ratio and the baseline risk, or simply the true log-odds in the treatment group compared to baseline risk (from the control group).

Nam et al. present a Bayesian formulation of Model A (see Chapter 7) and their work is the most in-depth Bayesian analysis for BRMA published so far [147]. In their applied example the two summary statistics of interest from each study were the log-odds ratio for developing asthma ($\tilde{Y}_1$) and the log-odds ratio of developing lower respiratory disease ($\tilde{Y}_2$), comparing children exposed and unexposed to passive smoking. Their dataset also contains the largest amount of missing data across all the publications I reviewed, with 51 of 59 studies only reporting one of the two outcomes. Interestingly, the authors again do not have $\rho_{wi}$ available, which represents the within-study correlation between the estimate of the log-odds ratio for asthma and the estimate of the log-odds ratio for lower respiratory disease. To help limit this problem, the authors perform sensitivity analyses using a range
of different prior distributions for the within- and between-study correlations. The values of the parameter estimates scarcely change regardless of the prior distributions used, except for the correlations themselves. However, the authors are cautious about this being a general result and they do not report the impact on the standard errors of the pooled effects. However, they do conclude that, for all the analyses considered, the overall BRMA results were practically identical to the URMA results and they suggest that BRMA appears to be most beneficial over URMA when the datasets are small and when some outcomes are missing at random (see Section 6.2).

In a rare application in a non-statistical journal, Glas et al. apply a BRMA to a systematic review of tumour markers used for the diagnosis of primary bladder cancer, where the sensitivity and specificity were of primary interest [167]. For each marker identified, the authors model the logit-transformed sensitivity ($\tilde{Y}_{ni}$) and logit-transformed specificity ($\tilde{Y}_{i2}$) values from each study as a bivariate normally distributed response, and thereby account for the correlation between the two outcomes. The same approach is used elsewhere [168]. However, in both of these publications, the authors have not detailed whether Model A or an alternative parameterisation is used, and they do not discuss any assumptions or report any results for $\rho_{wi}$, the within-study correlation between $\tilde{Y}_{ni}$ and $\tilde{Y}_{i2}$. I understand that a more technical paper by these authors will be published in the near future on this subject.

In one of the most recent and novel applications of BRMA, Thompson et al. apply BRMA models to genetic studies of coronary heart disease that use Mendelian randomisation, with the bivariate outcome of interest the genotype-phenotype association and the genotype-disease association [169]. The authors consider two slightly different parameterisations of BRMA in order to model these two associations and to subsequently derive information...
about the phenotype-disease association. Firstly Model A is used, and \( \hat{Y}_{i1} \) is an estimate of log-odds ratio of disease given genotype and \( \hat{Y}_{i2} \) is an estimate of the mean change in phenotype given genotype from each study \( i \). The within-study correlation (\( \rho_{wi} \)) between these summary statistics was assumed zero for each study because the difference in phenotype is often measured in a subset of the total number of subjects and the log-odds ratio of disease given phenotype is based on aggregate statistics. The key objective from their meta-analysis is to estimate the pooled or average ratio (\( \alpha = \beta_1 / \beta_2 \)) of these two estimates as this provides the log-odds ratio of the effect of phenotype on disease.

Interestingly, the authors report problems in estimating the between-study correlation from Model A, in particular the between-correlation is often estimated as 1. The authors suggest that further research into this problem is required, but to overcome it they use an alternative parameterisation of Model A at the between-study level, as follows:

\[
\begin{align*}
\theta_j & \sim N\left( \begin{pmatrix} \alpha \beta_2 \\ \beta_2 \end{pmatrix}, \Omega_2 \right) \\
\Omega_2 & = \begin{pmatrix} \beta_1^2 \tau_a^2 + \alpha^2 \tau_2^2 & \alpha \tau_2^2 \\ \alpha \tau_2^2 & \tau_2^2 \end{pmatrix}
\end{align*}
\]

The difference in this formulation from Model A arises from modelling the heterogeneities on the genotype-phenotype and phenotype-disease stages (rather than genotype-phenotype and genotype-disease) and critically assuming they are independent. This assumption means that in studies that find a large effect of genotype on disease, they will not tend to find relatively larger or smaller effects of that phenotype on disease. Importantly, the between-study correlation is still induced in the resultant heterogeneities on genotype-phenotype and genotype-disease, as can be seen in \( \Omega_2 \) above, but it does not have to be estimated directly. Although this reparameterised model works well for this special situation presented by Thompson et al, it is unlikely to be applicable in general, because strong within- and between-study correlation assumptions are required [169].
as far as I am aware, the only published work that explicitly uses a formal multivariate
meta-analysis for making indirect comparisons; for example, I am surprised not to have
seen the model used for making an indirect comparison of treatment A versus treatment B
given estimates of A versus treatment C and B versus C. The lack of literature may be
related to the fact that most primary studies only report one of A versus C or B versus C,
and therefore there is no opportunity to estimate either the within- or between-study
correlation between treatment-effects.

3.7 Summary and rationale for subsequent chapters

From reviewing the literature there have also been some recurring reasons why authors are
choosing multivariate meta-analysis rather than univariate meta-analysis, and these reasons
can be summarised as:

(i) It provides a single unified model within which all the pooled values of interest
can be estimated (e.g. $\beta_1$ and $\beta_2$ in Model A).

(ii) It allows the correlation between the related summary statistics to be
incorporated in the estimation of the pooled values. This may allow the
'borrowing of strength' when some studies do not report all the summary
statistics of interest (i.e. there is missing data), and increased precision of
parameter estimates.

(iii) It allows one to estimate the difference between pooled values (e.g. $(\beta_1 - \beta_2)$ in
Model A). For example, if one estimated the pooled log-odds for a treatment
group and the pooled log-odds of a control group, the model can also easily
estimate the difference between pooled values, i.e. the log-odds ratio.

(iv) As it allows one to calculate $(\beta_1 - \beta_2)$ it therefore also facilitates the use of the
model for making indirect comparisons [169].
(v) It allows one to assess baseline risk. For example, one can assess the shape of the bivariate relation between the true log-odds ratio \((\beta_1 - \beta_2)\) and the baseline risk \((\beta_1)\), which is simply the log-odds from the control group [148].

(vi) It allows one to perform joint hypothesis tests relating to some or all of the pooled values, as one can compare each of the pooled values estimated; for example one may investigate a hypothesis \(H_0: \text{all the pooled values are equal (e.g. } (\beta_1 - \beta_2) = 0)\).

For the remainder of the thesis I am going to concentrate on the BRMA approach of Model A in relation to (i) to (iii) above. In particular, I do not consider baseline risk any further, as this has been considered extensively elsewhere [159;170-173]. From reviewing the literature, it is clear that there are some particular methodological issues that need to be addressed regarding Model A, and an assessment of both the benefits and limitations of the approach is needed in a variety of contexts to bridge the gap between theory and application. In particular, there is a need to:

- **Obtain and understand analytic solutions (see Chapter 4).** The literature provides detailed estimation procedures but none provide or investigate analytic solutions for the parameter estimates of interest in Model A (i.e. \(\tilde{\beta}_1\) and \(\tilde{\beta}_2\)), and do not compare them to those from a URMA. This is important to help meta-analysts understand why BRMA results may differ from URMA results, and thereby facilitate more coherent inferences from the results obtained.

- **Consider the benefits and limitations of using Model A when one is interested in the pooled estimates themselves (see Chapter 4, 5 and 6).** For example, does one improve the mean-square error (MSE) and precision of the pooled estimates over those from a URMA? Previous work has focused mainly on the difference between pooled estimates \((\beta_1 - \beta_2)\), in relation to baseline risk, however in other
situations (e.g. synthesis of OS and DFS HRs across prognostic marker studies) \( \hat{\beta}_1 \) and \( \hat{\beta}_2 \) themselves will be of primary interest.

- **Identify under what circumstances Model A should be preferred to two independent univariate analyses (see Chapter 4, 5 and 6).** For example, should one always be seeking to use Model A rather univariate meta-analyses, or are there limitations in some situations?

- **Assess the benefits and limitations of Model A for both complete-case data (see Chapter 5) and missing data situations (see Chapter 6).** There has currently been very little application to missing data situations, where not all the summary statistics of interest are available for every study.

- **Provide an explanation regarding the problem of poorly defined estimation of the between-study correlation (see Chapter 5),** as suggested by Thompson et al. [169].

- **Consider any additional benefit of a Bayesian framework for Model A over a frequentist approach, and to consider the impact of prior distributions in general but especially for the between-study correlation (see Chapter 7),** as suggested by Nam et al. [147].

- **Explore how Model A can be applied when the within-study correlations are unavailable, and whether the approaches suggested are ultimately beneficial over two independent URMAs (see Chapter 8).** Such work has so far been limited [147;166], even though it is perhaps the most pressing methodological hurdle for using Model A in practice. It is also a particularly relevant issue for bivariate meta-analysis of OS and DFS HRs, for which the within-study correlation will rarely be available.
• Consider reparameterisation of Model A to allow a bivariate meta-analysis of OS and DFS HRs (see Chapter 8), including those from the neuroblastoma review, something that has not yet been considered in the literature.

• Consider if and how Model A can be applied when missing data is ‘not missing at random’ (Chapter 6), and if a bivariate meta-analysis framework can be useful for aiding an assessment of dissemination bias (see Chapter 9). The issues of different types of missing data and the threat of dissemination bias have currently received little attention in a bivariate meta-analysis setting.

By tackling these issues in the forthcoming chapters, I aim to clarify the merit or otherwise of Model A for those meta-analysts dealing with evidence synthesis where two correlated summary statistics are desired from each study. In particular, the ultimate objective from my work is to ascertain if, and how, it is appropriate to apply BRMA to the OS and DFS HRs from the neuroblastoma review. I have already alluded that this is not straightforward because of unavailable within-study correlations (see Sections 3.4.2 and 3.6.3), but, for example, application of Model A would certainly be desired if it facilitated the least biased evidence-based results from future prognostic markers reviews. This issue will be dealt with directly in Chapter 8, with specific consideration made to MYCN and the other potentially important prognostic markers in neuroblastoma. To reach this goal it is clearly important to firstly ascertain the benefits and limitations of Model A when the within-study correlations are actually known (or could validly be assumed zero), which will also make the research findings generally relevant beyond the prognostic marker field. I therefore emphasise here that the research in Chapter 4 to 7 will all assume that the within-study correlations are known (or could all be validly assumed zero). It is also worth adding that although my focus is on a bivariate random-effects model, the extension of my findings to a bivariate fixed-effects model, and to trivariate and other multivariate meta-analyses will also be made (see Section 6.7).
Chapter 4

AN ANALYTICAL ASSESSMENT OF MODEL A, WITH CRITICAL COMPARISON TO UNIVARIATE RANDOM-EFFECTS META-ANALYSIS

Chapter overview

In this chapter I will begin my critical assessment of Model A by deriving analytic solutions for the unknown parameters in Model A and comparing them with those from the univariate random-effects meta-analysis (URMA) which were derived earlier (see Section 1.7.2 and equations (1.4) and (1.5)). The Berkey et al. dataset will be used to demonstrate some of the findings [151], and throughout the chapter the assessments will be placed in the context of how Model A may facilitate meta-analysis of prognostic marker studies.

4.1 Estimation using restrictive iterative generalised least squares

To estimate the unknown parameters in Model A (i.e. $\beta_1$, $\beta_2$, $\tau_1^2$, $\tau_2^2$, $\tau_{12}$ in equation (3.3)) one can use restrictive iterative generalised least squares (RIGLS), which is equivalent to restricted maximum likelihood (REML) when a normal error structure is assumed [51]. For RIGLS, it helps to express the model in matrix form, using a design matrix $X$ for the fixed effects $\beta$ and a design matrix $Z$ for the random-effects $u$, where

$$\theta_y = \beta + u,$$

as follows:

$$Y = X\beta + Zu + e.$$ (4.1)
where,
\[
e \sim MVN(0, \Omega_1), \quad \Omega_1 = \text{diag} \begin{bmatrix} s_{11}^2 & \lambda_1 & \cdots & 0 \\
 \lambda_1 & s_{12}^2 & \cdots & 0 \\
 \vdots & \vdots & \ddots & \vdots \\
 0 & 0 & \cdots & \lambda_n \end{bmatrix}, \quad \Omega_2 = \text{diag} \begin{bmatrix} \tau_{11}^2 \\
 \tau_{12} \\
 \vdots \\
 \tau_{12} \\
 \tau_{22} \end{bmatrix}
\]

and \( \mathbf{u} = \begin{bmatrix} u_{11} \\
 u_{12} \\
 u_{21} \\
 u_{22} \\
 u_{n1} \\
 u_{n2} \end{bmatrix} \sim MVN(0, \Omega_2) \).

N.B. The parameters in \( \Omega_1 \) are assumed known, 'diag' indicates the matrix is diagonal in the square brackets with every other cell equal to zero, and 'I_{2n}' is the 2n by 2n identity matrix.

Let \( \mathbf{V} \) be the covariance matrix of \( \mathbf{Y} \), which is the sum of \( \Omega_1 \) and \( \Omega_2 \) as follows:

\[
\mathbf{V} = \text{var}(\mathbf{Y}) = \text{diag} \begin{bmatrix} \tau_{11}^2 + s_{11}^2 & \tau_{12} + \lambda_1 \\
 \tau_{12} + \lambda_1 & \tau_{22} + s_{12}^2 \\
 \vdots & \vdots \\
 \tau_{12} + \lambda_n & \tau_{22} + s_{22}^2 \\
 \tau_{12} + \lambda_n & \tau_{22} + s_{22}^2 \end{bmatrix}
\]

(4.2)

The RIGLS procedure minimises \( (\mathbf{Y} - \mathbf{X}\beta)^T \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{X}\beta) \) with respect to \( \beta \) by differentiating with respect to \( \beta \) and setting the first derivative to zero; this gives

\[-2\mathbf{X}^T \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{X}\beta) = 0.\]

Thus, rearranging obtains \( \hat{\beta} = \left(\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{Y} \), where \( \hat{\beta} \) is the GLS estimator of \( \beta \) and is also the best linear unbiased estimate of \( \beta \). The covariance matrix of \( \hat{\beta} \) can also be estimated by \( \left(\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X}\right)^{-1} \) [46]. Alongside \( \beta_1 \) and \( \beta_2 \), the between-study variance parameters (i.e. \( \tau_{11}^2 \), \( \tau_{12}^2 \) and \( \tau_{12} \)) are also unknown and therefore have to be estimated simultaneously. Let \( \mathbf{Y}^* = \overline{\mathbf{y}}^T + \mathbf{X}(\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \), where

\( \overline{\mathbf{y}} = \mathbf{Y} - \mathbf{X}\hat{\beta} - \Omega_1 \) are the between-study residuals and \( (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \) can be thought of as a correction to obtain unbiased covariance parameter estimates. That is, \( (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \) is...
the 'restricted' part of RIGLS, which is not a part of standard GLS; a simple example is the
RIGLS variance estimate of a sample mean which is equal to \( \sum (x_i - \bar{x})^2 \) divided by \( (n - 1) \),
where the restricted part can be seen in \'(n - 1)' rather than simply 'n', which would have
been obtained from a standard GLS estimation [51].

Let \( \mathbf{Y}'' \) be a single column vector formed by stacking the \( 2n \) columns of \( \mathbf{Y}^* \) on top of one
another in succession. Goldstein shows that \( E(\mathbf{Y}''') = \mathbf{\Omega}_2 = \mathbf{Z}^{'} \mathbf{\Psi} \) say, where \( \mathbf{Z}^{'} \) is the
design matrix for the variance components in \( \mathbf{\Omega}_2 \) and \( \mathbf{\Psi}^T = (\tau_1^2, \tau_2^2, \tau_{12}) \) [174]. The GLS
estimate for \( \mathbf{\Psi} \) is \( \hat{\mathbf{\Psi}} = (\mathbf{Z}^{'} \mathbf{\Psi}^{'+1} \mathbf{Z})^{+1} \mathbf{Z}^{'} \mathbf{\Psi}^{'+1} \mathbf{Y}''' \), where \( \mathbf{\Psi}^{'} \) is the covariance matrix of \( \mathbf{Y}''' \)
and is defined by \( \mathbf{\Psi}^{'} = \mathbf{v} \otimes \mathbf{v} \) where \( \otimes \) is the Kronecker product which multiplies every
element of the left hand matrix by each element of the right hand matrix [174].

The RIGLS procedure iterates between calculating \( \hat{\beta} = (X^T \Psi^{-1} X)^{-1} X^T \Psi^{-1} \mathbf{Y} \) and
\( \hat{\mathbf{\Psi}} = (\mathbf{Z}^{'} \mathbf{\Psi}^{'+1} \mathbf{Z})^{+1} \mathbf{Z}^{'} \mathbf{\Psi}^{'+1} \mathbf{Y}''' \), with the parameter estimates from the previous iteration
used to obtain updated estimates at the current iteration until convergence is obtained (e.g.
to 6 decimal places) between successive iterations for both \( \hat{\beta} \) and \( \hat{\mathbf{\Psi}} \).

4.2 Fitting Model A in SAS Proc Mixed using Cholesky decomposition

There has been detailed discussion elsewhere on how to fit Model A using SAS Proc
Mixed, and I include the syntax code in Appendix B1 [148]. However, I want to emphasise
a model fitting point that has only briefly been discussed before [148], and will become
increasingly important from Chapter 5 onwards. From the data one observes the total
variance between the \( \hat{Y}_{1i} \)s and the total variance between the \( \hat{Y}_{12} \)s, and the specification of
Model A decomposes these total variances into the sum of the within- and the between-
study variances for each outcome. Similarly, the total covariance between the $Y_{i1}$s and $Y_{i2}$s observed from the data is split into the sum of the within-covariance and the between-covariance. This can be seen in equation (4.2) by the formulation of the parameters in $V$.

To satisfy the conditions of a GLS procedure, one needs to ensure that the variance matrices used in the estimation are all positive definite. Essentially this means that the diagonal sub-matrices of $V$ are all positive definite, and therefore that the total covariance in each is defined properly so that the total correlation is $\leq 1$ or $\geq -1$. The within-study values are known but the between-study parameters (i.e. $r_1^2$, $r_2^2$ and $r_{12}$) have to be estimated. However, because the total variances and total covariance is split into within- and between-study values, the between-study parameters themselves are not naturally defined to limit the between-study correlation to be $\leq 1$ and $\geq -1$. Hence, the RIGLS procedure may seek a between-study correlation estimate outside this restricted range and one may obtain negative definite matrices, misleading answers or, most likely, non-convergence of the parameter estimates. This will become an increasing problem the more poorly defined the between-study parameter estimates are (for example when there are a small number of studies, or if the $s^2_{ij}$ are much larger than the between-study variances – see simulation results in Chapter 5).

To overcome this problem, one can use Cholesky’s Factorisation Theorem [175;176], which states that if a matrix $M$ is positive definite then there exists a unique lower triangular matrix $L$ with positive diagonal entries such that $M = LL^T$. For Model A, one requires $\Omega_2$ to be positive definite. Hence, for this to be the case, there must exist a unique lower triangular matrix $L$ such that $\Omega_2 = LL^T$. 

---

*R.D. Riley, Ph.D. Thesis Chapter 4* 127
Let $L = \begin{pmatrix} \alpha_{11} & 0 \\ \alpha_{12} & \alpha_{22} \end{pmatrix}$, then $\Omega_2 = \begin{pmatrix} \tau_1^2 & \tau_{12} \\ \tau_{12} & \tau_2^2 \end{pmatrix} = \begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{12} & \alpha_{22} + \alpha_{11} \end{pmatrix}$. If $\Omega_2$ is specified in this way in Model A, i.e. in terms of $\alpha_{11}$, $\alpha_{12}$ and $\alpha_{22}$, then one can subsequently obtain estimates of the original parameters of interest by:

$$
\tau_1^2 = \alpha_{11}, \quad \tau_{12} = \alpha_{12}, \quad \tau_2^2 = \alpha_{12} + \alpha_{22}
$$

Most importantly, this reparameterisation ensures that $\Omega_2$ is always non-negative definite because $\alpha_{11}^2 \alpha_{12}^2 \leq \alpha_{11}^2 (\alpha_{12}^2 + \alpha_{22}^2)$, and therefore $\tau_{12}^2 \leq \tau_1^2 \tau_2^2$ and the modulus of the between-study correlation is therefore always $\leq 1$. This reparameterisation for $\Omega_2$ is easily implemented in SAS by simply specifying type = $fa0(2)$ after the RANDOM statement, rather than type = uns (i.e. $\Omega_2$ is unstructured) as has been most commonly suggested in the literature (Appendix B1) [148].

4.3 Analytic solutions for the unknown parameters in Model A

In order to assess the difference between Model A and a URMA, and to understand why one may obtain different estimates and conclusions, it is important to obtain the analytic solutions to $\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y$ in Model A. The mathematical details to obtain the solutions are shown in Appendix B2, and for simplicity I only show the final 'simplified' equations below. Consider that there is complete-case data, i.e. $\hat{Y}_{i1}, \hat{s}_{i1}^2, \hat{Y}_{i2}$ and $\hat{s}_{i2}^2$ are available for every study. In this situation, the analytic solutions for the pooled estimates for outcome $j = 1 (\hat{\beta}_1)$ and outcome $j = 2 (\hat{\beta}_2)$ at each iteration of the RIGLS estimation are:

$$
\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \left( \frac{\hat{Y}_{i1}}{(\tau_1^2 + \hat{s}_{i1}^2)(\tau_1^2 + \hat{s}_{i2}^2)} - (\tau_{12} + \lambda) \right) \left( \frac{\hat{Y}_{i1}^2 + \hat{s}_{i1}^2}{(\tau_1^2 + \hat{s}_{i1}^2)(\tau_1^2 + \hat{s}_{i2}^2)} - (\tau_{12} + \lambda) \right)}{\sum_{i=1}^{n} \left( \frac{\hat{Y}_{i1}^2 + \hat{s}_{i1}^2}{(\tau_1^2 + \hat{s}_{i1}^2)(\tau_1^2 + \hat{s}_{i2}^2)} - (\tau_{12} + \lambda) \right) \left( \frac{\hat{Y}_{i1}^2 + \hat{s}_{i1}^2}{(\tau_1^2 + \hat{s}_{i1}^2)(\tau_1^2 + \hat{s}_{i2}^2)} - (\tau_{12} + \lambda) \right)}
$$

$$
\hat{\beta}_2 = \frac{\sum_{i=1}^{n} \left( \frac{\hat{Y}_{i2}}{(\tau_2^2 + \hat{s}_{i2}^2)(\tau_1^2 + \hat{s}_{i1}^2)} - (\tau_{12} + \lambda) \right) \left( \frac{\hat{Y}_{i2}^2 + \hat{s}_{i2}^2}{(\tau_2^2 + \hat{s}_{i2}^2)(\tau_1^2 + \hat{s}_{i1}^2)} - (\tau_{12} + \lambda) \right)}{\sum_{i=1}^{n} \left( \frac{\hat{Y}_{i2}^2 + \hat{s}_{i2}^2}{(\tau_2^2 + \hat{s}_{i2}^2)(\tau_1^2 + \hat{s}_{i1}^2)} - (\tau_{12} + \lambda) \right) \left( \frac{\hat{Y}_{i2}^2 + \hat{s}_{i2}^2}{(\tau_2^2 + \hat{s}_{i2}^2)(\tau_1^2 + \hat{s}_{i1}^2)} - (\tau_{12} + \lambda) \right)}
$$

R.D. Riley, Ph.D. Thesis Chapter 4
\[
\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \left( \hat{y}_i \right) \left( \hat{y}_i \right) - \left( \hat{y}_i \right)^2}{\sum_{i=1}^{n} \left( \hat{y}_i \right)^2 - \left( \hat{y}_i \right)^2}
\]

(4.4)

, where \( k = 1, \ldots, n \) representing the \( n \) studies, and \( k \) is used to distinguish the summation from \( 1 \) to \( n \) within the summation for \( i = 1, \ldots, n \).

Incidentally, the analytic solutions in equations (4.3) and (4.4) are the same whether ‘restricted’ or standard generalised least squares is used; it is only the analytic solutions for \( \hat{\tau}_1^2 \), \( \hat{\tau}_2^2 \), and \( \hat{\tau}_{12} \) that are altered by using the ‘restricted’ approach, and these will be discussed shortly. Firstly, at each iteration of the RIGLS procedure one can use the diagonal entries of \((X'V^{-1}X)^{-1}\) to obtain the analytic solutions for the variance of the pooled estimates:

\[
\text{var}(\hat{\beta}) = \frac{\sum_{i=1}^{n} \left( \hat{y}_i \right) \left( \hat{y}_i \right) - \left( \hat{y}_i \right)^2}{\sum_{i=1}^{n} \left( \hat{y}_i \right)^2 - \left( \hat{y}_i \right)^2}
\]

(4.5)

\[
\text{var}(\hat{\beta}) = \frac{\sum_{i=1}^{n} \left( \hat{y}_i \right) \left( \hat{y}_i \right) - \left( \hat{y}_i \right)^2}{\sum_{i=1}^{n} \left( \hat{y}_i \right)^2 - \left( \hat{y}_i \right)^2}
\]

(4.6)

Also, using the off-diagonal entries of \((X'V^{-1}X)^{-1}\), the covariance between the pooled estimates at each iteration is:

\[
\text{cov}(\hat{\beta}_1, \hat{\beta}_2) = \frac{\sum_{i=1}^{n} \left( \hat{y}_i \right) \left( \hat{y}_i \right) - \left( \hat{y}_i \right)^2}{\sum_{i=1}^{n} \left( \hat{y}_i \right)^2 - \left( \hat{y}_i \right)^2}
\]

(4.7)
As mentioned, \( \hat{\tau}^2_1, \hat{\tau}^2_2 \), and \( \hat{\tau}_{12} \) are also estimates, even though they are assumed known in the above solutions (N.B. a Bayesian approach is one possible option if one does want to take into account their uncertainty – see Chapter 7), and the RIGLS estimation procedure iterates between estimating the parameters in \( \beta \) (i.e. \( \beta_1 \) and \( \beta_2 \)) and \( \Omega \) (i.e. \( \tau^2_1, \tau^2_2 \), and \( \tau_{12} \)) until the estimates for each have converged to a pre-specified level (e.g. 6 decimal places). The analytic solutions for \( \tau^2_1, \tau^2_2 \), and \( \tau_{12} \) are very complex and unnecessary to discuss in detail here, except to say that all the solutions involve all the parameters from both \( j = 1 \) and \( j = 2 \). More detailed explanation about exactly how RIGLS estimates these between-study parameters is given in Appendix B3.

The analytic solutions from a URMA can be found by setting \( \hat{\tau}_{12} = 0 \) and \( \lambda_i = 0 \) in equations (4.3) to (4.7); for example, for \( j = 1 \) this produces the following URMA solutions that were originally shown in equations (1.5) and (1.6) of Section 1.7.2:

\[
\hat{\beta}_1 = \frac{\sum \left[ \frac{\tilde{Y}_{i}}{\tilde{\tau}^2_i + s_i^2} \right]}{\sum \left[ \frac{1}{\tilde{\tau}^2_i + s_i^2} \right]} \quad \text{and} \quad \text{var} \left( \hat{\beta}_1 \right) = \left( \sum \frac{1}{\tilde{\tau}^2_i + s_i^2} \right)^{-1}
\]

4.4 Weighting of each study in the analytic solutions from Model A

For a URMA one is able to neatly simplify the analytic solutions using weights, i.e.

\( w_i = \left( s_i^2 + \tau^2 \right)^{-1} \) (see equation (1.5) in Section 1.7.2), and the contribution of the \( i \)th study to the pooled estimate is \( 100\% \times w_i \sqrt{\frac{\sum w_i}{\sum w_i}} \). The weights are not so clear for Model A because the analytic solutions have evidently more complicated than those for the URMA.

Without loss of generality consider the pooled solution for outcome \( j = 1 \) given complete-case data (equation (4.3)), and note that it is effectively partitioned into two terms, one
involving \( \hat{Y}_{i1} \) and one involving \( \hat{Y}_{i2} \). Indeed, the contribution of each study \( i \) to the value of \( \hat{\beta}_1 \) is \( a_i \hat{Y}_{i1} + b_i \hat{Y}_{i2} \), where:

\[
a_i = \frac{1}{\sum_{i=1}^n \frac{\hat{\tau}_{i1}^2 + s_{i1}^2}{(\hat{\tau}_{i1}^2 + s_{i1}^2 - \hat{\tau}_{i1} + \lambda_i)^2} \left( \frac{\sum_{i=1}^n (\hat{\tau}_{i2}^2 + s_{i2}^2)(\hat{\tau}_{i1}^2 + s_{i1}^2) - (\hat{\tau}_{i2} + \lambda_i)(\hat{\tau}_{i1} + \lambda_i)^2}{(\hat{\tau}_{i1} + \lambda_i)^2} \right) - \sum_{i=1}^n \frac{(\hat{\tau}_{i1} + \lambda_i)^2}{(\hat{\tau}_{i1} + \lambda_i)^2}}
\]

\[
b_i = \frac{1}{\sum_{i=1}^n \frac{\hat{\tau}_{i2}^2 + s_{i2}^2}{(\hat{\tau}_{i2}^2 + s_{i2}^2 - \hat{\tau}_{i2} + \lambda_i)^2} \left( \frac{\sum_{i=1}^n (\hat{\tau}_{i2}^2 + s_{i2}^2)(\hat{\tau}_{i2}^2 + s_{i2}^2) - (\hat{\tau}_{i2} + \lambda_i)^2}{(\hat{\tau}_{i2} + \lambda_i)^2} \right) - \sum_{i=1}^n \frac{(\hat{\tau}_{i2} + \lambda_i)^2}{(\hat{\tau}_{i2} + \lambda_i)^2}}
\]

So for example, for 3 studies, \( \hat{\beta}_1 = a_1 \hat{Y}_{11} + b_1 \hat{Y}_{12} + a_2 \hat{Y}_{21} + b_2 \hat{Y}_{22} + a_3 \hat{Y}_{31} + b_3 \hat{Y}_{32} \).

In Appendix B4 I show that \( \sum_{i=1}^n (a_i + b_i) = 1 \). Hence, each study’s weighting to the value of \( \hat{\beta}_1 \) can be thought of as \( a_i + b_i \), or \( 100 \times (a_i + b_i) \)%.

In the URMA situation, where \( \hat{\tau}_{i2} = 0 \) and \( \lambda_i = 0 \), one obtains the well-known URMA weighting for each study of:

\[
a_i = \frac{1}{\sum_{i=1}^n \frac{\hat{\tau}_{i1}^2 + s_{i1}^2}{(\hat{\tau}_{i1}^2 + s_{i1}^2)}} = \frac{\sum_{i=1}^n w_i}{\sum_{i=1}^n \sum_{j=1}^n w_j}
\]

\[b_j = 0\]

Of course, the same principles as discussed above for \( \hat{\beta}_1 \) apply for the study weighting in \( \hat{\beta}_2 \). An example and comparison of percentage weight for each study from Model A and a URMA is considered in Section 4.7.1.

### 4.5 Interpretation of the analytic solutions for the pooled estimates

Importantly, the analytic solution for \( \hat{\beta}_1 \) in equation (4.3) shows that the pooled estimate from Model A for \( j = 1 \) (i.e. \( \hat{\beta}_1 \)) now not only takes into account the \( \hat{Y}_{i1} \)s, \( s_{i1}^2 \)s and \( \hat{\tau}_{i1}^2 \), but also incorporates the \( j = 2 \) data (i.e. the \( \hat{Y}_{i2} \)s, \( s_{i2}^2 \)s, and \( \hat{\tau}_{i2}^2 \)), as well as the covariance (i.e.
between studies and the covariance within each study (i.e. the $\lambda_i$s). Similarly the pooled estimate from Model $A$ for $j = 2$ (i.e. $\tilde{\beta}_2$) incorporates the $j = 1$ parameters (equation (4.3)). This is not true for the URMA solutions, where $\tilde{\beta}_1$ and $\tilde{\beta}_2$ only take into account the $j = 1$ and $j = 2$ parameters respectively (e.g. $\tilde{\beta}_1$ from a URMA only takes into account the $\tilde{Y}_{i1}$s, the $s_{i1}^2$s and $\tilde{\tau}_{11}^2$).

Although the analytic solutions for complete-case data are quite complex, they highlight some other important points that may otherwise go unnoticed. Concentrate without loss of generality on the $j = 1$ pooled solution once more (i.e. $\tilde{\beta}_1$ in equation (4.3)) and consider its $b_i$ component as specified in equation (4.9). This $b_i$ term indicates how the $j = 1$ pooled estimate takes into account the $j = 2$ summary statistics (i.e. the $\tilde{Y}_{i2}$s). In particular, by considering the bracket in the numerator of $b_i$, i.e.

$$
\left[ \sum_{k=1}^{n} \tilde{r}_{12} (s_{i1}^2 - s_{k1}^2) + \lambda_k (\tilde{r}_{11}^2 + s_{i1}^2) - \lambda_i (\tilde{r}_{12}^2 + s_{k1}^2) \right] = \text{bracket (a), say,}
$$

and by noting that bracket (a) can also be rewritten as

$$
\left[ \sum_{k=1}^{n} \left( r_{1}^2 + s_{i1}^2 \right) r_{12} + \lambda_i \right] - \left( r_{1}^2 + s_{k1}^2 \right) r_{12} + \lambda_i = \text{bracket (b), say,}
$$

then the following two key points arise:

1: The $j = 1$ pooled estimate only uses the $\tilde{Y}_{i2}$ values if

(i) $\tilde{r}_{12} + \lambda_i \neq 0$ for all $i$ (see top line in bracket (b)); and

(ii) $s_{i1}^2 - s_{k1}^2 \neq 0$ and $\lambda_i \neq \lambda_k$ for all $i$ and $k$ (see top line in bracket (a)).

2: The larger the between study covariance ($\tilde{r}_{12}$), the larger ($s_{i1}^2 - s_{k1}^2$) and the larger ($\lambda_i - \lambda_k$) the more influence $\tilde{Y}_{i2}$ can have on the $j = 1$ pooled result (see top line in bracket (a)).
Point (1) means that if \( \tilde{\beta}_1 \) is to take into account the \( \tilde{Y}_{i2} \)s, there must be both: (i) some overall correlation across studies between \( \tilde{Y}_{ii} \) and \( \tilde{Y}_{i2} \), and (ii) differences between the \( j = l \) within-study covariance matrices across studies. Point (1) part (ii) is not necessarily intuitive. When \( s_{i1}^2 - s_{i1}^2 = 0 \) and \( \lambda_i = \lambda_k \) for all \( i \) it would still seem sensible if Model A could 'borrow strength' from the related \( j = 2 \) data, perhaps taking into account the relative size of the \( s_{i2}^2 \)s for each \( i \); however, this is not the case. This emphasises why it is important to appreciate the mathematics underlying the estimation of the parameters in Model A.

One can also think of Points (1) and (2) in terms of the URMA weights, \( w_i = \frac{1}{s_i^2 + \hat{r}_i^2} \).

If these weights are the same for all studies and the within-study covariance is the same for all studies, then Model A will give exactly the same parameter estimates as a URMA, i.e. Model A reverts entirely to two independent URMA models (Point (1)). However, if some of the studies have different weighting in a URMA for \( j = l \) (so that \( s_{i1}^2 - s_{i1}^2 \neq 0 \) for all studies) then, where \( r_{i2} + \lambda_i \neq 0 \) for every \( i \), Model A will utilise the \( j = 2 \) data for the pooled estimate of \( j = l \); similarly the pooled estimate for \( j = 2 \) will only utilise the \( j = l \) data if the URMA weights for \( j = 2 \) are not all the same and \( r_{i2} + \lambda_i \neq 0 \) for all \( i \).

Furthermore, the larger the differences in the URMA weighting across studies (i.e. \( s_{i1}^2 - s_{i1}^2 \)) the more Model A will seek to incorporate the other related outcome data (Point (2)). It is worth noting here that in practice it is very common for at least some of the studies to have different weighting in a URMA, and thus it is highly unlikely that \( s_{i1}^2 - s_{i1}^2 = 0 \) for all \( i \) and \( k \).

Given these Points (1) and (2) it is evident that there is the potential for the weighting given to each study to be different in Model A than in a URMA and that the pooled...
estimates and their precision may also be different between models. This becomes increasingly possible the larger the differences in the URMA weights for each outcome and also the larger the within- and between-study correlation between the two summary statistics. In this situation, the model increasingly allows the 'borrowing of strength' between the two related outcomes, and this can be utilised by the pooled estimates.

4.6 Empirical Bayes or shrunken study estimates from Model A

Alongside the pooled estimates it also possible to obtain empirical Bayes or 'shrunken' estimates of the underlying true summary statistics in each study ($\theta_j = \beta_j + u_j$). This may be informative if the underlying value in one particular study is of interest, and the empirical Bayes estimates from a URMA were introduced in Section 1.7.2 through equation (1.8). Taking the same approach as then, one can use $\tilde{u} = \Omega^\prime Z^\prime V^{-1}(Y - X\tilde{\beta})$ to estimate the random-effects $u_j$, where $Y - X\tilde{\beta} = (r_{11}, r_{12}, r_{21}, r_{22}, \ldots, r_{n1}, r_{n2})$ are the residuals, i.e. the difference from the summary statistic from each study to the overall pooled estimate. Now for Model A:

$$\Omega^\prime Z^\prime V^{-1} = \Omega^\prime V^{-1}$$

$$= \text{diag} \begin{bmatrix} r_{11}^2 & r_{12}^2 & \cdots & r_{n1}^2 \\ r_{12}^2 & r_{22}^2 & \cdots & r_{n2}^2 \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1}^2 & r_{n2}^2 & \cdots & r_{nn}^2 \end{bmatrix} \text{ diag} \begin{bmatrix} \tau_1^2 + s_{11}^2 \left( \tau_1^2 + s_{11}^2 \right) - (r_{11} + \lambda_1)^2 \\ \vdots \\ \tau_1^2 + s_{11}^2 \left( \tau_1^2 + s_{11}^2 \right) - (r_{n1} + \lambda_1)^2 \\ \tau_2^2 + s_{22}^2 \left( \tau_2^2 + s_{22}^2 \right) - (r_{12} + \lambda_2)^2 \\ \vdots \\ \tau_{n1}^2 + s_{n1}^2 \left( \tau_{n1}^2 + s_{n1}^2 \right) - (r_{n1} + \lambda_{n1})^2 \end{bmatrix}$$

Therefore:

$$\tilde{u} = \Omega^\prime Z^\prime V^{-1}(Y - X\tilde{\beta})$$

$$= \begin{bmatrix} \bar{u}_{11} \\ \bar{u}_{12} \\ \vdots \\ \bar{u}_{n1} \end{bmatrix} \begin{bmatrix} \frac{\tau_1^2 + s_{11}^2 \left( \tau_1^2 + s_{11}^2 \right) - (r_{11} + \lambda_1)^2}{(r_{11} + s_{11}^2)(\tau_1^2 + s_{11}^2) - (r_{11} + \lambda_1)^2} + \frac{r_{12}}{r_{12}} \left( \frac{\tau_1^2 + s_{11}^2 \left( \tau_1^2 + s_{11}^2 \right) - (r_{12} + \lambda_1)^2}{(r_{11} + s_{11}^2)(\tau_1^2 + s_{11}^2) - (r_{12} + \lambda_1)^2} \right) + \frac{r_{11}}{r_{11}} \left( \frac{\tau_1^2 + s_{11}^2 \left( \tau_1^2 + s_{11}^2 \right) - (r_{11} + \lambda_1)^2}{(r_{11} + s_{11}^2)(\tau_1^2 + s_{11}^2) - (r_{11} + \lambda_1)^2} \right) \right) \\ \frac{\tau_2^2 + s_{22}^2 \left( \tau_2^2 + s_{22}^2 \right) - (r_{12} + \lambda_2)^2}{(r_{11} + s_{12}^2)(\tau_2^2 + s_{22}^2) - (r_{12} + \lambda_2)^2} + \frac{r_{12}}{r_{12}} \left( \frac{\tau_2^2 + s_{22}^2 \left( \tau_2^2 + s_{22}^2 \right) - (r_{12} + \lambda_2)^2}{(r_{11} + s_{12}^2)(\tau_2^2 + s_{22}^2) - (r_{12} + \lambda_2)^2} \right) + \frac{r_{11}}{r_{11}} \left( \frac{\tau_2^2 + s_{22}^2 \left( \tau_2^2 + s_{22}^2 \right) - (r_{12} + \lambda_2)^2}{(r_{11} + s_{12}^2)(\tau_2^2 + s_{22}^2) - (r_{12} + \lambda_2)^2} \right) \right) \\ \vdots \\ \frac{\tau_{n1}^2 + s_{n1}^2 \left( \tau_{n1}^2 + s_{n1}^2 \right) - (r_{n1} + \lambda_{n1})^2}{(r_{n1} + s_{n1}^2)(\tau_{n1}^2 + s_{n1}^2) - (r_{n1} + \lambda_{n1})^2} + \frac{r_{12}}{r_{12}} \left( \frac{\tau_{n1}^2 + s_{n1}^2 \left( \tau_{n1}^2 + s_{n1}^2 \right) - (r_{n1} + \lambda_{n1})^2}{(r_{n1} + s_{n1}^2)(\tau_{n1}^2 + s_{n1}^2) - (r_{n1} + \lambda_{n1})^2} \right) + \frac{r_{11}}{r_{11}} \left( \frac{\tau_{n1}^2 + s_{n1}^2 \left( \tau_{n1}^2 + s_{n1}^2 \right) - (r_{n1} + \lambda_{n1})^2}{(r_{n1} + s_{n1}^2)(\tau_{n1}^2 + s_{n1}^2) - (r_{n1} + \lambda_{n1})^2} \right) \right) \end{bmatrix}$$

(4.10)
The empirical Bayes estimates for Model A can then be obtained by \( \tilde{\theta}_{ij} = \tilde{\beta}_j + \tilde{\nu}_{ij} \). Of course, if \( \bar{r}_{12} = \lambda_i = 0 \) the URMA solution is obtained, i.e. \( \tilde{u}_{i1} = r_{11} \left( \frac{\bar{u}_i^2}{(\bar{u}_i^2 + \sigma_i^2)} \right) \).

To consider these solutions in more detail, take the empirical Bayes estimate for study 1.

The URMA solution for \( \tilde{u}_{i1} \) is always less than or equal to \( r_{11} \), and therefore, when compared to the original study estimate (i.e. \( \tilde{Y}_{i1} \)), \( \tilde{\theta}_{11} \) is 'shrunk' toward to the overall pooled value \( \bar{\beta}_1 \) (see Section 1.7.2). However, the empirical Bayes solution for Model A is much more difficult to interpret (equation (4.10)), as the introduction of the within- and between-study covariances (\( \lambda_i \) and \( \bar{r}_{12} \)) allows \( r_{11} \) and the \( j = 2 \) residual \( r_{12} \) to influence the solution for \( \tilde{u}_{i1} \). Indeed, this allows the possibility of \( \tilde{\theta}_{11} \) being further away from \( \bar{\beta}_1 \) than \( \tilde{Y}_{i1} \). However, the use of the term 'shrinkage' is still correct for the empirical Bayes estimates from Model A as long as one considers both outcomes jointly for each study, because the shrinkage toward the pooled estimates relates to the overall shrinkage across both outcomes in that study (i.e. the shrinkage is a two-dimensional issue). For example, if \( \tilde{\theta}_{11} \) is further away from \( \bar{\beta}_1 \) than \( \tilde{Y}_{i1} \), then to a greater degree \( \tilde{\theta}_{12} \) will be closer to \( \bar{\beta}_2 \) than \( \tilde{Y}_{12} \), and overall there will be shrinkage across the two dimensions. This is shown clearly in Section 4.7.4 during the application of Model A to the Berkey data, which now follows.

### 4.7 Application of Model A to the Berkey dataset

As an applied example of Model A, and to emphasise the analytic solutions identified so far in Chapter 4, I will now consider the dataset published by Berkey et al., containing two outcomes from each of 5 periodontal clinical trials (Table 4.1, 'original values') [151]. As introduced in Section 3.4, these outcomes are improvement (surgical minus non-surgical group) in probing depth and in attachment level, and summary statistics for both outcomes are available from all the 5 studies, together with all the 5 \( \lambda_i \) s. As well as these original...
observed values, I have also created a modified version of this dataset, where my only change was to set all the $s^2_{ij}$ values to 0.01 and all the $\lambda_i$s to zero (Table 4.1, ‘modified values’). This modified dataset will be used to reiterate Point (1) in Section 4.5.

Table 4.1: The originally observed and slightly modified data from Berkey et al., where PD = probing depth, AL = attachment level, URMA = a univariate rather than bivariate (Model A) random-effects meta-analysis [151].

<table>
<thead>
<tr>
<th>Study</th>
<th>% weighting of each study data in the estimation of $\hat{\beta}_{ij}$</th>
<th>Data</th>
<th>Original Variance Values</th>
<th>Modified Variance Values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Original data</td>
<td>Modified Data</td>
<td>Outcome</td>
<td>$\hat{y}_{ij}$</td>
</tr>
<tr>
<td>1</td>
<td>18.1%, 18.1%</td>
<td>20%, 20%</td>
<td>PD</td>
<td>0.470</td>
</tr>
<tr>
<td>1</td>
<td>AL</td>
<td>-0.320</td>
<td>0.0077</td>
<td>0.0100</td>
</tr>
<tr>
<td>2</td>
<td>19.9%, 20.2%</td>
<td>20%, 20%</td>
<td>PD</td>
<td>0.200</td>
</tr>
<tr>
<td>2</td>
<td>AL</td>
<td>-0.600</td>
<td>0.0008</td>
<td>0.0008</td>
</tr>
<tr>
<td>3</td>
<td>25.1%, 24.9%</td>
<td>20%, 20%</td>
<td>PD</td>
<td>0.400</td>
</tr>
<tr>
<td>3</td>
<td>AL</td>
<td>-0.120</td>
<td>0.0014</td>
<td>0.0014</td>
</tr>
<tr>
<td>4</td>
<td>23.8%, 23.6%</td>
<td>20%, 20%</td>
<td>PD</td>
<td>0.260</td>
</tr>
<tr>
<td>4</td>
<td>AL</td>
<td>-0.310</td>
<td>0.0015</td>
<td>0.0015</td>
</tr>
<tr>
<td>5</td>
<td>13.1%, 13.2%</td>
<td>20%, 20%</td>
<td>PD</td>
<td>0.560</td>
</tr>
<tr>
<td>5</td>
<td>AL</td>
<td>-0.390</td>
<td>0.0304</td>
<td>0.0304</td>
</tr>
</tbody>
</table>

N.B. $\lambda_i$ is the within-study covariance, $\rho_{wi}$ is the within-study correlation.

4.7.1 Results for the original Berkey dataset

For the original Berkey dataset, there does appear to be a reasonably strong between-study correlation between the summary statistics for each outcome ($\hat{\rho}_B = 0.61$). Furthermore, the pooled estimates for both outcomes in Model A are slightly closer to zero than in the URMA, indicating that the URMA solutions are possibly slightly overestimating the pooled values (Table 4.2). Also, the standard error of the pooled estimates has decreased in Model A, relating to the fact that $\hat{\tau}^2_i$ and $\hat{\tau}^2_j$ have also decreased. Although the overall conclusions from the URMA model remain unchanged, these results indicate some of the benefits of the bivariate random-effects meta-analysis (BRMA) approach. The differences in the URMA weighting across studies and the high correlation has allowed the two outcomes to ‘borrow strength’ from each other in Model A. This has caused the pooled

R.D. Riley, Ph.D. Thesis

Chapter 4

136
estimates to change slightly, and importantly it is has increased their associated precision.

This illustrates the discussion in Section 4.5 about how and why the Model A pooled estimates and their precision can be different to those from a URMA.

**Table 4.2:** RIGLS univariate (URMA) and bivariate (Model A) random-effects meta-analysis results for the original and modified Berkey data where PD = probing depth, AL = attachment level and s.e. = standard error [151].

<table>
<thead>
<tr>
<th></th>
<th>PD</th>
<th>AL</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model</td>
<td>$\hat{\beta}_1$ (s.e.)</td>
<td>$\tilde{s}^2_{11}$ (s.e.)</td>
<td>$\tilde{\rho}^2_{1}$ (s.e.)</td>
<td>$\tilde{r}_{12}$</td>
<td>$\tilde{\rho}_h$</td>
<td>$\text{corr}(\tilde{\beta}_1, \tilde{\beta}_2)$ (s.e.)</td>
</tr>
<tr>
<td><strong>Original Data</strong></td>
<td>URMA</td>
<td>0.361 (0.0592)</td>
<td>0.0119 (0.0885)</td>
<td>-0.346 (0.107)</td>
<td>0.0331 -</td>
<td>- - -</td>
<td>0.706 (0.107)</td>
</tr>
<tr>
<td></td>
<td>Model A</td>
<td>0.353 (0.0590)</td>
<td>0.0117 (0.0879)</td>
<td>-0.339 (0.107)</td>
<td>0.0327 0.0119 0.609</td>
<td>0.547</td>
<td>0.692 (0.074)</td>
</tr>
<tr>
<td><strong>Modified Data</strong></td>
<td>URMA</td>
<td>0.378 (0.0660)</td>
<td>0.0119 (0.0885)</td>
<td>-0.346 (0.111)</td>
<td>0.0331 -</td>
<td>- - -</td>
<td>0.724 (0.111)</td>
</tr>
<tr>
<td></td>
<td>Model A</td>
<td>0.378 (0.0660)</td>
<td>0.0119 (0.0877)</td>
<td>-0.325 (0.0769)</td>
<td>0.0329 0.0154 0.778</td>
<td>0.531</td>
<td>0.703 (0.0769)</td>
</tr>
</tbody>
</table>

The 'borrowing of strength' can be also seen by some slight changes to each study’s weighting in Model A compared to those in the URMA model (Table 4.1). Although the changes in weighting are small, $\tilde{Y}_{21}$ from study 2 now has a slightly increased weighting in the value of $\tilde{\beta}_1$ because $s^2_{21}$ is only moderately sized compared to the other four other $s^2_{ii}$ values ($i = 1,3,4$ or 5) but the related attachment level estimate ($\tilde{Y}_{22}$) in study 2 has a very small value of $s^2_{22}$ compared to most of the other $s^2_{ij}$ values, and so $\tilde{\beta}_1$ is able to utilise the strong related information coming from study 2. This illustrates the discussion in Section 4.4 about differences between URMA and Model A study weights.

Some authors are strongly against URMA approaches because those studies with a large $s^2_{ij}$ can have more weighting in a random-effects model than in a fixed-effects model, and they consider this to be philosophically wrong [48]. Given this, the same authors are likely
to be even more strongly against the BRMA approach because those studies with the larger $s_{ij}^2$ values can have even more weighting in Model A than in a URMA, and this is shown in the Berkey results for study 2. Personally I believe that the random-effects approach is a much more realistic model than a fixed-effects analysis because in most situations between-study heterogeneity will exist [50; 151]. Given this I also consider Model A to be a natural and sensible extension to use for meta-analysis. Of course, if there is heterogeneity one should also attempt to explain it wherever possible [40], and the extension of Model A to bivariate meta-regression will be considered in Section 6.7.4. Perhaps the key message from this Berkey example is that meta-analysts need to be aware of the underlying estimation mechanism in a meta-analysis model so that they can decide if they are comfortable with using the model in practice. In particular for Model A, the results for the original Berkey dataset show how strong within- and between-study correlation can allow a study to have more weighting in Model A than in a URMA, and meta-analysts need to be comfortable with this if they consider Model A to be suitable for their own datasets. The changes in weighting were only small for the original Berkey dataset, but in other situations the changes may be much more substantial (e.g. see the use of Model A when there is missing data in Chapter 6).

4.7.2 Results for the modified Berkey dataset

For the modified Berkey dataset, the probing depth results from Model A are now identical to those from a URMA, even though there is still a strong between-study correlation between probing depth and attachment level ($\hat{\rho}_B = 0.78$, see Table 4.2). This is because the value of $s_{ii}^2$ is the same for all 5 probing depth estimates and $\lambda_i = 0$ for all $i$, which means that there is no difference in the URMA weights across studies for probing depth and therefore no ‘borrowing of strength’ for this outcome is possible in Model A (this illustrates Point (1) of Section 4.5). However, the URMA weights for attachment level are
different, as the $s^2_{ij}$s are not the same for each study, and therefore estimation for
attachment level takes into account the probing depth data, which causes a slight decrease
in $\hat{\tau}^2_1$ and an increase in the precision of $\hat{\beta}_2$.

4.7.3 Results for $(\hat{\beta}_1 - \hat{\beta}_2)$

Both the original and modified Berkey data results show that it is possible to calculate
$(\hat{\beta}_1 - \hat{\beta}_2)$ from Model A if desired (Table 4.2). The pooled difference was not of interest
in the original Berkey publications, yet it is conceivable that $(\hat{\beta}_1 - \hat{\beta}_2)$ may be of interest if
one wanted some overall score across outcomes or wanted to assess the hypothesis that
$\hat{\beta}_1 = \hat{\beta}_2$. However, my main reasons for including $(\hat{\beta}_1 - \hat{\beta}_2)$ here are to demonstrate the
ability of Model A to calculate this value and to allow a comparison to the equivalent
URMA value.

The results show that the estimate $(\hat{\beta}_1 - \hat{\beta}_2)$ from Model A has a much smaller standard
error than that from the URMA, even for the modified dataset, because the $corr(\hat{\beta}_1, \hat{\beta}_2)$ is
being taken into account (Table 4.2). For example, for the original dataset, $(\hat{\beta}_1 - \hat{\beta}_2)$ from
the URMA is 0.706 with a standard error of 0.107. However, Model A gives an estimate of
0.694 with a much smaller standard error of 0.074 due to the incorporation of the high
value of $corr(\hat{\beta}_1, \hat{\beta}_2)$, which equals 0.547. This issue is discussed further in Section 4.8.6
and a simulation study to investigate the use of Model A for estimating $(\beta_1 - \beta_2)$ is
performed in Chapter 6.

4.7.4 Results for the empirical Bayes estimates

I have also obtained the empirical Bayes estimates for the original Berkey dataset when
applying Model A, and also those from a separate URMA for each outcome (Table 4.3).

The URMA empirical Bayes estimates are all shrunk toward to the pooled estimates.
However, this is not always true for Model A; for example the probing depth empirical Bayes estimate for study 3 is further away from $\hat{\beta}_1$ than $\bar{y}_3$. Nevertheless, the empirical Bayes estimates from Model A are still subject to shrinkage when one considers both outcomes collectively. For example, although $\hat{\theta}_3$ moves further away from $\hat{\beta}_1$ than $\bar{y}_3$, $\hat{\theta}_2$ moves to a greater degree toward $\hat{\beta}_2$ than $\bar{y}_3$, and therefore collectively for study 3 the empirical Bayes estimates are shrunk across the two dimensions. This illustrates the discussion in Section 4.6.

### Table 4.3: Original summary statistics and the RIGLS empirical Bayes study estimates from the univariate (URMA) and bivariate (Model A) random-effects meta-analysis of the (original) Berkey Dataset in Table 4.1, where PD = probing depth and AL = attachment level.

<table>
<thead>
<tr>
<th>Study</th>
<th>Outcome</th>
<th>Original Study Estimate</th>
<th>URMA Empirical Bayes Study Estimate</th>
<th>Model A Empirical Bayes Study Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PD</td>
<td>0.470</td>
<td>0.428</td>
<td>0.418</td>
</tr>
<tr>
<td>1</td>
<td>AL</td>
<td>-0.320</td>
<td>-0.325</td>
<td>-0.325</td>
</tr>
<tr>
<td>2</td>
<td>PD</td>
<td>0.200</td>
<td>0.253</td>
<td>0.230</td>
</tr>
<tr>
<td>2</td>
<td>AL</td>
<td>-0.600</td>
<td>-0.594</td>
<td>-0.591</td>
</tr>
<tr>
<td>3</td>
<td>PD</td>
<td>0.400</td>
<td>0.394</td>
<td>0.402</td>
</tr>
<tr>
<td>3</td>
<td>AL</td>
<td>-0.120</td>
<td>-0.129</td>
<td>-0.128</td>
</tr>
<tr>
<td>4</td>
<td>PD</td>
<td>0.260</td>
<td>0.280</td>
<td>0.286</td>
</tr>
<tr>
<td>4</td>
<td>AL</td>
<td>-0.310</td>
<td>-0.312</td>
<td>-0.308</td>
</tr>
<tr>
<td>5</td>
<td>PD</td>
<td>0.560</td>
<td>0.450</td>
<td>0.431</td>
</tr>
<tr>
<td>5</td>
<td>AL</td>
<td>-0.390</td>
<td>-0.369</td>
<td>-0.344</td>
</tr>
</tbody>
</table>

URMA Results: PD pooled estimate = 0.361, AL pooled estimate = -0.346
Model A Results: PD pooled estimate = 0.353, AL pooled estimate = -0.339

### 4.8 Further comparison of the parameter estimates from Model A to those from a univariate random-effects meta-analysis

I will now continue with a comparison of the complete-case analytic solutions from Model A with those equivalent solutions from a URMA. Firstly note that, in both a URMA model and Model A, RIGLS estimation theory suggests that the estimates of the pooled values and the between-study variance parameters will be [51]:
1) **Unbiased.** The expected value of each parameter estimate is the true parameter value.

2) **Consistent.** As \( n \) tends to infinity, the estimates tend to the true value. The mean-square error (MSE) also therefore tends to zero as \( n \) tends to infinity.

Furthermore, RIGLS estimates are equivalent to REML estimates in both a URMA and Model A because a normal error structure is assumed [174]. Therefore, they are also the minimum variance unbiased estimators of all the possible set of estimators for that model.

However, as a URMA model and Model A are different models (as Model A contains correlation parameters) it remains important to assess which one produces pooled RIGLS estimates with the greatest precision. RIGLS estimation theory states that RIGLS will produce unbiased parameter estimates in both models, but which one produces the smallest uncertainty in the pooled estimates required from the meta-analysis? This is particularly important as smaller standard errors may enable more reliable evidence-based conclusions to be formed, something that I ultimately want to facilitate for prognostic markers. The following ‘Remark 1’ considers this issue directly, and it is followed by five other remarks that arise from comparing the Model A and URMA analytic solutions.

### 4.8.1 Remark 1: The precision of \( \hat{\beta}_j \) is smallest in Model A on average

I will now compare the precision of \( \hat{\beta}_j \) from Model A with that from a URMA when there is complete-case data, and ideally I would like to show that on average:

\[
\text{var}(\hat{\beta}_j) \text{ from Model A} \leq \text{var}(\hat{\beta}_j) \text{ from the URMA model} \quad (4.11)
\]

Consider without loss of generality \( \hat{\beta}_i \). Now, the theory of RIGLS estimation suggests \( \hat{\tau}_1^2 \) and \( \hat{\tau}_2^2 \) will be the same from the URMA model and Model A on average, as both models should give unbiased estimates of \( \tau_1^2 \) and \( \tau_2^2 \) using RIGLS in principle [51]. Therefore, assuming the same value of \( \hat{\tau}_1^2 \) (\( = \tau_1^2 \) say) on both sides of equation (4.11), and also the same value of \( \hat{\tau}_2^2 \) (\( = \tau_2^2 \) say) on both sides of the equation, the proposal is that:
\[ \text{var}(\hat{\beta}_i) \text{ from Model A} \leq \text{var}(\hat{\beta}_i) \text{ from the URMA model, if and only if} \]
\[ \sum_{i=1}^{n} \frac{\left( r_{i1}^2 + s_{i1}^2 \right)}{\left( r_{i1}^2 + s_{i1}^2 \right) - (\tau_{i2} + \lambda_i)^2} \left[ \sum_{i=1}^{n} \frac{\left( r_{i1}^2 + s_{i1}^2 \right)}{\left( r_{i1}^2 + s_{i1}^2 \right) - (\tau_{i2} + \lambda_i)^2} \right] \leq \left( \sum_{i=1}^{n} \frac{1}{\left( r_{i1}^2 + s_{i1}^2 \right) - (\tau_{i2} + \lambda_i)^2} \right)^{-1} \] (4.12)

Is the inequality in equation (4.12) true? Rearranging and ensuring a common denominator gives:
\[ \left[ \sum_{i=1}^{n} \frac{\left( r_{i1}^2 + s_{i1}^2 \right)}{\left( r_{i1}^2 + s_{i1}^2 \right) - (\tau_{i2} + \lambda_i)^2} \right] \left[ \sum_{i=1}^{n} \frac{1}{\left( r_{i1}^2 + s_{i1}^2 \right) - (\tau_{i2} + \lambda_i)^2} \right] \leq \left( \sum_{i=1}^{n} \frac{1}{\left( r_{i1}^2 + s_{i1}^2 \right) - (\tau_{i2} + \lambda_i)^2} \right)^{-1} \] (4.13)

The denominator in equation (4.13) is always positive because \( \tau_i^2 \sigma_j^2 \geq \tau_i^2 \) (as the modulus of the between correlation coefficient is \( \leq 1 \)) and the other terms are all squared; thus for the inequality in equation (4.13) to hold the numerator must be \( \leq 0 \). Hence, consider just the numerator of equation (4.13); rearranging this, collecting together the summations and using a common denominator produces the following:
\[ \sum_{i=1}^{n} \sum_{k=1}^{n} \left( \frac{\left( r_{i1}^2 + s_{i1}^2 \right) - (\tau_{i2} + \lambda_i)^2 \left( \tau_{k1} + s_{k1}^2 \right) - (\tau_{k2} + \lambda_k)^2 \left( \tau_{i2} + \lambda_i \right)^2}{\left( r_{i1}^2 + s_{i1}^2 \right) - (\tau_{i2} + \lambda_i)^2 \left( r_{k1}^2 + s_{k1}^2 \right) - (\tau_{k2} + \lambda_k)^2 \left( \tau_{i2} + \lambda_i \right)^2} \right) \leq 0 \] (4.14)

The terms in equation (4.14) where \( i = k \) equate to zero. Now consider all the other terms where \( i \neq k \). These terms essentially come in pairs. For example, one term for \( i = 1 \) and \( k = 2 \), and also a second term where \( i = 2 \) and \( k = 1 \). Consider summing each of these paired terms. For example, summing the term for \( i = 1 \) and \( k = 2 \) with the term for \( i = 2 \) and \( k = 1 \) using a common denominator gives:
\[ \frac{- \left[ \left( \tau_{i2} \left( s_{i1}^2 - s_{k1}^2 \right) + \tau_{i1} \left( \lambda_{i1} - \lambda_{k1} \right) \right) + \left( s_{i1}^2 \lambda_{i2} + s_{k1}^2 \lambda_{k2} \right) \right]}{\left( r_{i1}^2 + s_{i1}^2 \right) \left( r_{k2}^2 + s_{k2}^2 \right) - (\tau_{i2} + \lambda_i)^2 \left( \tau_{k2} + \lambda_k \right)^2} \leq 0 \] (4.15)
The whole expression in equation (4.15) is always less than or equal to zero because the numerator is always negative or zero, whilst the denominator is always positive as all the four bracketed components are positive. Hence, the sum of all the pairs in the left of equation (4.14) will be less than or equal to zero, and thus equation (4.14) itself is true. Therefore, working backwards, the inequality in equation (4.11) must also be true, and thus for complete-case data this shows that \( \text{var}(\hat{\beta}_j) \text{ from Model A} \leq \text{var}(\hat{\beta}_j) \text{ from a URMA} \) on average, conditional on the assumption that \( \tau_1^2 \) and \( \tau_2^2 \) are unbiased in both models.

Of course, as this inequality indicates what will happen on average, by chance in a single dataset \( \tau_1^2 \) and \( \tau_2^2 \) may both be larger in Model A than in the equivalent URMA, thus potentially causing \( \text{var}(\hat{\beta}_j) \) from Model A to be larger in that particular dataset; however, in the long-run (i.e. on average) the inequality states that one will observe a smaller standard error of the pooled estimates from Model A than for the URMA. I emphasise again though that this finding is based on the assumption that \( \tau_1^2 \) and \( \tau_2^2 \) are unbiased for both URMA and Model A as RIGLS theory suggests (see Section 4.1).

4.8.2 Remark 2: The mean-square error of \( \hat{\beta}_j \) is smallest in Model A on average

The mean-square error (MSE) of \( \hat{\beta}_j = E \left[ (\hat{\beta}_j - \beta_j)^2 \right] \), i.e. the average squared-distance between the true and estimated pooled value, and this gives a combined measurement of the bias and precision of \( \hat{\beta}_j \). Again assuming that \( \hat{\beta}_j \) from both a URMA model and Model A are unbiased as RIGLS suggests, and given Remark 1 holds for complete-case data, there must also, by definition, be a smaller MSE for \( \hat{\beta}_j \) on average in Model A compared to a URMA when there is complete-case data. However, this result is again conditional on the assumption of unbiased parameter estimates in both models, in particular for \( \tau_1^2 \) and \( \tau_2^2 \). If Model A does increase precision and decrease MSE on
average it would be a highly desirable model for meta-analysts and those seeking
evidence-based conclusions (e.g. about prognostic markers), because, compared to a
URMA, potentially more reliable conclusions could be drawn from Model A about the
true values of the pooled values of interest. A simulation study of Model A versus URMA
is therefore needed to help establish whether Remarks 1 and 2 are indeed true (see
Chapters 5 and 6).

4.8.3 Remark 3: Remarks 1 and 2 may be true for \( \bar{\tau}_j^2 \) in Model A

Following the same principles as above, it is conceivable that the precision and MSE of \( \bar{\tau}_j^2 \)
are also improved in Model A on average. Given Model A is 'borrowing strength' between
outcomes through the incorporation of within- and between-study correlation, the same
findings for \( \bar{\tau}_j^2 \) as for \( \hat{\beta}_j \) would appear plausible. However, this again assumes that \( \bar{\tau}_i^2 \) and
\( \bar{\tau}_z^2 \) are unbiased in both models, and is much more difficult to consider analytically given
the complex nature of the analytic solutions for \( \bar{\tau}_j^2 \) in Model A (see Appendix B3). This
again emphasises the need for a simulation study to help assess the differences between
parameter estimates from Model A and a URMA.

4.8.4 Remark 4: Situations when Model A is equivalent to two independent URMAs

Consider just outcome \( j = 1 \), again without loss of generality. The numerator in equation
(4.15) of Section 4.8.1 indicates that \( \text{var}(\hat{\beta}_1) \) from Model A is equivalent to that from a
URMA for complete-case data when (i) \( \tau_{1i} + \lambda_i \neq 0 \) for all \( i \); or when (ii) \( s_{ii}^2 - s_{ki}^2 \neq 0 \) and
\( \lambda_i \neq \lambda_k \) for all \( i \) and \( k \). This concurs with Points (1) and (2) in Section 4.5, and reiterates
when Model A reverts back to two independent URMAs.
4.8.5 Remark 5: The potential benefits of Model A given missing data

Again consider just $j = 1$, without loss of generality. One reason why Model A would be identical to a URMA for outcome $j = 1$ is if $s^2_{ii}$ and $\lambda_i$ were the same for all studies (see Remark 4 above). However, when there is missing data (i.e. $\hat{Y}_{ni}$ is missing for some studies) this can never happen. Studies with missing data can be considered to have a $s^2_{ii} = \infty$ (see Section 6.1), and so the differences between the $s^2_{ii}$s for known and missing data is virtually infinity themselves. Furthermore, the numerator in equation (4.15) indicates that the $\text{var}(\hat{\beta}_i)$ from Model A will be increasingly smaller than the $\text{var}(\hat{\beta}_i)$ from a URMA when there are large differences between the $s^2_{ii}$s. 

This indicates that there is perhaps the greatest potential for 'borrowing of strength' in Model A when there is missing data. This is important, as missing summary statistics is a common problem, and was a particularly large problem for OS and DFS HRs from the neuroblastoma review (see Chapter 2). Missing data is considered further in Chapter 6.

4.8.6 Remark 6: Estimation of $\text{var}(\hat{\beta}_1, \hat{\beta}_2)$ using Model A compared to a URMA

As discussed throughout the literature review in Section 3.6.3, often of interest is the difference between the pooled estimates $(\hat{\beta}_1 - \hat{\beta}_2)$. For example, if one were interested in the pooled log-odds for a treatment group $(\hat{\beta}_1)$ and the pooled log-odds for a control group $(\hat{\beta}_2)$, one may also be interested in the pooled log-odds ratio $(\hat{\beta}_1 - \hat{\beta}_2)$ [148].

Now, the $\text{var}(\hat{\beta}_1, \hat{\beta}_2) = \text{var}(\hat{\beta}_1) + \text{var}(\hat{\beta}_2) - 2\text{cov}(\hat{\beta}_1, \hat{\beta}_2)$. However, if one only performed two independent URMAs, the $\text{cov}(\hat{\beta}_1, \hat{\beta}_2)$ would be unknown and one may have to, potentially wrongly, assume the $\text{cov}(\hat{\beta}_1, \hat{\beta}_2)$ was zero. However, the $\text{cov}(\hat{\beta}_1, \hat{\beta}_2)$ is available from Model A (see equation (4.7) in Section 4.3) and therefore Model A will
produce much more appropriate estimate of \( \text{var}(\tilde{\beta}_1 - \tilde{\beta}_2) \) as it will take \( \text{cov}(\tilde{\beta}_1, \tilde{\beta}_2) \) into account.

Where \( \text{cov}(\tilde{\beta}_1, \tilde{\beta}_2) \) is positive, one will reduce the \( \text{var}(\tilde{\beta}_1 - \tilde{\beta}_2) \) from that which is available from two independent URMA analyses; this was seen in the \( (\tilde{\beta}_1 - \tilde{\beta}_2) \) results from the Berkey example in Table 4.2. Where \( \text{cov}(\tilde{\beta}_1, \tilde{\beta}_2) \) is negative, one will most likely increase the \( \text{var}(\tilde{\beta}_1 - \tilde{\beta}_2) \) from that which is available from two independent URMA analyses. This is particularly important because the URMA models will be overstating the certainty of \( (\tilde{\beta}_1 - \tilde{\beta}_2) \), which may then lead to misleading evidence-based conclusions. Of course, by obtaining a more appropriate \( \text{var}(\tilde{\beta}_1 - \tilde{\beta}_2) \), Model A will therefore also obtain a more appropriate coverage of \( (\tilde{\beta}_1 - \tilde{\beta}_2) \), i.e. 95% confidence intervals will truly be 95% confidence intervals, not > 95% (as for where the URMA model overestimates \( \text{var}(\tilde{\beta}_1 - \tilde{\beta}_2) \)) and not < 95% (as for where the URMA model underestimates \( \text{var}(\tilde{\beta}_1 - \tilde{\beta}_2) \)).

For further illustration of this see the simulation studies for \( (\tilde{\beta}_1 - \tilde{\beta}_2) \) for both complete-case and missing data in Section 6.6.

Importantly, even when Model A is equivalent to two independent URMAs in terms of the individual parameter estimates and their precision (see Remark 4), those interested in \( (\tilde{\beta}_1 - \tilde{\beta}_2) \) should still use Model A in this situation because the \( \text{var}(\tilde{\beta}_1 - \tilde{\beta}_2) \) could potentially be very different to that from a URMA, again due to the incorporation of \( \text{cov}(\tilde{\beta}_1, \tilde{\beta}_2) \).
4.9 Summary and rationale for subsequent chapters

Chapter 4 has indicated how one can estimate the parameters in Model A and has provided the analytic solutions for the pooled estimates, with critical comparison to those from a URMA. In particular, the investigation of the analytic solutions has shown how and why the pooled estimates from Model A can potentially be different to those from a URMA. Furthermore, they suggest that Model A is capable of potentially more appropriate evidence-based conclusions about $\hat{\beta}_1$, $\hat{\beta}_2$ and $(\hat{\beta}_1 - \hat{\beta}_2)$, because the variance and MSE of the parameter estimates have the potential to be smaller when Model A is used rather than two independent URMA s, especially when there is missing data. The fact that benefits may exist not just for $(\hat{\beta}_1 - \hat{\beta}_2)$ but also for $\hat{\beta}_1$ and $\hat{\beta}_2$ is particularly important as this has not previously been demonstrated in detail (see Section 3.7), whilst $\hat{\beta}_1$ and $\hat{\beta}_2$ will often be of primary interest from an evidence synthesis of prognostic marker studies (e.g. pooled DFS and OS HRs). The research in this chapter has therefore gone some of the way to achieving a number of the targets specified in Section 3.7.

Investigations of the analytic solutions can, however, only go so far. In particular, there is a real need to actually fit Model A to some simulated datasets in order to assess the assumption that $\hat{\beta}_1$, $\hat{\beta}_2$ and particularly $\hat{\tau}_1^2$ and $\hat{\tau}_2^2$ are unbiased, as RIGLS theory would suggest it should be [51]. Furthermore, what if there truly is gain in precision of $\hat{\beta}_j$ from Model A on average, as Remark 1 suggests (Section 4.8.1), but that in reality it is only very small? In this situation, is it worth recommending researchers use the more complex Model A to estimate $\beta_j$ than a more standard URMA? Also, are the benefits of Model A mainly due to the within-study correlation, the between-study correlation, or do they depend on the relationship between the within- and between-study parameters? All these questions are very difficult to answer just by assessing the complex analytic solutions, but
they need to be addressed to help inform those considering using Model A in practice. Hence, to further assess the benefits and limitations of Model A, the next chapter considers an extensive simulation study. For this, I will continue to assume that the within-study correlations are known in each study. Of course this may not be true in reality, and it is a particular problem for any application of Model A to the OS and DFS HRs from the neuroblastoma review, and the issue of unknown within-study correlations will be considered directly in Chapter 8. However, it is firstly more important at this stage to provide a full assessment of Model A for when the within-study correlations are known. For instance, if there are some serious problems or limitations of Model A even when the within-study correlations are known, it is vital to be aware of these before considering if and how Model A can be used when the within-study correlations are unknown, as is the case for evidence synthesis of prognostic marker studies.
Chapter 5

A SIMULATION STUDY TO ASSESS THE BENEFITS AND LIMITATIONS OF MODEL A FOR COMPLETE-CASE DATA

Chapter overview

In this chapter I will perform an extensive simulation study of Model A, as it is clearly necessary at this stage of the thesis to assess Model A in a wide variety of situations and to potentially highlight any practical or computational problems not exposed by the analytic solutions in Chapter 4. For simplicity and clarity, I will only consider simulations where the true within- and between-study correlations were greater than or equal to zero. This is likely to be the most common situation in practice, however I will discuss the relation of my findings to negative correlation situations in Section 6.7.1. I will also only consider complete-case data in this chapter; missing data situations are assessed separately in Chapter 6. Furthermore, again for simplicity, I will only assess Model A for estimating \( \beta_j \) in this chapter; extensions to \( (\beta_1 - \beta_2) \) will be made in Section 6.6.

5.1 Previous simulation studies of bivariate meta-analysis models

There have been only two previous simulation studies of multivariate meta-analysis as far as I am aware. Berkey at al. compare a bivariate fixed-effects meta-regression with a bivariate random-effects meta-regression using a simulation of 5000 meta-analyses (each of 5 studies, all of which report both outcomes and all information, including the within-study correlation) and they show that the fixed-effects model can seriously underestimate the standard errors of the parameter estimates when a random-effects model should have been used [151]. An extension of Model A to bivariate random-effects meta-regression will be introduced in Chapter 6.7.4.
Sohn also performed simulations to compare the following four different models [177]:

1. A trivariate meta-analysis model (i.e. similar to Model A but with three outcomes rather than two), where the within-study correlations are known.
2. Three independent univariate random-effects meta-analysis (URMA) models.
3. Three independent univariate fixed-effects meta-analysis models.
4. Three independent ordinary least squares models, i.e. not a meta-analysis model, but one where each study has equal weighting and there is only variation between studies.

Multiple datasets were simulated for 16 studies from (1), and then all four models were fitted to each dataset, with the average pooled estimates across studies then compared across the four models. Sohn concludes that (3) is a poor approximation to (1) because the coverage of the pooled estimates are too small, which concurs with the findings from the aforementioned simulation results of Berkey et al. [151]. However, Sohn also states that, for the situations considered, (2) and (4) are practically identical to (1), and therefore if the within-study correlation is unknown models (2) and (4) are good alternatives to use. This latter result is particularly surprising, but the reasons why Sohn reaches these conclusions will become apparent from my simulations below and I will discuss this again in Chapter 5.6.

5.2 Simulations of Model A for complete-case data from n = 50 studies

In Section 5.2.1 I will describe in detail the simulation procedure used to assess Model A in relation to a URMA. I want to firstly emphasise here that all the simulations I perform in this thesis are initiated from the summary statistics level (using Model A, i.e. from level 2 in equation (3.2)) rather than the IPD level (e.g. from level 1 in equation (3.2)); for instance, for simulations of Model A I will describe in Section 5.2 how I generate \( \tilde{Y}_{i1} \) and \( \tilde{Y}_{i2} \) for \( n \) studies directly from Model A using known \( s^2 \) s, known within and between-
study correlations and other known parameter values of $\beta_1 = 0$, $\beta_2 = 2$, $\tau_1^2 = 0.25$, and $\tau_2^2 = 0.25$. A similar approach to a simulation assessment of Model A is taken by both Sohn [177] and Berkey et al. [151] in their respective multivariate meta-analysis papers (see Section 5.1), as they also generate the summary statistics rather than the IPD. The generation of summary statistics (i.e. the $\tilde{Y}_i$ s and the $\tilde{Y}_{i2}$ s) directly from Model A is a general way of obtaining $\tilde{Y}_i$ s and $\tilde{Y}_{i2}$ s that are suitable for bivariate meta-analysis (i.e. as they are generated from Model A itself, one can be assured that they are correlated and have the bivariate normal distribution required for Model A to be suitable). However, if one wanted to be more specific and generate $\tilde{Y}_i$ s and $\tilde{Y}_{i2}$ s that closely reflect those summary statistics (e.g. the log-HR for OS and the log-HR for DFS) that arise from survival studies then, rather than directly generating these summary statistics for each study from Model A, it may be more appropriate to firstly generate the IPD for each study and to then use this raw data to estimate the $\tilde{Y}_i$ and $\tilde{Y}_{i2}$ required. The main reason for this is that inherent relationships more specific to summary statistics from survival data would be observed in the $\tilde{Y}_i$ s and $\tilde{Y}_{i2}$ s produced from IPD. For instance, for the synthesis of log-HRs, the size of $s_{ij}^2$ is likely to be related to the size of $\tilde{Y}_j$ because a $\tilde{Y}_j$ far away from the true underlying pooled value is more likely to have a relatively high $s_{ij}^2$ than those $\tilde{Y}_j$ very close to the pooled value. Furthermore, there may also be some correlation between $s_{i1}^2$ and $s_{i2}^2$ (i.e. between the standard errors of the two summary statistics), particularly for OS and DFS as the standard error for the log-HR takes into account the number of events and, by definition, the event of ‘death’ is likely to be incorporated in the DFS standard error as well as the OS standard error.
In the simulations in this thesis I have not generated $\bar{Y}_y$ s that are specific to survival data.

For instance, I do not include any correlation between the $\bar{Y}_y$ s and the $s_y^2$ s when generating the $\bar{Y}_y$ s from Model A (see Section 5.2.1), and I also assume independence between $s_{i1}^2$ and $s_{i2}^2$. I therefore acknowledge here that the results and conclusions from the simulations in this thesis only relate directly to summary statistics that have these properties, i.e. where the summary statistics (the $\bar{Y}_y$ s) and their standard error ($s_y^2$) are uncorrelated and the $s_{i1}^2$ and $s_{i2}^2$ are independent in each study.

The key question is therefore, can the simulation results and conclusions be generalised to other types of summary statistics, such as those more specific to survival studies? To answer this question one would ideally require further simulation assessments that involve generating IPD from survival studies and then using this IPD to obtain the $\bar{Y}_y$ s and $s_y^2$ s required. However, the generation of IPD is likely to be non-trivial and it would add complexity to the statistical modelling and increase the computational time required. In particular, one would most likely require the IPD to be simulated from a frailty model [156;157], and thus additional assumptions would be required regarding the distribution of survival times and the censoring exhibited by patients in each study. Such assumptions are not required when simply generating the summary statistics directly from Model A, as I have done in this thesis.

Although further research is undoubtedly required to ascertain the generalisability of the simulation results in this thesis to specific summary statistics from survival studies, there are two indications that generalisability is potentially plausible. Firstly, even if one did have correlated $\bar{Y}_y$ s and $s_y^2$ s across studies, the correlation between $\bar{Y}_y$ and $s_y^2$ is not actually taken into consideration by Model A. This is in contrast to other meta-analysis
models and methods where the correlation between $\bar{Y}_j$ and $s_{ij}^2$ is important and utilised.

One such example of this is the use of funnel plots to assess dissemination bias, where the $\bar{Y}_j$s are plotted against some function of their standard error, and so if no correlation is considered in the simulated $\bar{Y}_j$s and $s_{ij}^2$s then this will have a large impact on the relationships observed on the funnel plots that are produced (see Chapter 9). Similarly, Model A does not acknowledge the correlation between $s_{i1}^2$ and $s_{i2}^2$ even when it exists.

The second indication that generalisability is potentially plausible can be seen in Sections 8.5.1 and 8.5.2, where I specifically apply a number of BRMAs to the OS and DFS log-HRs extracted from the neuroblastoma prognostic marker studies of Chapter 2. Encouragingly the results I obtain from these BRMAs show consistency with the conclusions about the benefit of BRMA over URMA identified by the simulations throughout the thesis (e.g. when there is missing data across studies, the BRMA produces pooled estimates with considerably smaller standard error than those pooled estimates from the URMA). Further investigation as to the generalisability of the conclusions from the simulations in this thesis would make very pertinent further research, and to this end the simulation of IPD to specifically reflect survival studies may be particularly important.

I will now introduce and consider simulations of Model A and URMA for two situations, one involving $n = 50$ studies (see Section 5.2.1) and then one involving $n = 5$ studies (see Section 5.3.1). These relate to a very large meta-analysis and a very small (but more common) meta-analysis respectively, and for both situations two summary statistics are of interest from each study and there is no missing data.
5.2.1 The simulation procedure for n = 50 studies

The procedure used for the simulations of n = 50 studies is now outlined.

STEP 1

I sampled 50 $s_i^2$ s and 50 $s_{i2}^2$ s for the n = 50 studies from

\[
\begin{pmatrix}
\ln(s_i^2) \\
\ln(s_{i2}^2)
\end{pmatrix} \sim N\left(\begin{pmatrix} 0.25 \\ 0.25 \end{pmatrix}, \begin{pmatrix} 0.25 & 0 \\ 0 & 0.25 \end{pmatrix}\right)
\]

The minimum value of all the $s_i^2$ s was 0.0026, the maximum was 16.26, the mean was 1.07, the median was 0.25 and the range was similar for $s_i^2$ and $s_{i2}^2$. The average value of $s_i^2$ may be considered quite large but the range was chosen to reflect the size of $s_i^2$ and $s_{i2}^2$ from the OS and DFS loge(HR) estimates from the neuroblastoma review in Chapter 2.

STEP 2

On 1000 occasions I generated 50 $\widetilde{Y}_i$ s and 50 $\widetilde{Y}_{i2}$ s for each of four different settings separately. The only parameter values that were different across settings were the within-study correlation (assumed the same in each of the 50 studies, i.e. $\rho_{wi} = \rho_w$ for all i) and between-study correlation ($\rho_B = \tau_{12}/\tau_1\tau_2$); the other known parameter values stayed the same. These settings were:

Setting (i): $\rho_w = 0; \rho_B = 0; \beta_1 = 0, \beta_2 = 2, \tau_1^2 = 0.25, \tau_2^2 = 0.25$ and the 50 $s_i^2$ s and 50 $s_{i2}^2$ s from Step 1.

Setting (ii): $\rho_w = 0; \rho_B = 0.8; \beta_1 = 0, \beta_2 = 2, \tau_1^2 = 0.25, \tau_2^2 = 0.25$ and the 50 $s_i^2$ s and 50 $s_{i2}^2$ s from Step 1.

Setting (iii): $\rho_w = 0.8; \rho_B = 0; \beta_1 = 0, \beta_2 = 2, \tau_1^2 = 0.25, \tau_2^2 = 0.25$ and the 50 $s_i^2$ s and 50 $s_{i2}^2$ s from Step 1.
Setting (iv): \( \rho_w = 0.8, \rho_g = 0.8; \beta_1 = 0, \beta_2 = 2, \tau_1^2 = 0.25, \tau_2^2 = 0.25 \) and the 50 \( s_{i1}^2 \) s and 50 \( s_{i2}^2 \) s from Step 1

The simulations proceeded as follows. For each setting, 50 \( \theta_{i1} \) s and 50 \( \theta_{i2} \) s were randomly generated 1000 times using the 'rmvnorm' function in S-Plus assuming \( \theta_{ij} \sim N(\beta_j, \Omega_2) \) as in Model A [Appendix B5]. For example, this meant that for Setting (i) 50 \( \theta_{i1} \) s and 50 \( \theta_{i2} \) s were generated 1000 times using:

\[
\theta_{ij} \sim N(\beta_j, \Omega_2), \quad \text{where} \quad \beta_j = \begin{pmatrix} 0 \\ 2 \end{pmatrix} \quad \text{and} \quad \Omega_2 = \begin{pmatrix} 0.25 & 0 \\ 0 & 0.25 \end{pmatrix}
\]

Then for each of the four settings separately, using the 50 \( \theta_{i1} \) s and the 50 \( \theta_{i2} \) s, and assuming the 50 \( s_{i1}^2 \) s and the 50 \( s_{i2}^2 \) s from Step 1 were known, I generated 50 \( Y_{i1} \) s and 50 \( Y_{i2} \) s 1000 times using the 'rmvnorm' function in S-Plus assuming \( Y_{ij} \sim N(\theta_{ij}, \delta_i) \) as in Model A [Appendix B5]. For example, for Setting (i) 50 \( Y_{i1} \) s and 50 \( Y_{i2} \) s were generated 1000 times using:

\[
Y_{ij} \sim N(\theta_{ij}, \delta_i), \quad \text{where} \quad \delta_i = \begin{pmatrix} s_{i1}^2 & 0 \\ 0 & s_{i2}^2 \end{pmatrix},
\]

and \( s_{ij}^2 \) and \( \theta_{ij} \) were known from Steps 1 and 2 respectively.

N.B. I emphasise here that the same 50 \( s_{i1}^2 \) s and the same 50 \( s_{i2}^2 \) s were used to generate the 50 \( Y_{i1} \) s and 50 \( Y_{i2} \) s on each of the 1000 occasions in each of the four settings.

**STEP 3**

Using the \( \bar{Y}_{i1} \) s and \( \bar{Y}_{i2} \) s generated and the \( s_{i1}^2 \) s, \( s_{i2}^2 \) s and \( \rho_w \) assumed known, both a URMA and Model A were fitted using SAS Proc Mixed to each of the 1000 simulations.
for each of Settings (i) to (iv), and the unknown parameters (i.e. $\beta_1$, $\beta_2$, $\tau_1^2$, $\tau_2^2$, $\rho_8$) were estimated using RIGLS each time. Hence at the end of Step 3, assuming all simulations converged, 1000 estimates for each of the parameters would be available from Model A and similarly 1000 estimates for each of the parameters would be available from the URMA.

**STEP 4**

For each of the four settings, the parameter estimates from the simulations were compared to the known parameter values from which the data were generated, and then the parameter estimates from the URMA were directly compared to those from Model A. To do this I calculated the:

- average parameter estimates across all the simulations (to check for bias)
- coverage of the 95% confidence intervals for the pooled estimates
- average variance and MSE of the pooled estimates

I also recorded other pertinent information such as the number of occasions where $\hat{\tau}_j^2 = 0$ in each model.

### 5.2.2 Degrees of freedom for the coverage of the pooled estimates

One comparison of interest was the coverage of the 95% confidence intervals for $\beta_j$. I calculated such confidence intervals using:

$$\beta_j \pm \left( t_{n_j-1}(0.05) \times \sqrt{\text{var}(\beta_j)} \right)$$

where $n_j$ is the number of studies for which outcome $j$ is available. The use of the $t$-distribution, rather than a normal distribution, is recommended in small samples, especially because $\text{var}(\beta_j)$ is based on $\hat{\tau}_1^2$ and $\hat{\tau}_2^2$ which are estimates themselves (N.B. The $t$-distribution does not take into account the uncertainty in $\hat{\tau}_j^2$; however, a Bayesian
approach is particularly suitable for this – see Chapter 7) [178]. For both Model A and the URMA model, the degrees of freedom used was \((n_j - 1)\) and this is equivalent to the total number of studies minus the number of fixed-effect parameters for outcome \(j\). However this is only an approximation, because the true degrees of freedom are very complex and will take into account the individual study weights (i.e. the \(s^2_n\) and the \(s^2_i\)) [151; 178-181]. Indeed, as the simulations will show, there are occasions where the size of the \(s^2_j\)s will ensure that \(\hat{\tau}^2\) and \(\hat{\tau}^2\) are poorly estimated, and in such situations the degrees of freedom used above may not be a very good approximation. In other situations, for example in a meta-analysis of a large number of studies, \(\hat{\tau}^2\) and \(\hat{\tau}^2\) will be more well-defined and the approximate degrees of freedom used will be much closer to the true value. The use of this approximate \(t\)-distribution is common practice in the meta-analysis literature, however I acknowledge that it is not ideal.

5.2.3 Simulation results and assessment of URMA and Model A for \(n = 50\) studies

The simulation results are shown in Table 5.1 for each of Settings (i) to (iv) as described in Section 5.2.1. Looking at the average parameter estimates across all simulations, the URMA model gives mean values of \(\hat{\beta}_1\), \(\hat{\beta}_2\), \(\hat{\tau}^2\) and \(\hat{\tau}^2\) that are very close to the true values in all of Settings (i) to (iv), as is to be expected as RIGLS estimation should produce unbiased parameter estimates in theory [51]. The coverage is also close to 95% for the pooled estimates. The unbiased pooled estimates and the apt coverage highlights why URMA is a suitable method for forming evidence-based results given complete-case data, and also emphasises the need to assess the benefits and limitations of Model A in relation to the URMA approach.
Table 5.1: Simulation results from the univariate (URMA) and bivariate (Model A) random-effects meta-analysis models for complete-case data from 50 studies. A description of the simulation procedure is given in Section 5.2.1, and the true parameter values to compare the results to are $\beta_1 = 0$, $\beta_2 = 2$, $\tau_1^2 = 0.25$, $\tau_2^2 = 0.25$.

In Model A the within-study correlation ($\rho_w$) was known and the same for each study, whilst the between-study correlation ($\rho_B$) was estimated. The true values of $\rho_w$ and $\rho_B$ in each simulation are shown in the table.

<table>
<thead>
<tr>
<th>Meta-analysis model</th>
<th>$\rho_w$</th>
<th>$\rho_B$</th>
<th>Converged out of 1000</th>
<th>Mean of $\hat{\beta}_1$ (s.e. of mean)</th>
<th>Mean/ Med s.e. of $\hat{\beta}_1$</th>
<th>MSE of $\hat{\beta}_1$</th>
<th>Mean of $\hat{\beta}_2$ (s.e. of mean)</th>
<th>Mean/ Med s.e. of $\hat{\beta}_2$</th>
<th>MSE of $\hat{\beta}_2$</th>
<th>No. of 95% CIs including $\beta_1$ (%)</th>
<th>Mean/ Med $\hat{\tau}_1^2$ (no. of $\hat{\tau}_1^2 = 0$)</th>
<th>Mean/ Med $\hat{\tau}_2^2$ (no. of $\hat{\tau}_2^2 = 0$)</th>
<th>Mean of $\hat{\rho}_B$</th>
<th>$\hat{\rho}_B = -1$ (%)</th>
<th>$\hat{\rho}_B = 1$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>URMA</td>
<td>0</td>
<td>0</td>
<td>1000</td>
<td>-0.0050 (0.101)</td>
<td>0.102/ 0.103</td>
<td>948 (94.8%)</td>
<td>2.001 (0.104)</td>
<td>0.107/ 0.107</td>
<td>0.0108</td>
<td>954 (95.4%)</td>
<td>0.247/ 0.237 (0)</td>
<td>0.255/ 0.246 (1)</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Model A</td>
<td>0</td>
<td>0</td>
<td>1000</td>
<td>-0.0048 (0.102)</td>
<td>0.101/ 0.104</td>
<td>947 (94.7%)</td>
<td>2.001 (0.104)</td>
<td>0.106/ 0.107</td>
<td>0.0108</td>
<td>955 (95.5%)</td>
<td>0.247/ 0.235 (0)</td>
<td>0.255/ 0.246 (0)</td>
<td>-0.001 2 (0.2%)</td>
<td>4 (0.4%)</td>
<td></td>
</tr>
<tr>
<td>URMA</td>
<td>0</td>
<td>0.8</td>
<td>1000</td>
<td>-0.0044 (0.101)</td>
<td>0.102/ 0.102</td>
<td>953 (95.3%)</td>
<td>2.001 (0.107)</td>
<td>0.106/ 0.106</td>
<td>0.0114</td>
<td>944 (94.4%)</td>
<td>0.250/ 0.241 (0)</td>
<td>0.246/ 0.241 (1)</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Model A</td>
<td>0</td>
<td>0.8</td>
<td>1000</td>
<td>-0.0040 (0.0999)</td>
<td>0.100/ 0.100</td>
<td>953 (95.3%)</td>
<td>2.001 (0.104)</td>
<td>0.104/ 0.104</td>
<td>0.0107</td>
<td>954 (95.4%)</td>
<td>0.251/ 0.243 (0)</td>
<td>0.249/ 0.243 (0)</td>
<td>0.795 252 (0%)</td>
<td>0 (25.2%)</td>
<td></td>
</tr>
<tr>
<td>URMA</td>
<td>0.8</td>
<td>0</td>
<td>1000</td>
<td>0.0018 (0.105)</td>
<td>0.101/ 0.102</td>
<td>941 (91.1%)</td>
<td>2.003 (0.105)</td>
<td>0.106/ 0.105</td>
<td>0.0113</td>
<td>951 (95.1%)</td>
<td>0.245/ 0.238 (0)</td>
<td>0.246/ 0.235 (1)</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Model A</td>
<td>0.8</td>
<td>0</td>
<td>1000</td>
<td>0.0027 (0.103)</td>
<td>0.0987/ 0.0992</td>
<td>937 (93.7%)</td>
<td>2.003 (0.105)</td>
<td>0.103/ 0.102</td>
<td>0.0109</td>
<td>948 (94.8%)</td>
<td>0.245/ 0.238 (0)</td>
<td>0.245/ 0.235 (0)</td>
<td>-0.080 23 (2.3%)</td>
<td>0 (9%)</td>
<td></td>
</tr>
<tr>
<td>URMA</td>
<td>0.8</td>
<td>0.8</td>
<td>1000</td>
<td>-0.0044 (0.101)</td>
<td>0.102/ 0.102</td>
<td>945 (95.4%)</td>
<td>1.993 (0.107)</td>
<td>0.106/ 0.106</td>
<td>0.0114</td>
<td>946 (94.6%)</td>
<td>0.248/ 0.240 (0)</td>
<td>0.247/ 0.238 (0)</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Model A</td>
<td>0.8</td>
<td>0.8</td>
<td>1000</td>
<td>-0.0031 (0.094)</td>
<td>0.095/ 0.095</td>
<td>951 (95.1%)</td>
<td>1.995 (0.097)</td>
<td>0.0972/ 0.0966</td>
<td>0.0094</td>
<td>939 (93.9%)</td>
<td>0.250/ 0.240 (0)</td>
<td>0.248/ 0.235 (0)</td>
<td>0.786 0 (2.0%)</td>
<td>20 (0%)</td>
<td></td>
</tr>
</tbody>
</table>

N.B. $\hat{\rho}_B$ was not applicable when one or both of $\hat{\tau}_1^2$ and $\hat{\tau}_2^2$ was zero. MSE = mean-square-error, CIs = confidence intervals, s.e. = standard error

The 95% CIs for the pooled estimates were calculated using a t-distribution with 49 degrees of freedom.
Model A also produces mean values of $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\tau}^2_1$, $\hat{\tau}^2_2$, and now $\hat{\rho}_B$ that are very close to their true values for each of Settings (i) to (iv) (Table 5.1). However, one worrying aspect is seen in Setting (ii) (i.e. for $\rho_w = 0$ and $\rho_B = 0.8$) where a relatively large proportion (25%) of the $\hat{\rho}_B$ values are equal to 1. On average, the gain in precision and reduction in MSE of the pooled estimates from Model A is negligible in all situations when compared to the URMA model (Table 5.1, Figure 5.1). This is perhaps expected because the number of studies is very large. Although still negligible, the benefits of Model A are best where both $\rho_w$ and $\rho_B$ are high.

**Figure 5.1:** Univariate versus bivariate (Model A) random-effects meta-analysis results for outcome $j = 1$ for (i) $\hat{\beta}_1$, (ii) s.e.( $\hat{\beta}_1$) and (iii) $\hat{\tau}^2_1$ from 1000 simulations of 50 studies as described in Section 5.2.1, with results shown in Table 5.1. The true value of $\beta_1$ was 0, the true value of $\tau^2_1$ was 0.25, and for Model A the true within- and between-study correlations were 0.8. The solid lines show line of equality, dotted lines show standard linear regression line through the points.
5.3 Simulations of Model A for complete-case data from \( n = 5 \) studies

5.3.1 The simulation procedure for \( n = 5 \) studies

The simulation procedure of Section 5.2.1 for Settings (i) to (iv) was repeated, except now only 5 studies were used in each meta-analysis instead of 50. Hence, 5 \( \tilde{Y}_n \) and 5 \( \tilde{Y}_{n+} \) were simulated 1000 times for each of Settings (i) to (iv), using 5 \( s^2_i \) and 5 \( s^2_{i+} \) sampled as in ‘Step 1’ of Section 5.2.1 and true parameter values as before of \( \beta_1 = 0, \beta_2 = 2, \tau_1^2 = 0.25, \) and \( \tau_2^2 = 0.25 \) (N.B. the same 5 \( s^2_i \) and \( s^2_{i+} \) were used in each of the \( n = 5 \) simulations considered). The 5 \( s^2_i \) had a minimum value of 0.108, a maximum of 0.200, a mean of 0.155 and a median of 0.154; the 5 \( s^2_{i+} \) had a minimum value of 0.048, a maximum of 0.743, a mean of 0.245 and a median of 0.141.

5.3.2 URMA simulation results and assessment for \( n = 5 \) studies

Averaging across all simulations, the URMA model gives mean values of \( \tilde{\beta}_1 \) and \( \tilde{\beta}_2 \) very close to the true values \( \beta_1 \) and \( \beta_2 \) in each of Settings (i) to (iv) (Table 5.2), again indicating the suitability of URMA for evidence synthesis. The mean values of \( \tilde{\tau}_j^2 \) are upwardly biased on average, inflated beyond their true value. However, this is simply due to the large number of \( \tilde{\tau}_j^2 \) that are set to zero, rather than being given a negative value (which can sometimes happen when there are a small number of studies, as by chance the overall variation may be less than that specified by the known \( s^2_i \) and \( s^2_{i+} \)). Hence, the upward bias in \( \tilde{\tau}_j^2 \) is a consequence of the constraint on the lower bound of \( \tilde{\tau}_j^2 \), and it is not indicative that URMA is an inappropriate model (the \( n = 50 \) results in Section 5.2.3 show that URMA is suitable for evidence synthesis of complete-case data, and it is a commonly applied model in the meta-analysis literature [43]).
Table 5.2: Simulation results from the univariate (URMA) and bivariate (Model A) random-effects meta-analysis models for complete-case data from 5 studies. A description of the simulation procedure is given in Section 5.3.1, and the true parameter values to compare the results to are $\beta_1 = 0$, $\beta_2 = 2$, $\tau_1^2 = 0.25$, $\tau_2^2 = 0.25$.

In Model A the within-study correlation ($\rho_w$) was known and the same for each study, whilst the between-study correlation ($\rho_B$) was estimated. The values of $\rho_w$ and $\rho_B$ in each simulation are shown in the table.

The 95% CIs for the pooled estimates were calculated using a t-distribution with 4 degrees of freedom.

<table>
<thead>
<tr>
<th>Meta-analysis model</th>
<th>$\rho_w$</th>
<th>$\rho_B$</th>
<th>Converged out of 1000</th>
<th>Mean of $\hat{\beta}_1$ (s.e. of mean)</th>
<th>Mean of $\hat{\beta}_2$ (s.e.)</th>
<th>MSE of $\hat{\beta}_1$</th>
<th>No. of 95% CIs for $\hat{\beta}_1$ including $\beta_1$</th>
<th>MSE of $\hat{\beta}_2$</th>
<th>No. of 95% CIs for $\hat{\beta}_2$ including $\beta_2$</th>
<th>Mean of $\tau_1^2$ (no. of $\tau_1^2 = 0$)</th>
<th>Mean of $\tau_2^2$ (no. of $\tau_2^2 = 0$)</th>
<th>Mean of $\rho_B$</th>
<th>No. of $\rho_B = -1$ (%)</th>
<th>No. of $\rho_B = 1$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>URMA</td>
<td>0</td>
<td>0</td>
<td>1000</td>
<td>-0.0015 (0.285)</td>
<td>0.267 (0.257)</td>
<td>0.0810 (96.0%)</td>
<td>1.994 (0.298)</td>
<td>0.267 (0.259)</td>
<td>0.0887 (94.0%)</td>
<td>0.264 (0.289)</td>
<td>0.256 (0.181)</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Model A</td>
<td>0</td>
<td>0</td>
<td>998</td>
<td>-0.0020 (0.285)</td>
<td>0.274 (0.261)</td>
<td>0.0809 (96.7%)</td>
<td>1.994 (0.299)</td>
<td>0.269 (0.257)</td>
<td>0.0894 (95.3%)</td>
<td>0.256 (0.191)</td>
<td>0.274 (0.187)</td>
<td>-0.027 (29.6%)</td>
<td>295</td>
<td>289</td>
</tr>
<tr>
<td>URMA</td>
<td>0</td>
<td>0.8</td>
<td>999</td>
<td>-0.0015 (0.277)</td>
<td>0.271 (0.259)</td>
<td>0.0769 (97.3%)</td>
<td>1.995 (0.288)</td>
<td>0.263 (0.251)</td>
<td>0.0826 (94.2%)</td>
<td>0.254 (0.185)</td>
<td>0.255 (0.168)</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Model A</td>
<td>0</td>
<td>0.8</td>
<td>1000</td>
<td>-0.0017 (0.277)</td>
<td>0.279 (0.266)</td>
<td>0.0767 (98.1%)</td>
<td>1.992 (0.286)</td>
<td>0.268 (0.253)</td>
<td>0.0819 (95.7%)</td>
<td>0.274 (0.204)</td>
<td>0.274 (0.181)</td>
<td>0.639 (10.3%)</td>
<td>605</td>
<td>-</td>
</tr>
<tr>
<td>URMA</td>
<td>0.8</td>
<td>0.8</td>
<td>1000</td>
<td>-0.0083 (0.289)</td>
<td>0.271 (0.257)</td>
<td>0.0834 (97.9%)</td>
<td>1.984 (0.303)</td>
<td>0.262 (0.250)</td>
<td>0.0921 (93.2%)</td>
<td>0.254 (0.179)</td>
<td>0.251 (0.166)</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Model A</td>
<td>0.8</td>
<td>0.8</td>
<td>998</td>
<td>-0.0094 (0.284)</td>
<td>0.262 (0.250)</td>
<td>0.0805 (97.7%)</td>
<td>1.984 (0.285)</td>
<td>0.254 (0.243)</td>
<td>0.0870 (92.9%)</td>
<td>0.262 (0.193)</td>
<td>0.259 (0.184)</td>
<td>0.651 (12.4%)</td>
<td>535</td>
<td>-</td>
</tr>
<tr>
<td>URMA</td>
<td>0.8</td>
<td>0</td>
<td>999</td>
<td>0.0040 (0.285)</td>
<td>0.268 (0.255)</td>
<td>0.0814 (96.2%)</td>
<td>2.004 (0.283)</td>
<td>0.269 (0.256)</td>
<td>0.0803 (95.5%)</td>
<td>0.246 (0.174)</td>
<td>0.269 (0.180)</td>
<td>0.195 (89)</td>
<td>289</td>
<td>261</td>
</tr>
<tr>
<td>Model A</td>
<td>0.8</td>
<td>0</td>
<td>997</td>
<td>0.0061 (0.289)</td>
<td>0.272 (0.250)</td>
<td>0.0826 (97.1%)</td>
<td>2.004 (0.289)</td>
<td>0.272 (0.257)</td>
<td>0.0832 (95.4%)</td>
<td>0.269 (0.196)</td>
<td>0.287 (0.191)</td>
<td>-0.073 (35.9%)</td>
<td>358</td>
<td>174</td>
</tr>
</tbody>
</table>

N.B. $\hat{\rho}_B$ was not applicable when one or both of $\tau_1^2$ and $\tau_2^2$ was zero. MSE = mean-square-error, CIs = confidence intervals, s.e. = standard error.
Even though the mean values of $\hat{\tau}_j^2$ from the URMA are slightly too high, of most importance for this thesis will be how they compare to the values of the equivalent $\hat{\tau}_j^2$ from Model A (see Section 5.3.3 onwards).

Interestingly the coverage of the URMA pooled estimates is slightly different for each outcome, with $\hat{\beta}_1$ having a coverage slightly too large (> 95%) and $\hat{\beta}_2$ a coverage slightly too small (< 95%) on the whole. For $\hat{\beta}_2$, one possible explanation why the coverage is slightly too small is that, although on average the $\hat{\tau}_j^2$ s are upwardly biased, the distribution of $\hat{\tau}_2^2$ from all the 1000 simulations is positively skewed and therefore more than 50% of the $\hat{\tau}_2^2$ are actually smaller than the true value of 0.25, thus leading to confidence intervals that are on average also too small (see the median values of $\hat{\tau}_2^2$ in Table 5.2 for evidence of this skewness). This can also be seen, but to a lesser extent, for many of the $n = 50$ results (Table 5.1). As discussed in Chapter 5.2.2, to calculate the coverage of the pooled estimates I used an approximate $t$-distribution and it appears that this approximation is worse where $\hat{\tau}_j^2$ is poorly defined (emphasised by the skewed distribution of $\hat{\tau}_j^2$ across simulations), and this possibly explains why there are slight discrepancies in the expected coverage.

For $j = 1$, the mean and median values of the $s_i^2$ s I sampled are very similar and this subsequently leads to a distribution of the $\hat{\tau}_1^2$ across the simulations that is more symmetrical than for $j = 2$, with the mean and median values of $\hat{\tau}_1^2$ much closer together than those for $\hat{\tau}_2^2$. These small differences in the $s_i^2$ s (and therefore the weighting) between studies are potentially why a more well-defined estimate of the between-study variance is obtained for $j = 1$, thus also removing the problem of slightly small coverage as
described for outcome \( j = 2 \). So why is the coverage of \( \hat{\beta}_1 \) slightly greater than 95% on average? One possible explanation is that the mean value of \( \hat{\tau}_1^2 \) is also slightly too high across studies, due to the number of \( \hat{\tau}_1^2 \) set to zero, and this inflation in \( \hat{\tau}_1^2 \) may be the cause of the wider confidence intervals. Although these explanations are important, from this point forward, for both \( j = 1 \) and \( j = 2 \), the key interest for this thesis is how the coverage of \( \hat{\beta}_j \) from Model A compares to the coverage of \( \hat{\beta}_j \) from the URMA.

5.3.3 Model A simulation results and assessment for \( n = 5 \) studies

Immediately apparent from the Model A simulation results is that \( \hat{\rho}_g \) is commonly equal to 1 in each of the Settings (i) to (iv), even where the true \( \rho_w \) and \( \rho_g \) are zero; similarly a value of –1 is obtained a large proportion of the time (Table 5.2). This problem leads to the mean value of \( \hat{\rho}_g \) across all simulations to be somewhat different than the true value \( \rho_g \), and worryingly it is also associated with a large upward bias in \( \hat{\tau}_1^2 \) and \( \hat{\tau}_2^2 \). This upward bias is much further than the upward bias from the URMA estimates (see Section 5.3.2), and this is despite a much smaller number of \( \hat{\tau}_1^2 \) and \( \hat{\tau}_2^2 \) being set to zero due to the constraint on the lower bound of \( \hat{\tau}_j^2 \). This is disconcerting because, unlike for the URMA, the upward bias in \( \hat{\tau}_j^2 \) observed for Model A is not just because of the constraint imposed on the lower bound for \( \hat{\tau}_j^2 \). Even though \( \hat{\beta}_j \) is still unbiased across the Model A simulations, the inflation of \( \hat{\tau}_1^2 \) and \( \hat{\tau}_2^2 \) across simulations leads to a coverage of \( \hat{\beta}_j \) that is commonly greater than 95% in Model A and which is often much larger than that from the URMA.
As was the case for $n = 50$ studies, the gain in precision and reduction of MSE from Model A for $\hat{\beta}_1$ and $\hat{\beta}_2$ is still very small, although most evident when $\rho_w$ and $\rho_B$ are both high (Setting (iv): $\rho_w = \rho_B = 0.8$). However, in the other Settings (i) to (iii), the average standard error and MSE of $\hat{\beta}_j$ across simulations is actually larger in Model A than in the URMA. This appears to contradict Remarks 1 and 2 of Section 4.8.1 and Section 4.8.2, which were derived analytically. This problem is again a consequence of the upward bias in the RIGLS estimates of $\tau^2_j$. I noted in Section 4.8 that the theory behind RIGLS estimation suggests that the RIGLS estimate $\hat{\tau}^2_j$ should be unbiased in both a URMA and Model A [51], and thus both models should give the correct underlying value of $\tau^2_j$ on average. Remark 1 and Remark 2 were derived conditional on this assumption being true.

The simulations have shown that this assumption appears to be true for the URMA when ignoring the artificial upward bias induced by the lower bound constraint on $\hat{\tau}^2_j$ (see Sections 5.2.3 and 5.3.2); however, this assumption is not true for Model A, which gives an upwardly biased $\hat{\tau}^2_j$ on average relating to the problem of $\hat{\rho}_B$ often being equal to 1 or -1 (which acts in addition to the artificial upward bias induced by the lower bound constraint on $\hat{\tau}^2_j$). Having a larger $\hat{\tau}^2_j$ in Model A than in the URMA on average allows the standard error and MSE of $\hat{\beta}_j$ to also be larger in Model A on average, thus causing Remarks 1 and 2 to fail.

Importantly, the observation that there is still only very small benefit of Model A over the URMA, in terms of gain in precision and reduction in MSE of $\hat{\beta}_j$, is still true even when comparing just the subset of simulations for which Model A did not give $\hat{\rho}_B$ equal to either 1 or -1 (Table 5.3).
Table 5.3: Comparison of univariate (URMA) versus bivariate (Model A) random-effects meta-analysis simulation results just for the 339 simulations where \( \hat{\rho}_B \) was not equal to either 1 or -1 for the complete-case results for \( n = 5 \) and setting \( \rho_w = \rho_B = 0.8 \) (for full results see Table 5.2).

<table>
<thead>
<tr>
<th>Meta-analysis model</th>
<th>Mean of ( \tilde{\beta}_1 ) (s.e. of mean)</th>
<th>Mean s.e. of ( \tilde{\beta}_1 )</th>
<th>MSE of ( \tilde{\beta}_1 )</th>
<th>Mean of ( \tilde{\beta}_2 ) (s.e. of mean)</th>
<th>Mean s.e. of ( \tilde{\beta}_2 )</th>
<th>MSE of ( \tilde{\beta}_2 )</th>
<th>Mean ( \tau^2_1 )</th>
<th>Mean ( \tau^2_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>URMA</td>
<td>-0.009 (0.016)</td>
<td>0.299</td>
<td>0.0834</td>
<td>1.978 (0.016)</td>
<td>0.302</td>
<td>0.0895</td>
<td>0.331</td>
<td>0.339</td>
</tr>
<tr>
<td>Model A</td>
<td>-0.011 (0.016)</td>
<td>0.294</td>
<td>0.0830</td>
<td>1.975 (0.016)</td>
<td>0.299</td>
<td>0.0880</td>
<td>0.336</td>
<td>0.348</td>
</tr>
</tbody>
</table>

5.4 Simulations of complete-case data from \( n = 5 \) and \( n = 50 \) studies for low and high values of \( \tau^2_1 \) and \( \tau^2_2 \)

Section 4.1 showed that the variance-covariance matrix (V) of \( Y \) is formed by the sum of the within- and between-study parameters (equation (4.2)). Hence, it would seem particularly important to consider simulations of Model A for different ratios of the between-study variance to the within-study variance.

5.4.1 The simulation procedure for investigating various ratios of \( \tau^2_j \) to \( s^2_{ij} \)

The simulations considered so far, for both \( n = 5 \) and \( n = 50 \) studies, have used \( s^2_{ii} \)'s and \( s^2_{ij} \)'s with a median value reasonably close to the true value of 0.25 for \( \tau^2_1 \) and \( \tau^2_2 \). Keeping the same \( s^2_{ii} \)'s and \( s^2_{ij} \)'s as used previously for \( n = 5 \) and \( n = 50 \) studies, I now considered simulations as for Section 5.2.1 and Section 5.3.1 but where the true \( \tau^2_1 \) and \( \tau^2_2 \) are much smaller (0.0025) and much larger (1.5) than the median \( s^2_{ii} \)'s and \( s^2_{ij} \)'s. As the simulations in Sections 5.2 and 5.3 clearly indicated that the largest benefits of Model A are when \( \rho_w \) and \( \rho_B \) are high, I decided to only consider Setting (iv) where \( \rho_w = \rho_B = 0.8 \), which is the setting most likely to show the benefits of Model A if there are any.
5.4.2 Simulation results for low $\tau_1^2$ and low $\tau_2^2$ in relation to the $s_{ij}^2$'s

These simulations were not performed for $n = 5$ because, even for $n = 50$, Model A often failed to converge (Table 5.4). This is a consequence of the $s_{ij}^2$'s and the $s_{ij}^2$'s being large in comparison to $\tau_1^2$ and $\tau_2^2$, and the RIGLS estimation fails to produce a converged value of $\hat{\tau}_1^2$, $\hat{\tau}_2^2$, and $\hat{\rho}_B$. As the $s_{ij}^2$'s are the major cause of the overall variation, it is hard for the RIGLS procedure to estimate the between-study parameters. This is related to the common problem of flat profile likelihoods for $\hat{\tau}_1^2$, $\hat{\tau}_2^2$, and $\hat{\rho}_B$ (see Section 5.6.2 for further discussion on this). A profile likelihood is obtained by maximising over the other parameters (e.g. $\beta_1$, $\beta_2$, $\tau_1^2$, and $\tau_2^2$) for each possible value of the parameter of interest (e.g. $\rho_B$). The poor estimation of the between-study parameters is highlighted further in Figure 5.2(i) by the very flat histogram of all the 1000 $\hat{\tau}_2^2$ values obtained from the 1000 simulations of Model A.

For the $n = 50$ simulations of Model A, $\hat{\rho}_B$ being equal to either 1 or $-1$ is now an even worse problem, with 78% of the simulations providing one of these two values. As a consequence, the average $\hat{\rho}_B$ across simulations is severely underestimating the true value $\rho_B$. Furthermore, the average $\hat{\tau}_1^2$ and $\hat{\tau}_2^2$ are again upwardly biased from Model A (Table 5.4(i)).
Table 5.4: Investigating the importance of the size of $\tau^2_j$ relative to the $s^2_j$s, as described in Section 5.4. Simulation results from the univariate (URMA) and bivariate (Model A) random-effects meta-analysis models for complete-case data for 50 and 5 studies for true parameter values of $\beta_1 = 0$, $\beta_2 = 2$, and (i) $\tau^2_1 = 0.0025$, (ii) $\tau^2_1 = \tau^2_2 = 1.5$, and (iii) $\tau^2_1 = 0.0025$, $\tau^2_2 = 1.5$. The 95% confidence intervals (CIs) for the pooled estimates were calculated using a t-distribution with (n-1) degrees of freedom.

(i) $\tau^2_1 = \tau^2_2 = 0.0025$ (on average much smaller than the $s^2_j$s):

<table>
<thead>
<tr>
<th>n</th>
<th>Meta-analysis model</th>
<th>$\rho_W$</th>
<th>$\rho_B$</th>
<th>Converged out of 1000</th>
<th>Mean of $\hat{\beta}_1$ (s.e. of mean)</th>
<th>Mean/ Med s.e. of $\hat{\beta}_1$</th>
<th>MSE of $\hat{\beta}_1$</th>
<th>No. of 95% Cls for $\beta_1$ including $\beta_1$ (%)</th>
<th>Mean of $\hat{\beta}_2$ (s.e.)</th>
<th>Mean/ Med s.e. of $\hat{\beta}_2$</th>
<th>MSE of $\hat{\beta}_2$</th>
<th>No. of 95% Cls for $\beta_2$ including $\beta_2$ (%)</th>
<th>Mean/ Med $\tau^2_1$ (no. of $\tau^2_1 = 0$)</th>
<th>Mean/ Med $\tau^2_2$ (no. of $\tau^2_2 = 0$)</th>
<th>Mean of $\hat{\rho}_B$</th>
<th>No. of $\hat{\rho}_B = -1$ (%)</th>
<th>No. of $\hat{\rho}_B = 1$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>URMA</td>
<td>0.8</td>
<td>0.8</td>
<td>998</td>
<td>0.0012/ (0.049)</td>
<td>0.0456/ 0.0385</td>
<td>0.00243</td>
<td>930 (93.2%)</td>
<td>2.003/ (0.065)</td>
<td>0.067/ 0.0636</td>
<td>0.00427</td>
<td>962 (96.4%)</td>
<td>0.0069/ 0 (229)</td>
<td>0.0133/ 0 (143)</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>50</td>
<td>Model A</td>
<td>0.8</td>
<td>0.8</td>
<td>951*</td>
<td>0.0013/ (0.039)</td>
<td>0.0357/ 0.0338</td>
<td>0.00157</td>
<td>864 (90.1%)</td>
<td>2.002/ (0.048)</td>
<td>0.048/ 0.0481</td>
<td>0.00233</td>
<td>901 (94.7%)</td>
<td>0.0052/ 0 (143)</td>
<td>0.007/ 0 (21.9)</td>
<td>0.359</td>
<td>208</td>
<td>534</td>
</tr>
</tbody>
</table>

* Convergence was quite difficult, even though there were 50 studies, because $\tau^2_j$ was small relative to the $s^2_j$s, and so $\hat{\tau}^2_j$ and $\hat{\rho}_B$ were very poorly defined and struggled to converge. Hence the simulations were not performed for $n = 5$.

(ii) $\tau^2_1 = 0.0025$ (on average much smaller than the $s^2_j$s), $\tau^2_2 = 1.5$ (on average larger than the $s^2_j$s):

<table>
<thead>
<tr>
<th>No. of studies</th>
<th>Meta-analysis model</th>
<th>$\rho_W$</th>
<th>$\rho_B$</th>
<th>Converged out of 1000</th>
<th>Mean of $\hat{\beta}_1$ (s.e. of mean)</th>
<th>Mean/ Med s.e. of $\hat{\beta}_1$</th>
<th>MSE of $\hat{\beta}_1$</th>
<th>No. of 95% Cls for $\beta_1$ including $\beta_1$ (%)</th>
<th>Mean of $\hat{\beta}_2$ (s.e.)</th>
<th>Mean/ Med s.e. of $\hat{\beta}_2$</th>
<th>MSE of $\hat{\beta}_2$</th>
<th>No. of 95% Cls for $\beta_2$ including $\beta_2$ (%)</th>
<th>Mean/ Med $\tau^2_1$ (no. of $\tau^2_1 = 0$)</th>
<th>Mean/ Med $\tau^2_2$ (no. of $\tau^2_2 = 0$)</th>
<th>Mean of $\hat{\rho}_B$</th>
<th>No. of $\hat{\rho}_B = -1$ (%)</th>
<th>No. of $\hat{\rho}_B = 1$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>URMA</td>
<td>0.8</td>
<td>0.8</td>
<td>1000</td>
<td>-0.0008/ (0.049)</td>
<td>0.0460/ 0.0385</td>
<td>0.0024</td>
<td>934 (93.6%)</td>
<td>2.004/ (0.203)</td>
<td>0.199/ 0.200</td>
<td>0.0413</td>
<td>947 (94.7%)</td>
<td>0.00738/ 0 (271)</td>
<td>1.495/ 0 (143)</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>50</td>
<td>Model A</td>
<td>0.8</td>
<td>0.8</td>
<td>980*</td>
<td>-0.0023/ (0.046)</td>
<td>0.0398/ 0.0356</td>
<td>0.0021</td>
<td>887 (90.5%)</td>
<td>2.001/ (0.198)</td>
<td>0.192/ 0.193</td>
<td>0.0392</td>
<td>928 (94.7%)</td>
<td>0.0089/ 1.511/ 0 (10.8)</td>
<td>1.502/ 0 (21.9)</td>
<td>0.564</td>
<td>106</td>
<td>512</td>
</tr>
</tbody>
</table>

* Even for $n = 50$, convergence was again difficult due to the low $\tau^2_j$ in relation the $s^2_j$s, and so $n = 5$ simulations were not considered here.
\( \tau_1^2 = \tau_2^2 = 1.5 \) (on average larger than the \( s_i^2 \) s)

<table>
<thead>
<tr>
<th>No. of studies (n)</th>
<th>Meta-analysis model</th>
<th>( \rho_W )</th>
<th>( \rho_B )</th>
<th>Converged out of 1000</th>
<th>Mean of ( \hat{\beta}_1 ) (s.e. of mean)</th>
<th>Mean of ( \hat{\beta}_1 ) (s.e.)</th>
<th>No. of 95% CIs for ( \beta_1 ) including ( \beta_1 ) (%)</th>
<th>Mean of ( \hat{\beta}_2 ) (s.e.)</th>
<th>MSE of ( \hat{\beta}_2 ) (s.e.)</th>
<th>No. of 95% CIs for ( \beta_2 ) including ( \beta_2 ) (%)</th>
<th>Mean of ( \hat{\rho}_B ) (no. of ( \hat{\tau}_2^2 = 0 ))</th>
<th>MSE of ( \hat{\rho}_B ) (no. of ( \hat{\tau}_2^2 = 0 ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>URMA</td>
<td>0.8</td>
<td>0.8</td>
<td>1000</td>
<td>-0.001 (0.195)</td>
<td>0.038 (95.6%)</td>
<td>1.995 (0.196)</td>
<td>0.0383 (94.9%)</td>
<td>1.515/1.510/1.486 (0)</td>
<td>0.0383 (94.9%)</td>
<td>1.515/1.510/1.486 (0)</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>Model A</td>
<td>0.8</td>
<td>0.8</td>
<td>1000</td>
<td>0.0010 (0.169)</td>
<td>0.0356 (95.7%)</td>
<td>1.995 (0.196)</td>
<td>0.0358 (95.6%)</td>
<td>1.509/1.514/1.469 (0)</td>
<td>0.0358 (95.6%)</td>
<td>1.509/1.514/1.469 (0)</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>URMA</td>
<td>0.8</td>
<td>0.8</td>
<td>1000</td>
<td>0.0189 (0.579)</td>
<td>0.335 (94.5%)</td>
<td>2.011 (0.588)</td>
<td>0.346 (94.1%)</td>
<td>1.512/1.525/1.238 (10)</td>
<td>0.346 (94.1%)</td>
<td>1.512/1.525/1.238 (10)</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Model A</td>
<td>0.8</td>
<td>0.8</td>
<td>1000</td>
<td>0.0183 (0.579)</td>
<td>0.335 (94.6%)</td>
<td>2.012 (0.590)</td>
<td>0.348 (94.1%)</td>
<td>1.511/1.522/1.216 (10)</td>
<td>0.348 (94.1%)</td>
<td>1.511/1.522/1.216 (10)</td>
<td></td>
</tr>
</tbody>
</table>

N.B. \( \hat{\rho}_B \) was not applicable when one or both of \( \hat{\tau}_1^2 \) and \( \hat{\tau}_2^2 \) was zero. MSE = mean-square-error, CIs = confidence intervals, s.e. = standard error.

Figure 5.2: (i) Distribution of \( \hat{\tau}_2^2 \) and \( \hat{\rho}_B \) from the Model A simulations of Table 5.4(i) for \( n = 50 \) where \( \hat{\tau}_2^2 \) is considerably smaller than the \( s_i^2 \) s, and the true values for \( \tau_2^2 \) and \( \rho_B \) are 0.0025 and 0.8 respectively.

(ii) Distribution of \( \hat{\tau}_2^2 \) and \( \hat{\rho}_B \) from the Model A simulations of Table 5.4(i) for \( n = 50 \) where \( \hat{\tau}_2^2 \) is considerably larger than the \( s_i^2 \) s, and the true values for \( \tau_2^2 \) and \( \rho_B \) are 1.5 and 0.8 respectively.
It is difficult to compare the simulation results from Model A to those from the URMA, as the URMA simulations have a large number of \( \hat{\tau}_1^2 \) and \( \hat{\tau}_2^2 \) equal to zero (due to the constraint on the lower bound for this parameter), and this itself causes a large upward bias in the mean values of \( \hat{\tau}_1^2 \) and \( \hat{\tau}_2^2 \). The \( \hat{\tau}_1^2 \) and \( \hat{\tau}_2^2 \) from Model A are also on average inflated above the true value (due to bias induced by the lower bound constraint and also the \( \hat{\rho}_B \) equal to 1 or -1 problem), although their average value across simulations is smaller than those from a URMA. The key observation to note from this is that, despite \( \hat{\tau}_1^2 \) and \( \hat{\tau}_2^2 \) being smaller in Model A on average, the gain in precision and reduction in MSE from Model A is still negligible. To look at this more I compared just those simulation results where neither model gave \( \hat{\tau}_1^2 \) or \( \hat{\tau}_2^2 \) equal to zero, and I found the benefit of Model A was still very small as the two models gave almost identical results. For the setting assessed, this indicates that Model A has negligible benefit over two independent URMAs for estimating \( \beta_j \) when there is complete-case data.

5.4.3 Simulation results for low \( \tau_1^2 \) and high \( \tau_2^2 \) in relation to the \( s_{\beta}^2 \)s

These simulations were again not performed for \( n = 5 \) due to the common problem of non-convergence. Even for \( n = 50 \), convergence was again difficult for Model A but not to the same extent as the situation just discussed in Section 5.4.2 (Table 5.4(ii)). The results show there is still the major problem of \( \hat{\rho}_B \) being equal to either 1 or -1, with 63% of simulations providing one of these values, again a slight improvement to that in Section 5.4.2. Both \( \hat{\tau}_1^2 \) and \( \hat{\tau}_2^2 \) are upwardly biased in relation to their true values of 0.0025 and 1.5 respectively. However, relative to these true values, the inflation would appear worse for \( \hat{\tau}_1^2 \) which is inflated 256% above its true value whereas \( \hat{\tau}_2^2 \) is only inflated 0.73% above its true value (Table 5.4(ii)). It would appear that the high \( \hat{\tau}_2^2 \) is better defined than the low \( \hat{\tau}_1^2 \) and is affected to a lesser degree by \( \hat{\rho}_B \) being either 1 or -1.
The benefit of using Model A over the URMA model, in terms of precision and MSE of $\hat{\beta}_j$, is again very small for both outcomes but interestingly slightly better for outcome $j = 1$, where the $s_i^2$'s are much larger than $\tau_i^2$. This relates back to Points (1) and (2) in Section 4.5, and to Remarks 4 and 5 in Sections 8.4 and 8.5, where it was shown that the larger the differences between the URMA weights (i.e. the differences between the $s_i^2$'s across studies) the more Model A can be beneficial (e.g. in terms of gain in precision) because it has greater ability to 'borrow strength' across outcomes. For outcome $j = 1$ all the $s_i^2$'s are large and therefore differences between them are also larger, thus allowing more 'borrowing of strength'; however, for $j = 2$ the $s_i^2$'s are all small and therefore have much smaller differences between them, thus limiting the opportunity to 'borrow strength'.

5.4.4 Simulation results for high $\tau_1^2$ and high $\tau_2^2$ in relation to the $s_i^2$'s

Both a URMA and Model A give mean values of $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\tau}_1^2$ and $\hat{\tau}_2^2$ across the simulations that are very close to their true values (Table 5.4(iii)). Furthermore, I noticed that Model A converged much faster and more easily than in previous simulations, and the profile likelihoods for $\hat{\tau}_1^2$, $\hat{\tau}_2^2$ and $\hat{\rho}_B$ are well-defined. To further emphasise the well-defined between-study parameter estimates, I present in Figure 5.2(ii) a clearly peaked histogram of all the 1000 $\hat{\tau}_2^2$ values obtained from the 1000 simulations (to be contrasted with Figure 5.2(i)). Interestingly, the problem of $\hat{\rho}_B$ being equal to either 1 or −1 has also disappeared for $n = 50$, and although it still occurs for $n = 5$ the related bias in $\hat{\tau}_1^2$, $\hat{\tau}_2^2$ and the coverage is now evidently small, because these are all close to their expected values.

Despite the good estimation and convergence properties, for both $n = 50$ and $n = 5$ the simulation results show that the benefit of using Model A to estimate $\beta_j$ over two
independent URMA$\bar{s}$s is again negligible. For example, the standard error and MSE are practically identical on average. This is because the small size of the $s_{i}^{2}$s in relation to $\hat{\tau}_{j}^{2}$ means that the within-study variation of the $\hat{Y}_{i}^{2}$s is negligible in relation to the between-study variation. Hence, there is a very similar weighting of each study toward the pooled estimates because the relatively large $\hat{\tau}_{j}^{2}$s dominate the study weights, and the $s_{i}^{2}$s play little part in the estimation. This concurs with Remark 4 in Section 4.8.4, and also with Points (1) and (2) in Section 4.5, which indicated that when there are small differences between the URMA study weights (i.e. $w_{i} = \left(s_{i}^{2} + \tau^{2}\right)^{-1}$) the Model A results will be very similar to those from a URMA.

5.5 Main conclusions from the complete-case data simulations of Model A

The primary finding from the $n = 50$ and $n = 5$ simulations is that for complete-case data there is negligible benefit of Model A over two independent URMA$\bar{s}$s in terms of estimating the $\beta_{j}$s themselves. This is true for a variety of ratios of the within- and between-study variances, and in Appendix B6 I show that this finding also holds when considering similar simulations for $n = 25$. There are perhaps two main reasons for the negligible benefit of Model A for complete-case data. Firstly, even when the assumption that $\hat{\tau}_{j}^{2}$ is unbiased in Model A appears true (e.g. Table 5.1), and so Remark 1 and Remark 2 in Sections 4.8.1 and 4.82 are true, the gain in precision and reduction in MSE are actually only very small for complete-case data. Secondly, these small benefits are often reduced further by the problem of estimating the between-study correlation in Model A. The simulations show that when $\hat{\rho}_{b}$ is equal to either 1 or -1 there is an upward bias in the $\hat{\tau}_{j}^{2}$s from Model A on average, and this inflation in the $\hat{\tau}_{j}^{2}$s subsequently causes an increase in the variance and MSE of the $\hat{\beta}_{j}$s on average, reducing the small benefits of
Model A further. The coverage of the $\hat{\beta}_j$ s from Model A is also affected by this, but it is difficult to assess exactly how because I have had to use an approximate $t$-distribution to calculate the confidence intervals. Although this distribution is a good approximation when parameter estimates are well-defined (e.g. when the $\hat{\tau}_j^2$ s are much larger than the $s^2$ s, see Section 5.4.4), when there are problems of flat likelihoods for $\hat{\rho}_B$ and bias in $\hat{\tau}_j^2$ these appear to weaken the approximation and make the coverage somewhat different than 95% on average (see results in Table 5.2), making an evaluation of the coverage in my simulations very difficult.

Placing the findings into a meta-analysis context, the URMA simulation results indicate that given complete-case data this model is suitable for evidence synthesis because it produces unbiased pooled estimates with suitable coverage (e.g. see Table 5.1). The bias sometimes observed in the URMA $\hat{\tau}_j^2$ s across simulations was simply induced by the lower bound constraint of zero on this parameter (see Section 5.3.2). In addition to the URMA approach, the bivariate random-effects meta-analysis (BRMA) approach of Model A would also appear suitable for evidence synthesis given complete-case data where $\hat{\rho}_B$ is not equal to either 1 or $-1$, as the model produces unbiased pooled estimates with suitable coverage (e.g. see Table 5.1). However, where Model A estimates $\rho_B$ as 1 or $-1$ one must be cautious about using any Model A results as they may involve an upwardly biased, non-zero $\hat{\tau}_j^2$ (which is therefore not simply because of the lower bound constraint of zero for this parameter; see Section 5.3.3). Changes in $\hat{\tau}_j^2$ can alter the weighting of each study in the meta-analysis and thus biased $\hat{\tau}_j^2$ s may lead to inappropriate (i.e. potentially biased) pooled estimates, and thus potentially misleading evidence-based conclusions; this issue will be discussed further in Section 5.6.4. However, it is also important to stress again two
important points here. Firstly, the simulation results provide *no* evidence of bias in any of the Model A parameter estimates when the problem of $\hat{\rho}_B$ being equal to either 1 or −1 does *not* occur (e.g. see Table 5.1, for Settings (i) and (iv), where $\hat{\rho}_B$ is rarely equal to 1 or −1). Secondly, for complete-case data there is still negligible benefit of Model A over URMA even for those situations where $\hat{\rho}_B$ is not equal to either 1 or −1 (see Table 5.3).

Finally, my complete-case simulation results explain why Sohn surprisingly states that three independent URMAs and even three independent standard GLS models are comparable to a trivariate random-effects meta-analysis model in terms of the parameter estimates and their precision (see Section 5.1) [177]. Firstly, the author only considers complete-case data, and so this concurs with my finding that there is little or no benefit of multivariate meta-analysis over independent URMAs for complete-case data when the $\hat{\rho}_j$'s are of interest. Furthermore, all the $s^2_j$'s used by Sohn are very small, with little difference between them. Hence, there will be little ‘borrowing of strength’ in Model A (see results from Section 5.4.4) and the URMA weights will all be similar, with the within-study variation (i.e. $s^2_j$'s) not very influential in the overall estimation of parameters. Indeed, non-influential within-study variation also indicates why Sohn reports that the trivariate model and the three URMAs provide similar results to three standard GLS models, because the GLS model itself does not have within-study variance parameters and only has between-study parameters (see Section 5.1).
5.6 A critical assessment of the reasons for and the impact of a between-study correlation estimate of 1 or $-1$

Given the extent of the problem in my simulations, it is perhaps surprising that an estimate of $\rho$ equal to 1 or $-1$ has only been reported twice in the multivariate meta-analysis literature that I am aware of [148; 169]. To look at this further I performed three separate simulations based on the $s^2$s and $\hat{\tau}^2$s reported in three datasets used in the literature so far [147; 148; 151], and I still observed $\hat{\rho}$ to be equal to either 1 or $-1$ for a large proportion of the simulations in each case. For example, I replicated the simulations of Berkey et al. and observed that Model A estimated $\rho$ as 1 or $-1$ for 54% of the simulations, which may be related to why 114 of the authors’ original 5000 simulations struggled to reach convergence (these simulations and their results will be considered in detail in Section 6.7.4) [151]. Thompson et al. provide an alternative parameterisation of the BRMA approach of Model A in order to remove the problem of needing to estimate $\rho$ (see Section 3.6.3), however their model requires strong additional assumptions, such as $\rho_w = 0$, and which may not be plausible in many situations [169].

Before considering whether or not to look at simulations of Model A involving missing data, it is clearly important to understand why $\rho$ is often estimated as 1 or $-1$. After an in-depth investigation of my simulation results, two potential reasons (‘Reason (I)’ and ‘Reason (II)’) have been identified. I emphasise the word ‘potential’ here because these conjectures need to be backed up using formal mathematics and thus must only be treated as two conceivable reasons why the problem is occurring.
5.6.1 **Reason (I) why $\rho_B$ is often estimated as 1 or -1:**

A value of $\hat{\rho}_B$ above 1 or below -1 is required to obtain the overall correlation observed between the $\tilde{Y}_{i1}$s and the $\tilde{Y}_{i2}$s.

I will now discuss the underlying philosophy behind why I think Reason (I) is one plausible explanation. Given the $\tilde{Y}_y$ s, Model A can estimate the overall variance of the $\tilde{Y}_{i1}$s, the overall variance of the $\tilde{Y}_{i2}$s, and the overall covariance between the $\tilde{Y}_{i1}$s and $\tilde{Y}_{i2}$s.

The specification of Model A decomposes the overall variances and overall covariance into the sum of the within- and between-study parameters (see matrix V in Section 4.1). As the within-study parameters ($s_{i1}^2$, $s_{i2}^2$ and $\rho_{wi}$) are known and therefore fixed, Model A only estimates the between-study parameters ($\tau_1^2$, $\tau_2^2$, and $\rho_B$) in order to obtain the overall variance and overall covariance indicated by the $\tilde{Y}_y$ s. In terms of the overall correlation this means that, as the within-study correlations are known and specified, Model A must estimate the between-study correlation ($\rho_B$) in order to obtain the overall correlation observed between the $\tilde{Y}_{i1}$s and the $\tilde{Y}_{i2}$s. However, a problem arises because $\hat{\rho}_B$ is not inherently defined between -1 and 1 by Model A, and it is only the Cholesky parameterisation of the parameters that prevents it from going beyond this range (see Section 4.2). Indeed, in order to obtain the overall correlation indicated by the $\tilde{Y}_y$ s, the RIGLS estimation procedure may seek a $\hat{\rho}_B$ above 1 or below -1, but obviously this is not allowed (as non positive-definite matrices would arise, see Section 4.2) and so the RIGLS estimate is constrained to be 1 or -1 respectively. For example, if the $\tilde{Y}_y$ s indicate a very high overall correlation but the within-study correlation is specified as highly negative, the RIGLS estimation could plausibly attempt to estimate $\rho_B$ greater than 1 in order to compensate for the low within-study correlation specified; however, the highest RIGLS value allowed for $\hat{\rho}_B$ is 1, essentially the maximum of the profile likelihood for $\hat{\rho}_B$ in the
range -1 to 1, and so this will be the converged value provided by Model A. The strong association between $\hat{\rho}_B$ and the overall correlation is illustrated in Table 5.5 by considering Pearson’s correlation coefficient between the $\hat{Y}_{i1}$s and the $\hat{Y}_{i2}$s as a measure of the overall correlation.

Table 5.5: Evidence of the strong relationship between the between-study correlation estimate ($\hat{\rho}_B$) from Model A and the estimate of the overall correlation between the $\hat{Y}_{i1}$s and the $\hat{Y}_{i2}$s as measured by Pearson’s correlation coefficient. The results are for the 1000 simulations of 5 studies in Table 5.2 where $\rho_w = 0.8$ and $\rho_B = 0.8$ (Setting (iv) of Section 5.3.1).

<table>
<thead>
<tr>
<th>Pearson correlation between the $\hat{Y}<em>{i1}$s and the $\hat{Y}</em>{i2}$s</th>
<th>Mean $\hat{\rho}_B$</th>
<th>No. of simulations</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt; 0.9</td>
<td>0.988</td>
<td>45</td>
</tr>
<tr>
<td>&gt; 0.5</td>
<td>0.671</td>
<td>373</td>
</tr>
<tr>
<td>&gt; 0</td>
<td>0.243</td>
<td>707</td>
</tr>
<tr>
<td>&lt; 0</td>
<td>-0.860</td>
<td>284</td>
</tr>
<tr>
<td>&lt; -0.2</td>
<td>-0.918</td>
<td>189</td>
</tr>
<tr>
<td>&lt; -0.5</td>
<td>-0.999</td>
<td>72</td>
</tr>
<tr>
<td>&lt; -0.9</td>
<td>-1</td>
<td>6</td>
</tr>
</tbody>
</table>

If Reason (I) is plausible, and thus the RIGLS procedure indeed desires to go beyond the range of -1 to 1 for $\hat{\rho}_B$, then one may also conceive that there is a ‘true’ maximum likelihood value for $\hat{\rho}_B$ above 1 or below -1, which the RIGLS procedure is trying to reach. This idea is strengthened by Figure 5.3 which shows a slight curvature in the profile log-likelihood of $\hat{\rho}_B$ for two $n = 5$ simulations, one where $\hat{\rho}_B$ is equal to -1 and one where $\hat{\rho}_B$ is equal to 1, and the log-likelihoods both appear to be tending toward an optimum point beyond 1 or -1. However, I emphasise that this is purely a conjecture and the possibility of a peak beyond the range -1 to 1 ideally needs to be considered analytically, perhaps by obtaining and maximising the mathematical expressions for the profile likelihood of $\hat{\rho}_B$. However, to further demonstrate why I think it is a plausible
hypothesis, I repeated the \( n = 50 \) complete-case simulations for Setting (iv) (where \( \rho_w = \rho_b = 0.8 \)) but deliberately wrongly set \( \rho_w \) equal to zero for all studies. This led to an enormous 85\% of the simulations producing \( \hat{\rho}_b \) equal to 1 (and none to \(-1\)), with the estimates trying to compensate for the under-specification of the within-study correlation and seeking the overall correlation indicated by the \( \hat{Y}_y \) s.

**Figure 5.3:** Profile log-likelihood for the estimate of the between-study correlation from two of the simulations ((i) id = 341, (ii) id = 646) where \( n = 5 \) and the within-study correlation was 0 and the true between-study correlation was 0.8 (see Table 5.2). In both, the profile likelihood is curved possibly tending to a maximum likelihood value for the between-study correlation beyond the upper limit of 1 (Figure (i)) or below the lower limit of \(-1\) (Figure (ii)).

5.6.2 **Reason (II) why \( \rho_B \) is often estimated as 1 or \(-1\):**

The maximum likelihood value of \( \hat{\rho}_B \) is 1 or \(-1\) because there is little information regarding the between-study parameters

In addition to Reason (I) I have also identified another potential explanation (denoted Reason (II)) why \( \rho_B \) is often estimated as 1 or \(-1\). For some datasets the information from the \( \hat{Y}_y \) s and the \( s_y \) s may be insufficient for the RIGLS procedure to ascertain what value \( \hat{\rho}_B \) should be. This is in contrast to the proposed Reason (I), where the ample information is dictating that the RIGLS procedure seeks a \( \hat{\rho}_B \) outside the range \(-1\) to 1. However, for both Reason (I) and Reason (II) the outcome is a converged value of \( \hat{\rho}_B \) as either 1 or \(-1\).
**Figure 5.4:** A common 'flat' profile log-likelihood for the between-study correlation from one of the \( n = 5 \) simulations where the within-study correlation was 0 and the true between-study correlation was 0.8 (see Table 5.2). The profile likelihood for the between-study correlation is flat and gradually increases toward 1 (in (i)) and toward −1 (in (ii)), caused by little and insufficient information about the between-study parameters from the data ('Reason (II)').

Reason (II) is essentially a problem of **flat profile likelihoods** (Figure 5.4), which is a common problem for other types of multi-level models that seek to utilise the correlation between outcomes (e.g. joint repeated measures and survival models) [182]. When \( \hat{\rho}_B \) has a flat profile likelihood there is no peaked or optimal point to clearly indicate the maximum likelihood value. However, as the profile likelihood is flat and gradually increases up to the value of 1 (Figure 5.4(i)) or down to the value of −1 (Figure 5.4(ii)), the converged RIGLS value of \( \hat{\rho}_B \) from Model A will simply be 1 or −1 respectively, i.e. the value that maximises the profile likelihood for \( \hat{\rho}_B \) in the range of −1 to 1 in this situation.

A flat profile likelihood for \( \hat{\rho}_B \) is associated with the problem of poorly defined \( \bar{\tau}_i^2 \) and \( \bar{\tau}_j^2 \) in Model A. For example, the simulation results indicate that when the \( s_{i1}^2 \) s and \( s_{i2}^2 \) s are large in relation to \( \bar{\tau}_i^2 \) and \( \bar{\tau}_j^2 \), \( \hat{\rho}_B \) is poorly estimated (see Section 5.4.2), even for a large (and in practice unrealistic) number of studies. For example, I repeated the simulations in Section 5.4.2 where the \( s_{i1}^2 \) s were large in relation to the \( \bar{\tau}_j^2 \) s but used 200 studies rather than 50, and found that 45% of the simulations still gave a \( \hat{\rho}_B \) equal to 1 or −1 despite the large number of studies available. This is again due to the relatively large \( s_{i1}^2 \) s and \( s_{i2}^2 \) s.
dominating the overall variance and thus there is little information to estimate the between-study parameters (see discussion in Section 5.4.2).

5.6.3 The impact of Reasons (I) and (II) on the other parameter estimates

I have presented two possible explanations for why $\rho_B$ is often estimated as 1 or -1 in Model A, and I want to emphasise that there is a clear difference between the proposed Reason (I) and Reason (II). The profile likelihood plots for Reason (II) in Figure 5.4 are very flat, and the log-likelihood is quite similar for $\hat{\rho}_B = 0$ compared to $\hat{\rho}_B = 1$ (a difference of about 0.5). This is in contrast to the slightly curved profile likelihood plots for Reason (I) in Figure 5.3, which give more noticeably different values of the log-likelihood for $\hat{\rho}_B = 0$ compared to $\hat{\rho}_B = 1$ (difference of about 2). This difference is important because it indicates that the impact of Reason (I) on the other parameter estimates (i.e. $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\tau}_1^2$ and $\hat{\tau}_2^2$) has the potential to be much greater than Reason (II).

The flat likelihood of Reason (II) implies that the other parameter estimates will be reasonably similar regardless of the value of $\hat{\rho}_B$, whereas the steeper and curved likelihood for Reason (I) implies the other parameter estimates may change more considerably. The small impact of Reason (II) is further justified by flat likelihoods for $\hat{\rho}_B$ being most commonly observed when $\hat{\tau}_1^2$ and $\hat{\tau}_2^2$ are small in relation to the within-study variances (Section 5.4.2); in this situation, the $s_i^2$'s dominate the overall variance and overall correlation, and therefore the between-study parameters have a very small impact on the final results, thus indicating that a $\hat{\rho}_B$ of 1 or -1 due to Reason (II) will also have small impact.

To further emphasise the small impact of Reason (II) compared to Reason (I), I performed a sensitivity analysis to two datasets, one where the proposed Reason (I) causes a $\hat{\rho}_B$ of -1
and one where the proposed Reason (II) causes a $\hat{\rho}_B$ of $-1$. The sensitivity assessment compared how the Model A results change for a range of different values of $\hat{\rho}_B$. For the dataset affected by Reason (II), the sensitivity results were very consistent for all values of $\hat{\rho}_B$ used. However, for the dataset affected by Reason (I), the sensitivity results showed far greater parameter changes across the different values of $\hat{\rho}_B$ (Table 5.6).

Table 5.6: Sensitivity analysis assessing Model A results across a range of different fixed $\hat{\rho}_B$ values for (i) a single dataset where Reason (I) is proposed to be a problem, and (ii) a single dataset where Reason (II) is proposed to be the problem. The two datasets (id no. 646 and no. 134 respectively) were taken from the 1000 datasets simulated by Model A for $n = 5$ and complete-case data, where $\rho_w = 0$ and $\rho_{B} = 0.8$ (as in Table 5.2). Both datasets gave a $\hat{\rho}_B$ of $-1$ when the between-study correlation was estimated by Model A.

<table>
<thead>
<tr>
<th>$\hat{\rho}_B$</th>
<th>Profile log- like</th>
<th>$\hat{\beta}_1$ (s.e.)</th>
<th>$\hat{\beta}_2$ (s.e.)</th>
<th>$\hat{\tau}_1^2$</th>
<th>$\hat{\tau}_2^2$</th>
<th>$\hat{\rho}_B$</th>
<th>Profile log- like</th>
<th>$\hat{\beta}_1$ (s.e.)</th>
<th>$\hat{\beta}_2$ (s.e.)</th>
<th>$\hat{\tau}_1^2$</th>
<th>$\hat{\tau}_2^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>-7.90</td>
<td>0.077</td>
<td>2.745</td>
<td>0.240</td>
<td>0.246</td>
<td>-1</td>
<td>-6.45</td>
<td>-0.107</td>
<td>2.161</td>
<td>0.0534</td>
<td>0.0867</td>
</tr>
<tr>
<td>(0.279)</td>
<td>(0.276)</td>
<td>(0.276)</td>
<td>(0.276)</td>
<td>(0.276)</td>
<td>(0.276)</td>
<td>(0.201)</td>
<td>(0.206)</td>
<td>(0.206)</td>
<td>(0.206)</td>
<td>(0.206)</td>
<td>(0.206)</td>
</tr>
<tr>
<td>-0.8</td>
<td>-8.05</td>
<td>0.078</td>
<td>2.749</td>
<td>0.221</td>
<td>0.221</td>
<td>-0.8</td>
<td>-6.45</td>
<td>-0.103</td>
<td>2.147</td>
<td>0.0507</td>
<td>0.0861</td>
</tr>
<tr>
<td>(0.273)</td>
<td>(0.270)</td>
<td>(0.270)</td>
<td>(0.270)</td>
<td>(0.270)</td>
<td>(0.270)</td>
<td>(0.199)</td>
<td>(0.207)</td>
<td>(0.207)</td>
<td>(0.207)</td>
<td>(0.207)</td>
<td>(0.207)</td>
</tr>
<tr>
<td>-0.6</td>
<td>-8.25</td>
<td>0.079</td>
<td>2.749</td>
<td>0.209</td>
<td>0.205</td>
<td>-0.6</td>
<td>-6.45</td>
<td>-0.101</td>
<td>2.134</td>
<td>0.0474</td>
<td>0.0864</td>
</tr>
<tr>
<td>(0.268)</td>
<td>(0.265)</td>
<td>(0.265)</td>
<td>(0.265)</td>
<td>(0.265)</td>
<td>(0.265)</td>
<td>(0.0198)</td>
<td>(0.207)</td>
<td>(0.207)</td>
<td>(0.207)</td>
<td>(0.207)</td>
<td>(0.207)</td>
</tr>
<tr>
<td>-0.4</td>
<td>-8.45</td>
<td>0.081</td>
<td>2.747</td>
<td>0.203</td>
<td>0.196</td>
<td>-0.4</td>
<td>-6.45</td>
<td>-0.099</td>
<td>2.122</td>
<td>0.0434</td>
<td>0.0874</td>
</tr>
<tr>
<td>(0.266)</td>
<td>(0.263)</td>
<td>(0.263)</td>
<td>(0.263)</td>
<td>(0.263)</td>
<td>(0.263)</td>
<td>(0.0196)</td>
<td>(0.209)</td>
<td>(0.209)</td>
<td>(0.209)</td>
<td>(0.209)</td>
<td>(0.209)</td>
</tr>
<tr>
<td>-0.2</td>
<td>-8.65</td>
<td>0.083</td>
<td>2.745</td>
<td>0.202</td>
<td>0.192</td>
<td>-0.2</td>
<td>-6.45</td>
<td>-0.097</td>
<td>2.111</td>
<td>0.0382</td>
<td>0.0892</td>
</tr>
<tr>
<td>(0.266)</td>
<td>(0.262)</td>
<td>(0.262)</td>
<td>(0.262)</td>
<td>(0.262)</td>
<td>(0.262)</td>
<td>(0.193)</td>
<td>(0.210)</td>
<td>(0.210)</td>
<td>(0.210)</td>
<td>(0.210)</td>
<td>(0.210)</td>
</tr>
<tr>
<td>0</td>
<td>-8.85</td>
<td>0.085</td>
<td>2.742</td>
<td>0.204</td>
<td>0.192</td>
<td>0</td>
<td>-6.4</td>
<td>-0.096</td>
<td>2.101</td>
<td>0.0305</td>
<td>0.0915</td>
</tr>
<tr>
<td>(0.267)</td>
<td>(0.262)</td>
<td>(0.262)</td>
<td>(0.262)</td>
<td>(0.262)</td>
<td>(0.262)</td>
<td>(0.189)</td>
<td>(0.211)</td>
<td>(0.211)</td>
<td>(0.211)</td>
<td>(0.211)</td>
<td>(0.211)</td>
</tr>
<tr>
<td>0.2</td>
<td>-9.15</td>
<td>0.088</td>
<td>2.739</td>
<td>0.209</td>
<td>0.196</td>
<td>0.2</td>
<td>-6.35</td>
<td>-0.099</td>
<td>2.101</td>
<td>0</td>
<td>0.0915</td>
</tr>
<tr>
<td>(0.269)</td>
<td>(0.264)</td>
<td>(0.264)</td>
<td>(0.264)</td>
<td>(0.264)</td>
<td>(0.264)</td>
<td>(0.171)</td>
<td>(0.211)</td>
<td>(0.211)</td>
<td>(0.211)</td>
<td>(0.211)</td>
<td>(0.211)</td>
</tr>
<tr>
<td>0.4</td>
<td>-9.5</td>
<td>0.093</td>
<td>2.735</td>
<td>0.215</td>
<td>0.203</td>
<td>0.4</td>
<td>-6.30</td>
<td>-0.099</td>
<td>2.101</td>
<td>0</td>
<td>0.0915</td>
</tr>
<tr>
<td>(0.271)</td>
<td>(0.266)</td>
<td>(0.266)</td>
<td>(0.266)</td>
<td>(0.266)</td>
<td>(0.266)</td>
<td>(0.171)</td>
<td>(0.211)</td>
<td>(0.211)</td>
<td>(0.211)</td>
<td>(0.211)</td>
<td>(0.211)</td>
</tr>
<tr>
<td>0.6</td>
<td>-9.9</td>
<td>0.100</td>
<td>2.731</td>
<td>0.221</td>
<td>0.209</td>
<td>0.6</td>
<td>-6.25</td>
<td>-0.099</td>
<td>2.101</td>
<td>0</td>
<td>0.0915</td>
</tr>
<tr>
<td>(0.270)</td>
<td>(0.267)</td>
<td>(0.267)</td>
<td>(0.267)</td>
<td>(0.267)</td>
<td>(0.267)</td>
<td>(0.171)</td>
<td>(0.211)</td>
<td>(0.211)</td>
<td>(0.211)</td>
<td>(0.211)</td>
<td>(0.211)</td>
</tr>
<tr>
<td>0.8</td>
<td>-10.05</td>
<td>0.140</td>
<td>2.742</td>
<td>0</td>
<td>0.192</td>
<td>0.8</td>
<td>-6.20</td>
<td>-0.099</td>
<td>2.101</td>
<td>0</td>
<td>0.0915</td>
</tr>
<tr>
<td>(0.171)</td>
<td>(0.262)</td>
<td>(0.262)</td>
<td>(0.262)</td>
<td>(0.262)</td>
<td>(0.262)</td>
<td>(0.171)</td>
<td>(0.211)</td>
<td>(0.211)</td>
<td>(0.211)</td>
<td>(0.211)</td>
<td>(0.211)</td>
</tr>
<tr>
<td>1.0</td>
<td>-10.05</td>
<td>0.140</td>
<td>2.742</td>
<td>0</td>
<td>0.192</td>
<td>1.0</td>
<td>-6.15</td>
<td>-0.099</td>
<td>2.101</td>
<td>0</td>
<td>0.0915</td>
</tr>
<tr>
<td>(0.171)</td>
<td>(0.262)</td>
<td>(0.262)</td>
<td>(0.262)</td>
<td>(0.262)</td>
<td>(0.262)</td>
<td>(0.171)</td>
<td>(0.211)</td>
<td>(0.211)</td>
<td>(0.211)</td>
<td>(0.211)</td>
<td>(0.211)</td>
</tr>
</tbody>
</table>

The simulation results in Sections 3.2 to 3.4 clearly show that on average $\hat{\tau}_1^2$ and $\hat{\tau}_2^2$ are upwardly biased when $\hat{\rho}_B$ equals 1 or $-1$, and, given the above discussion, this is likely to
be primarily due to the proposed Reason (I). The upward bias in $\hat{\tau}_j^2$ when $\hat{\rho}_B$ is equal to 1 is indicated in Figure 5.5, which shows two profile likelihoods for $\hat{\tau}_j^2$ that I commonly observed when $\hat{\rho}_B$ was 1. Both these profile likelihoods are characterised by relatively rapid increases in the value of $\hat{\tau}_j^2$ as $\hat{\rho}_B$ becomes exceedingly close to the converged value of 1. Similar rapid increases in $\hat{\tau}_j^2$ were also observed for those occasions where $\hat{\rho}_B$ was equal to −1. This rapid inflation in $\hat{\tau}_j^2$ is one possible explanation why $\hat{\tau}_j^2$ is upwardly biased on average when $\hat{\rho}_B$ is equal to either 1 or −1. Indeed, as the between-study covariance is defined by $\tau_{12} = \rho_B \tau_1 \tau_2$, when Reason (I) leads to $\hat{\rho}_B$ being constrained to 1 (below the value the RIGLS procedure desires) or −1 (above the value the RIGLS procedure desires), it is plausible that the values of $\hat{\tau}_1^2$ and $\hat{\tau}_2^2$ will be forced higher to compensate and achieve an increase in the size of $\hat{\tau}_{12}$.

Figure 5.5: Two common profile likelihoods for $\hat{\tau}_j^2$ where $\hat{\rho}_B$ is equal to 1

(i) $\hat{\tau}_1^2$ is gradually increasing, before more rapidly increasing as $\hat{\rho}_B$ gets close to 1

![Graph](image1)

(ii) $\hat{\tau}_1^2$ is gradually decreasing before more rapidly increasing as $\hat{\rho}_B$ gets close to 1

![Graph](image2)
In summary, the important message is clearly that there is an upward bias in $\hat{\tau}_j^2$ when $\rho_b$ is estimated as 1 or -1 on average. I have put forward two potential explanations for why this problem occurs, of which the proposed Reason (I) is likely to be the most influential. I am aware that these two proposed reasons should ideally be backed up by formal mathematical arguments, and thus they should only be treated as potential explanations. However, the multi-dimensional nature of the problem makes analytic consideration quite complex, as there are five unknown parameters and each parameter estimate has some impact on the other four parameters. Although this issue would make interesting further research, a more pressing concern is to translate the problems, especially that of biased $\hat{\tau}_j^2$s, back into the meta-analysis context of deciding when to use Model A to form evidence-based results and conclusions, and to ultimately ascertain if and how Model A can be used for the evidence synthesis of prognostic marker studies.

5.6.4 The dangers of making evidence-based conclusions from Model A when $\rho_b$ is estimated as 1 or -1

For one specific dataset, an upward bias in the $\hat{\tau}_j^2$s could lead to inappropriate (i.e. biased) pooled estimates and therefore misleading evidence-based recommendations for clinical practice. For instance, an inflation in the $\hat{\tau}_j^2$s will cause those studies with the largest $s_{ij}^2$s to have more weighting in the estimate of the pooled value than they would otherwise have done. This is because as the $\hat{\tau}_j^2$s increase the individual study weights become more similar across studies (see weights described in Section 4.4), and so this change in weighting may cause biased, potentially misleading pooled estimates to be made. This is obviously disconcerting and means that whenever Model A estimates $\rho_b$ as 1 or -1 researchers should be extremely cautious about using the pooled estimates to make evidence-based conclusions or recommendations. The danger of inappropriate conclusions
is potentially greater when the proposed Reason (I), rather than Reason (II), is causing a
\( \hat{\rho}_B \) of 1 or -1, as Reason (I) has a greater potential impact on the parameter estimates and
thus a greater scope for introducing bias (Section 5.6.3). Furthermore, whenever there is
bias in \( \hat{\tau}_j^2 \) it is likely to be most influential on the pooled estimates when the \( \hat{\tau}_j^2 \) s are
similar in size to the \( s^2_{\hat{\tau}_j} \) s; this is because:

- When both \( \hat{\tau}_j^2 \) s are much larger than the \( s^2_{\hat{\tau}_j} \) s, a \( \hat{\rho}_B \) equal to either 1 or -1 introduces
  only very small bias in the \( \hat{\tau}_j^2 \) s; the simulation results show this in Section 5.4.4.
- When both \( \hat{\tau}_j^2 \) s are much smaller than the \( s^2_{\hat{\tau}_j} \) s, a \( \hat{\rho}_B \) equal to either 1 or -1 introduces
  a relatively large bias in the \( \hat{\tau}_j^2 \) s. However, if the inflated \( \hat{\tau}_j^2 \) s are still much smaller
  than the \( s^2_{\hat{\tau}_j} \) s, the impact of both \( \hat{\rho}_B \) and the biased \( \hat{\tau}_j^2 \) s will be very small because the
  \( s^2_{\hat{\tau}_j} \) s and \( \rho_w \) dominate the estimation procedure. The simulations results show this in
  Chapter 5.4.2.
- When the \( \hat{\tau}_j^2 \) s are similar in size to the \( s^2_{\hat{\tau}_j} \) s, a \( \hat{\rho}_B \) equal to either 1 or -1 will cause a
  relatively large bias in the \( \hat{\tau}_j^2 \) s and, given they are a similar size to the \( s^2_{\hat{\tau}_j} \) s, the biased
  \( \hat{\tau}_j^2 \) s can also have a large influence during the estimation of \( \beta_j \). Hence, this is the
  situation of most concern for making biased or misleading evidence-based conclusions.

In order to help clarify why there are dangers of using Model A when \( \hat{\rho}_B \) is equal to either
1 or -1, I have produced a summary box to highlight the key problems when estimating
parameters in Model A (Figure 5.6). Furthermore, I have also summarised the potential
impact on the other parameter estimates when \( \rho_B \) is estimated as 1 or -1 (Figure 5.7).

These summary boxes go some way to understanding the concerns raised by Thompson et
al. regarding the estimation of parameters in Model A [169], which was one of my key
objectives specified in Section 3.7. The dangers of using Model A when $\rho_B$ is estimated as 1 or $-1$ will be further illustrated in Sections 7.5.3, 8.3.5 and 8.3.6.

Figure 5.6: Summary of the key problems for estimating the parameters in Model A, with particular consideration of the between-study correlation ($\rho_B$).

Convergence Issues for Model A

- Model A will most readily converge when $S^2_{n1}$ and $S^2_{n2}$ are small in comparison to $\tau^2_1$ and $\tau^2_2$. In this situation $\hat{\tau}^2_1$, $\hat{\tau}^2_2$ and $\hat{\rho}_B$ are well-defined, i.e. the profile likelihood has a clear peaked maximum point for these parameters (Figure 5.2(ii) illustrates this).

- Model A will often fail to converge when $S^2_{n1}$ and $S^2_{n2}$ are large in comparison to $\tau^2_1$ and $\tau^2_2$. Even when convergence is achieved in this situation, $\hat{\tau}^2_1$, $\hat{\tau}^2_2$ and in particular $\hat{\rho}_B$ are poorly defined, with flat likelihoods common (Figure 5.2(i) illustrates this).

- The larger the number of studies, the better defined $\hat{\tau}^2_1$, $\hat{\tau}^2_2$ and $\hat{\rho}_B$ will be. However, where $S^2_{n1}$ and $S^2_{n2}$ are large in comparison to $\tau^2_1$ and $\tau^2_2$ an unrealistic number of studies may be needed to obtain well-defined $\hat{\tau}^2_1$, $\hat{\tau}^2_2$ and $\hat{\rho}_B$ (see Section 5.6.2).

The Problem of Estimating the Between-Study Correlation ($\rho_B$) in Model A

- $\hat{\rho}_B$ is not inherently defined between $-1$ and 1 by Model A, and it is only the Cholesky parameterisation of the parameters that prevents it from going beyond this range (see Section 4.2).

- $\hat{\rho}_B$ is commonly equal to either 1 or $-1$ in Model A, increasingly more so the larger the $S^2_{n1}$s and the $S^2_{n2}$s are in comparison to $\tau^2_1$ and $\tau^2_2$.

- I have identified two potential reasons why $\rho_B$ can be estimated as 1 or $-1$:

  REASON (I) - the RIGLS procedure appears to seek a value for $\hat{\rho}_B$ above 1 or below $-1$, but the Cholesky parameterisation does not allow $\hat{\rho}_B$ to go beyond the range $-1$ to 1, and therefore $\hat{\rho}_B$ from Model A converges at either 1 or $-1$ respectively, i.e. the value at the end of the constrained range (see Section 5.6.1).

  REASON (II) - the between-study parameter estimates are poorly defined and a flat likelihood occurs for $\hat{\rho}_B$, which is gradually increasing up to the value of 1 or $-1$ and thus causes $\hat{\rho}_B$ to converge at one of these values (see Section 5.6.2).

- A plot of the profile likelihood for $\hat{\rho}_B$ may help ascertain whether the proposed Reason (I) or Reason (II) is the cause of $\hat{\rho}_B$ being equal to either 1 or $-1$. A flat profile likelihood is likely to indicate Reason (II), whereas it will be steeper and more curved for Reason (I) (see Section 5.6.3).
Figure 5.7: The impact of $\rho_B$ being estimated as 1 or −1 on the other parameter estimates in Model A

<table>
<thead>
<tr>
<th>The impact of $\rho_B$ being estimated as 1 or −1 in Model A</th>
</tr>
</thead>
<tbody>
<tr>
<td>• When $\rho_B$ is estimated as 1 or −1 there will on average be an upward bias in $\tau_1^2$ and $\tau_2^2$.</td>
</tr>
<tr>
<td>• Importantly, bias in $\tau_j^2$ could lead to misleading meta-analysis results because it changes the weighting of each study toward the pooled estimates; for example, an upwardly biased $\tau_j^2$ would allow smaller studies to have more weighting toward the values of the pooled estimates than they ought, and thus misleading pooled estimates may be obtained.</td>
</tr>
<tr>
<td>• Whenever Model A gives a $\hat{\rho}_B$ equal to 1 or −1 researchers should therefore be extremely cautious about using the pooled estimates to form evidence-based conclusions or recommendations.</td>
</tr>
<tr>
<td>• The danger of biased, misleading conclusions is particularly high when the $\tau_j^2$ s are similar in size to the $s_j^2$ s, because in this situation the $\tilde{\tau}_j^2$ may both be highly affected by bias and highly influential toward the value of the pooled estimates.</td>
</tr>
</tbody>
</table>

5.7 Summary and rationale for subsequent chapters

The extensive simulations in Chapter 5 have highlighted some important issues regarding Model A that were not obvious during the analytical investigations in Chapter 4. In particular, for complete-case data there is negligible benefit of Model A over two independent URMAs in terms of estimating the $\beta_j$s themselves. Furthermore, Model A commonly estimates $\rho_B$ to be 1 or −1 and this leads to $\tilde{\tau}_1^2$ and $\tilde{\tau}_2^2$ being upwardly biased, which in turn can cause biased pooled estimates and therefore potentially misleading evidence-based conclusions. These issues may therefore limit the use of Model A in practice. However, on a more positive note, $\rho_B$ is not always estimated as 1 or −1 in Model A and there are no problems of bias on these occasions (for example, see the simulation results in Section 5.4.4). Furthermore, this chapter has only considered complete-case data and the $\hat{\beta}_j$s themselves; it is possible that the major benefits of Model
A will be seen when there is some missing data (see Section 4.8.5) and when \((\hat{\beta}_1 - \hat{\beta}_2)\) is of interest (see Section 4.8.6).

Whilst not ignoring the problem of unavailable within-study correlations (see Chapter 8), this chapter has clearly shown that even if Model A can be applied to OS and DFS HRs from prognostic marker studies, it will have little benefit over two independent URMAs when there is complete-case data for estimating the \(\beta_j\) s. An assessment of missing data is therefore clearly needed at this stage, particularly as this is a common problem for meta-analysis of prognostic marker studies (see Figure 3.2), and was the main motivation for initially investigating bivariate meta-analysis models. Chapter 6 will therefore critically assess the use of Model A in relation to missing data, and will also seek to extend all findings to when \((\hat{\beta}_1 - \hat{\beta}_2)\) is of interest. The understanding gained from Chapter 5 about the estimation problems of Model A will be increasingly important during the forthcoming assessments in Chapter 6, as I keep working toward ascertaining if and how BRMA models can be utilised in the evidence synthesis of prognostic marker studies.
Chapter 6

A CRITICAL ASSESSMENT OF MODEL A FOR MISSING DATA AND OTHER EXTENSIONS

Chapter overview

In this chapter I will assess the use of Model A beyond the findings in Chapter 5 for complete-case data. Firstly, I will assess the benefits of Model A over two independent univariate random-effects meta-analyses (URMAs) for estimating $\beta_j$ when there is missing data, and then I will consider extensions to the case when $(\hat{\beta}_1 - \hat{\beta}_2)$ is of interest. The chapter will then continue by briefly considering how all the findings extend to other situations, such as negative correlation, fixed- rather than random-effects analyses, and trivariate and other higher order multivariate models. The chapter concludes by extending Model A to a bivariate meta-regression.

6.1 Using Model A given missing summary statistics across studies

Model A is an example of a mixed model, which has the major benefit of being able to accommodate partially available response data [152]. For my purposes, this means that Model A can incorporate those studies where a summary statistic for only one of the two outcomes is known (e.g. study 1 can be included even if only one of $\bar{Y}_{11}$ and $\bar{Y}_{12}$ is available). This is not possible in a URMA, where a study will contribute no information if the summary statistic for the single outcome of interest is not available (e.g. study 1 would not be included in a URMA for outcome 1 if $\bar{Y}_{11}$ was not available). This is very important because those interested in multiple summary statistics from each study are highly unlikely to obtain all of them from every study, and this was particularly evident in the prognostic marker review in neuroblastoma (see Chapter 2). Thus, where a summary statistic is
missing for an outcome (e.g. DFS) there is the opportunity to use Model A to ‘borrow
strength’ from the correlated summary statistic that is available from a related outcome
(e.g. OS).

Consider the use of Model A when there are only two studies (i.e. \( n = 2 \)), with \( \hat{Y}_{i2} \)
available for both studies but \( \hat{Y}_{i1} \) only available for study 2. This is a hypothetical situation,
as one would not actually be able to estimate the between-study parameters in this case but
if I assume \( \tau_1^2, \tau_2^2, \) and \( \tau_{12} \) are known it serves as a simple illustrative example. In this
situation, the analytic solution for \( \hat{\beta}_j \) from Model A can be derived as in Section 4.1 using
\[
\hat{\beta} = \left( X^T V^{-1} X \right)^{-1} X^T V^{-1} Y,
\]
where now the rows and columns of \( X, V \) and \( Y \) relating to \( i = 1, j = 1 \) are omitted because they do not contribute anything toward the overall likelihood;
the following solution for \( \hat{\beta}_1 \) is obtained:
\[
\hat{\beta}_1 = \frac{\hat{Y}_{21} - (\hat{Y}_{22} - \hat{Y}_{12})(\tau_{12} + \lambda_2)}{2\tau_2^2 + s_{22}^2 + s_{12}^2}
\]  
(6.1)

This solution is equivalent to having complete-case data for both studies, but with \( \hat{Y}_{11} \)
having a \( s_{11}^2 = \infty \). In other words, equation (6.1) is the limit of the complete case \( \hat{\beta}_1 \)
answer from equation (4.3) for \( n = 2 \) as \( s_{11}^2 \rightarrow \infty \), i.e.
\[
\begin{align*}
\left[ \frac{Y_{11}}{(\tau_1^2 + s_{11}(\tau_1^2 + \lambda_2^2) - (\tau_1 + \lambda_2)^2)} + \frac{Y_{12}}{(\tau_1^2 + s_{12}(\tau_1^2 + \lambda_2^2) - (\tau_1 + \lambda_2)^2)} \right] \\
\left[ \frac{Y_{21}}{(\tau_2^2 + s_{21}(\tau_2^2 + \lambda_2^2) - (\tau_1 + \lambda_2)^2)} + \frac{Y_{22} - \hat{Y}_{12}}{(\tau_2^2 + s_{22}(\tau_2^2 + \lambda_2^2) - (\tau_1 + \lambda_2)^2)} \right] \\
\left[ \frac{\tau_{12} - \hat{Y}_{12}}{(\tau_1^2 + s_{12}(\tau_2^2 + \lambda_2^2) - (\tau_1 + \lambda_2)^2)} + \frac{\lambda_2(\tau_1^2 + s_{12}^2) - \lambda_1(\tau_2^2 + s_{22}^2)}{(\tau_1^2 + s_{12}^2)(\tau_2^2 + s_{22}^2) - (\tau_1 + \lambda_2)^2} \right]
\end{align*}
\]  
(6.2)

Knowing that missing \( j = 1 \) summary statistics essentially means their \( s_{11}^2 \)s are equivalent
to infinity is very important because Remark 4 in Section 4.8.4, together with Points (1)
and (2) in Section 4.5, showed that the larger the differences between the URMA weights (i.e. $s_{ji}^2 - s_{ki}^2$) the more Model A can ‘borrow strength’ from the $j = 2$ data. For those studies where $j = 1$ is missing their URMA weighting is zero as their $s_{ji}^2$ is infinity. Hence, the difference between $s_{ji}^2$ for an available summary statistic and $s_{ki}^2$ for an unavailable summary statistic is itself essentially infinity, and therefore there is a maximum opportunity to ‘borrow strength’ by using Model A when there is missing data (see also Remark 5 in Section 8.5).

Considering equation (6.1) further, if $(\tau_{12} + \lambda_2) = 0$ then $\hat{\beta}_1 = \tilde{Y}_{21}$, i.e. the URMA estimate. However if $(\tau_{12} + \lambda_2) \neq 0$ then $\hat{\beta}_1$ will incorporate the data from outcome $j = 2$ (i.e. $\hat{Y}_{12}$ and $\hat{Y}_{22}$). Furthermore, a larger $(\hat{Y}_{22} - \hat{Y}_{12})$, a smaller $s_{12}^2$ and $s_{22}^2$, and a larger $(\tau_{12} + \lambda_2)$ allows greater potential for ‘shrinkage’ of $\hat{\beta}_1$ away from $\hat{Y}_{21}$. This is sensible as when $\hat{Y}_{22}$ is highly likely to be very different than $\hat{Y}_{12}$, and given high within- and between-study correlation, the unknown $\hat{Y}_{11}$ is also highly likely to be very different than the known $\hat{Y}_{21}$. For example let $\hat{Y}_{11}$ be missing, $\hat{Y}_{12} = -1$, $\hat{Y}_{21} = 1$ and $\hat{Y}_{22} = 1$. Furthermore, let $s_{12}^2 = s_{21}^2 = s_{22}^2 = 1$, $\tau_2 = 1$, and $\tau_{12} = \lambda_2 = 0.5$ (i.e. the within-study correlation for study 2 and the between-study correlation are both 0.5). Both URMA and Model A give a pooled estimate of $\tilde{\beta}_2 = 0$. However, the URMA gives $\hat{\beta}_1 = 1$ whereas the solution from Model A (using equation (6.1)) is somewhat different, as follows:

$$\hat{\beta}_1 = \tilde{Y}_{21} - \frac{(\hat{Y}_{22} - \hat{Y}_{12})(\tau_{12} + \lambda_2)}{2\tau_2^2 + s_{22}^2 + s_{12}^2} = 1 - \frac{(1-(-1))(0.5+0.5)}{4} = 0.5$$

The difference between the URMA and Model A results is due to $\hat{\beta}_1$ from Model A ‘borrowing strength’ from the known and related information from outcome $j = 2$, and this shrinks it away from the URMA solution toward zero. Of course, I have assumed $\tau_1^2$, $\tau_2^2$, 

---

R.D. Riley, Ph.D. Thesis  Chapter 6 189
and $\tau_{ij}$ are known for simplicity here, but in reality they must also be estimated. Indeed, the between study parameter estimates can also ‘borrow strength’ in Model A in a similar way to the pooled estimates, with shrinkage away from the URMA solutions a possibility (see Chapter 6.4).

### 6.2 Types of missing data

When dealing with missing data it is important to be clear about the three basic ways in which it typically occurs [183]. The differences between the three types of missing data are subtle, and I will try to relate them to Model A by considering, without loss of generality, the situation where some data is missing for $j = i$:

**Type 1 - Data are missing completely at random (MCAR):** *MCAR arises when the missing variable does not depend on the value of the variable itself, and the missingness also does not depend on the values of other variables in the database.*

For the purposes of Model A this relates to where the missing $\tilde{Y}_{ij}$ is not missing due to either the value of $\tilde{Y}_{ij}$ itself or the related $\tilde{Y}_{ij}$ value. In this situation the known summary statistics across studies will be representative of the underlying population, and therefore the estimate of $\beta_i$ using only the known data will be unbiased on average from Model A and also from a URMA. Indeed, the use a URMA inherently assumes that any missing data is MCAR and thus that $\hat{\beta}_i$ from a URMA represents the truth on average. Furthermore, MCAR infers that the relationship between known $\tilde{Y}_{ij}$s and known $\tilde{Y}_{ij}$s will be the same as the relationship between (i) known $\tilde{Y}_{ij}$s and missing $\tilde{Y}_{ij}$s, and also (ii) known $\tilde{Y}_{ij}$s and missing $\tilde{Y}_{ij}$s, and so Model A is applicable for the situation of MCAR.

**Type 2 - Data are missing at random (MAR):** *MAR arises when the mechanism causing a variable to be missing depends on the known variables, and so the pattern of missing data is traceable or predictable from the values of the other variables in the database.*
For the purposes of Model A this relates to where the pattern of missing $\tilde{Y}_{ni}$s is predictable from the related and known $\tilde{Y}_{i2}$ values. For example, if $\tilde{Y}_{ni}$ and $\tilde{Y}_{i2}$ relate to two highly correlated outcomes (e.g. DFS and OS) some authors may decide to only publish the summary statistic for one of the outcomes (e.g. just present $\tilde{Y}_{i2}$ for OS), regardless of their values, and the missing outcome is then predictable from the known outcome. In this situation, if one uses the known summary statistics then $\hat{\beta}_i$ will be unbiased in Model A and also in a URMA. A more extreme example of MAR is when $\tilde{Y}_{ni}$ and $\tilde{Y}_{i2}$ are both reported if they are positive, but only $\tilde{Y}_{i2}$ is reported if they are both negative. In this situation, one can still trace the pattern of missing $\tilde{Y}_{ni}$s from the known $\tilde{Y}_{i2}$s and so the missing data mechanism is MAR. In terms of meta-analysis, if one uses a URMA in this situation then $\hat{\beta}_i$ is likely to be biased because the known $\tilde{Y}_{ni}$s are not representative of the overall evidence. However, if one uses Model A than the bias in $\hat{\beta}_i$ is likely to be smaller as it could potentially 'borrow strength' from all the related $\tilde{Y}_{i2}$s that are available (see shrinkage example in Section 6.1). Model A may therefore be most beneficial where data is MAR.

**Type 3 - Data are not missing at random (NMAR):** NMAR arises when the missing variable does depend on the value of the variable itself, and the pattern of missing data is not traceable or predictable from other variables in the database.

For the purposes of Model A this relates to when the missing $\tilde{Y}_{ni}$ cannot be predicted from the related $\tilde{Y}_{i2}$, and is missing due to the value of $\tilde{Y}_{ni}$ itself. Importantly, in this situation $\hat{\beta}_i$ will be biased in both a URMA and Model A on average, and Model A cannot be used to ‘borrow strength’ because the missingness is not related to the available $\tilde{Y}_{i2}$s and so these cannot be used to ‘borrow strength’ about the unavailable $\tilde{Y}_{ni}$s. Summary statistics
NMAR in an evidence-based review may often be caused by dissemination biases (see Sections 1.6 and 3.2, and also Chapter 9 later) [42], such as publication bias and biased within-study selective reporting [41;64].

Given these three definitions, it is very important to emphasise that Model A is only applicable when the missing summary statistic is MCAR or MAR in those studies only providing one of the two outcomes [152]. Given data is MCAR or MAR in these studies, then Model A can ‘borrow strength’ about the missing outcome’s summary statistic using the known outcome’s summary statistic. Indeed, previous applications of Model A in the meta-analysis literature have all assumed MAR where some studies have one outcome missing [147;169]. Of course, there may also be some studies where the summary statistics for neither outcome were available (e.g. see the discussion about the MYCN dataset in Section 3.2) but such studies are not taken into account by Model A. It is also possible that MCAR or MAR is the true mechanism in those studies only providing one outcome, whilst NMAR is the true mechanism in those studies providing neither outcome. Methods for assessing missing data and dissemination bias after fitting Model A will be considered in Chapter 9.

Given this discussion, one therefore needs to clearly justify the assumption of MCAR or MAR in those studies providing only outcome before implementing Model A in practice, and if such missing data is considered NMAR then Model A is not suitable for evidence synthesis. To demonstrate this, consider the hypothetical datasets ‘M1’ and ‘M2’ in Table 6.1, which shows the true and reported values of summary statistics from 11 studies for two outcomes, say DFS and OS for simplicity. Also assume the within-study correlation is known to be 0.9 in each study. In dataset M1, the DFS summary statistics for studies 7 to 11 were deliberately omitted because the summary statistics had a negative sign and so only the OS results were reported. Furthermore, the relationship between OS and DFS
summary statistics in those studies where both were reported is entirely different to the relationship in those studies where only OS was reported (see dataset M2). Hence, for dataset M1 the missing data is NMAR (i.e. $\tilde{Y}_{i1}$ is not predictable from the known $\tilde{Y}_{i2}$s) and so Model A is inappropriate for evidence synthesis in this situation. If one did wrongly use Model A for dataset M1 then the results obtained are even more misleading answers than those from the URMA (Table 6.2). For example, it further overestimates the DFS pooled estimate, it overestimates the between-study correlation (worryingly leading to $\hat{\rho}_B$ equal to 1), and it further underestimates the between study variance. Even though this is perhaps an extreme example, it suffices to illustrate the dangers of wrongly using Model A when data is NMAR, and not MCAR or MAR.

Table 6.1: A hypothetical dataset of 11 studies to show the problem when missing data is NMAR. Shown are the reported summary statistics known to the meta-analysts (M1) alongside the true values of those unreported summary statistics unknown to the meta-analyst (M2).

<table>
<thead>
<tr>
<th>Study</th>
<th>DFS $\tilde{Y}_{i1}$</th>
<th>DFS $s_{i1}^2$</th>
<th>OS $\tilde{Y}_{i2}$</th>
<th>OS $s_{i2}^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>0.5</td>
<td>5</td>
<td>0.5</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>0.5</td>
<td>4</td>
<td>0.5</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0.5</td>
<td>3</td>
<td>0.5</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>0.5</td>
<td>2</td>
<td>0.5</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.5</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0.5</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>7</td>
<td>NA</td>
<td>NA</td>
<td>5</td>
<td>0.5</td>
</tr>
<tr>
<td>8</td>
<td>NA</td>
<td>NA</td>
<td>4</td>
<td>0.5</td>
</tr>
<tr>
<td>9</td>
<td>NA</td>
<td>NA</td>
<td>3</td>
<td>0.5</td>
</tr>
<tr>
<td>10</td>
<td>NA</td>
<td>NA</td>
<td>2</td>
<td>0.5</td>
</tr>
<tr>
<td>11</td>
<td>NA</td>
<td>NA</td>
<td>1</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Table 6.2: Univariate (URMA) and bivariate (Model A) random-effects meta-analysis results for dataset M1

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>$\tilde{\beta}_1$ (s.e.)</th>
<th>$\tau_1^2$</th>
<th>$\tilde{\beta}_2$ (s.e.)</th>
<th>$\tau_2^2$</th>
<th>$\hat{\rho}_B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRUE</td>
<td>URMA</td>
<td>0 (1)</td>
<td>10.50</td>
<td>2.727 (0.506)</td>
<td>2.32</td>
<td>-</td>
</tr>
<tr>
<td>TRUE</td>
<td>Model A</td>
<td>0 (1)</td>
<td>10.50</td>
<td>2.727 (0.506)</td>
<td>2.32</td>
<td>0.314</td>
</tr>
<tr>
<td>REPORTED</td>
<td>URMA</td>
<td>2.500 (0.760)</td>
<td>3.00</td>
<td>2.727 (0.506)</td>
<td>2.32</td>
<td>-</td>
</tr>
<tr>
<td>REPORTED</td>
<td>Model A</td>
<td>2.711 (0.554)</td>
<td>2.47</td>
<td>2.727 (0.506)</td>
<td>2.32</td>
<td>1.0</td>
</tr>
</tbody>
</table>
6.3 Using Model A when the type of missing data is unknown

6.3.1 Are the missing OS and DFS summary statistics missing at random?

I think it is appropriate at this stage to consider whether, in those studies only providing one outcome, the unavailable OS and DFS HR estimates from the neuroblastoma review can be considered MCAR or MAR. There are a number of potential reasons why these missing OS and DFS summary statistics could fall into the MCAR or MAR categories:

- Some authors believe that OS is misleading, especially when there are other possible causes of death, and so do not report OS.
- Some authors may not report DFS because of the subjectivity in determining recurrence (and time of recurrence).
- OS and DFS summary statistics can often be very similar, so including both outcome results can often be deemed unnecessary. For example, for neuroblastoma patients the close relationship between OS and DFS estimates is a reflection of the extremely poor outcome for those patients experiencing a recurrence of neuroblastoma, and so authors may opt to only present one outcome [22].
- Lack of resources may mean that OS and DFS information was often not recorded, and thus the omission of these results may not be deliberate. For example, DFS information may not be recorded in the IPD and so authors may not have been able to estimate the DFS HR.
- The poor reporting of prognostic marker studies may simply be due to a consequence of poor statistical practice rather than a deliberate attempt to miss out some summary statistics or outcomes.

However, despite these possible reasons, there is still a great potential for the missing HRs from the neuroblastoma review to be NMAR; for example, within-study selective reporting may be causing one of the OS or DFS HRs to be unavailable in some studies (this will be
considered explicitly in Chapter 9). Of most concern is that the relationship between OS and DFS HR estimates in those studies where both outcomes are available is actually different than in those studies where only one outcome is available. This threat cannot be ruled out, even though MCAR or MAR may be valid for many studies due to the reasons above. Hence, to assume that the missing summary statistics are only MCAR or MAR is a very strong and potentially inappropriate assumption to make for OS and DFS HRs from prognostic marker studies.

6.3.2 Using Model A as a sensitivity analysis

Given Section 6.3.1, I believe that the use of Model A for the OS and DFS HRs from the neuroblastoma review may be more suitable as a sensitivity analysis in order to assess the robustness of the URMA results when one assumes the data is either MCAR or MAR in those studies only providing one outcome. Of course, the URMA results themselves also make the assumption that all missing data is MCAR (Section 6.2). When using Model A as a sensitivity analysis one must not consider the results from Model A in isolation as the additional assumption of MAR in those studies only providing one outcome may not be valid; thus, the Model A results should only be used to assess if and how the URMA conclusions would change if MAR (or MCAR) was plausible in these studies. The application of Model A to the prognostic marker datasets in this context is considered in Chapter 8, where the additional problem of unknown within-study correlations will also be considered. Of course, there is an additional issue for which a sensitivity analysis is required; that is, neither a URMA nor Model A would be valid if some missing data were NMAR in those studies for which neither outcome were available. Hence, in Chapter 9 I will also consider alternative methods for sensitivity analysis of the prognostic marker URMA results when all the missing summary statistics are assumed NMAR, and similarly for sensitivity analysis of the Model A results when MAR is assumed in those studies providing one outcome but NMAR is assumed in those studies providing neither outcome.
6.4 A simulation study to assess Model A for missing data

To assess the benefits of Model A over two independent URMAs when data truly is MCAR or MAR, I have performed a simulation study similar to that for complete-case data in Chapter 5 except that now I also created missing data. In all the following simulations Model A was known to be the correct model to use as complete-case data was generated directly from Model A before the ‘missing’ data for just one outcome was removed. Hence, the data was genuinely known to be MCAR or MAR in the simulations that now follow. Note though that again this section will only assess \( \tilde{\beta}_j \) and not \( \tilde{\beta}_1 - \tilde{\beta}_2 \), which will be considered separately in Section 6.6.

6.4.1 Data is missing completely at random for outcome \( j = 2 \)

The MCAR simulations are exactly as described in Chapter 5.2.1, except that there is only complete-case data for outcome \( j = 1 \) and now half the studies have a missing summary statistic for outcome \( j = 2 \). The studies with \( \tilde{Y}_{i2} \) missing were randomly selected, and the simulations were performed for \( n = 50 \) and \( n = 10 \) for each of the following four settings:

Setting (i): \( \rho_w = 0.8, \rho_b = 0.8 \), and the \( \tau_j^2 \)'s large in relation to the \( s_{ij}^2 \)'s

Setting (ii): \( \rho_w = 0.8, \rho_b = 0.8 \), and the \( \tau_j^2 \)'s similar in size to the \( s_{ij}^2 \)'s

Setting (iii): \( \rho_w = 0, \rho_b = 0.8 \), and the \( \tau_j^2 \)'s large in relation to the \( s_{ij}^2 \)'s

Setting (iv): \( \rho_w = 0, \rho_b = 0.8 \), and the \( \tau_j^2 \)'s similar in size to the \( s_{ij}^2 \)'s

I emphasise here that these simulations again assume the within-study correlations are known. Also, simulations for \( \tau_j^2 \) smaller than the \( s_{ij}^2 \)'s were not considered because in this situation the complete-case simulations had struggled to converge and the between-study parameters actually have very little impact because the overall variation is dominated by the \( s_{ij}^2 \)'s (see Section 5.4.2).
Averaging across all the simulation results, unbiased estimates of $\beta_1$ and $\beta_2$ were obtained for the URMA and for Model A in all of Settings (i) to (iv) (Table 6.3). Furthermore, for both $n = 50$ and $n = 10$, and again for all settings assessed, there was negligible benefit of Model A over a URMA for $\hat{\beta}_1$ as there was complete-case data for outcome $j = 1$ (this concurs with the conclusions from the complete-case simulations in Chapter 5). However, for outcome $j = 2$, the average precision of $\hat{\beta}_2$ noticeably increased in Model A, and this was despite an inflated $\hat{\tau}_1^2$ and particularly $\hat{\tau}_2^2$ on average, which was again caused by $\rho_B$ often being equal to either 1 or -1 (see Section 5.6.3). Similarly the average MSE of $\hat{\beta}_2$ was noticeably smaller in Model A than in the URMA. These results emphasise that Model A is beneficial (i.e. it gives a larger precision and smaller MSE on average) over a URMA for estimating $\beta_j$ when there is data MCAR and $\rho_B$ is not estimated as 1 or -1.

The gain in precision and reduction in MSE from Model A was larger for those simulations where $\rho_w$ and $\rho_B$ were both high, though even when $\rho_w$ was zero there was a reasonable improvement in precision and MSE where $\rho_B$ was still high. This was particularly true for the simulations where $\tau_1^2$ and $\tau_2^2$ were much larger than the $s_y^2$s, as the benefits (e.g. gain in precision) in the $\rho_w = 0$ results (Settings (iii) and (iv)) were very similar to those in the $\rho_w = 0.8$ results (Settings (i) and (ii), see Table 6.3(ii)). This is due to the relatively large $\hat{\tau}_j^2$s dominating the estimation.
Table 6.3: Simulation results from the univariate (URMA) and bivariate (Model A) random-effects meta-analysis models for 50 and 10 studies, where outcome \( j = 1 \) was available from all studies but outcome \( j = 2 \) was **missing completely at random (MCAR)** from half the studies in each case. The true parameter values to compare the results to are \( \beta_1 = 0, \beta_2 = 2, \tau_1^2 = 0.25, \tau_2^2 = 0.25 \). In Model A the within-study correlation (\( \rho_w \)) was known and the same for each study, whilst the between-study correlation (\( \rho_B \)) was estimated. The values of \( \rho_w \) and \( \rho_B \) in each simulation are shown in the table. Two settings were investigated for different sizes of the \( \tau_j^2 \) in relation to the \( \tau_j^2 \).

(i) **On average the \( s_j^2 \)s are similar in size to the \( \tau_j^2 \)**

| No. of studies (n) | Meta-analysis model | \( \rho_w \) | \( \rho_B \) | Converged out of 1000 | Mean of \( \hat{\beta}_1 \) (s.e. of mean) | Mean of \( s.e. \) of \( \hat{\beta}_1 \) | MSE of \( \hat{\beta}_1 \) (s.e.) | No. of 95% CIs for \( \hat{\beta}_1 \) including \( \beta_1 \) (%) | Mean of \( \hat{\beta}_2 \) (s.e.) | Mean of \( s.e. \) of \( \hat{\beta}_2 \) | MSE of \( \hat{\beta}_2 \) (s.e.) | No. of 95% CIs for \( \hat{\beta}_2 \) including \( \beta_2 \) (%) | Mean of \( \tau_1^2 \) (no. of \( \tau_1^2 = 0 \)) | Mean of \( \tau_2^2 \) (no. of \( \tau_2^2 = 0 \)) | Mean of \( \rho_B \) (no. of \( \rho_B = 1 \)) (%) | No. of \( \rho_B \) (no. of \( \rho_B = 1 \)) (%) |
|----------------------|----------------------|---------------|---------------|-----------------------|---------------------------------|----------------------------|------------------------|---------------------------------|---------------------------------|----------------------------|------------------------|---------------------------------|---------------------------------|---------------------------------|------------------------|---------------------------------|---------------------------------|
| 50 URMA 0.8 0.8 1000 | -0.0044 (0.147) | 0.102 | 0.0102 | 946 (94.6%) | 1.995 | 0.146 | 0.0216 | 949 (94.9%) | 0.253 | 0.248 | - | - | - | - |
| 50 Model A 0.8 0.8 1000 | -0.0045 (0.196) | 0.0987 | 0.00096 | 946 (94.6%) | 1.993 | 0.119 | 0.0150 | 946 (94.6%) | 0.249 | 0.252 | 0.794 | 2 (0.2%) | 95 (9.6%) | - | - | - |
| 50 URMA 0 0.8 1000 | -0.0045 (0.101) | 0.102 | 0.0102 | 956 (95.6%) | 1.999 | 0.145 | 0.0228 | 942 (94.2%) | 0.250 | 0.247 | - | - | - | - |
| 50 Model A 0 0.8 1000 | -0.0041 (0.143) | 0.101 | 0.0104 | 958 (95.8%) | 1.997 | 0.137 | 0.0203 | 947 (94.7%) | 0.251 | 0.253 | 0.788 | 1 (0.1%) | 356 (35.5%) | - | - | - |
| 10 URMA 0.8 0.8 1000 | -0.0072 (0.224) | 0.217 | 0.0502 | 935 (93.5%) | 1.984 | 0.262 | 0.0921 | 932 (93.2%) | 0.250 | 0.251 | - | - | - | - |
| 10 Model A 0.8 0.8 999 | -0.0083 (0.223) | 0.212 | 0.0498 | 933 (93.4%) | 1.986 | 0.225 | 0.0708 | 932 (93.3%) | 0.255 | 0.251 | 0.672 | 105 (10.5%) | 540 (54.1%) | - | - | - |
| 10 URMA 0 0.8 1000 | -0.0007 (0.213) | 0.218 | 0.0451 | 939 (93.9%) | 1.995 | 0.263 | 0.0825 | 942 (94.2%) | 0.256 | 0.255 | - | - | - | - |
| 10 Model A 0 0.8 997 | -0.0010 (0.213) | 0.222 | 0.0451 | 962 (96.5%) | 1.993 | 0.255 | 0.0797 | 963 (95.6%) | 0.270 | 0.275 | 0.636 | 101 (10.1%) | 600 (60.2%) | - | - | - |
(ii) On average the $s_f^2$ s are much smaller in size to the $\tau_j^2$.

<table>
<thead>
<tr>
<th>No. of studies (n)</th>
<th>Meta-analysis model</th>
<th>$\rho_W$</th>
<th>$\rho_B$</th>
<th>Converged out of 1000</th>
<th>Mean of $\hat{\beta}_1$ (s.e. of mean)</th>
<th>Mean s.e. of $\hat{\beta}_1$</th>
<th>MSE of $\hat{\beta}_1$</th>
<th>No. of 95% CIs for $\hat{\beta}_1$ including $\beta_0$ (%)</th>
<th>Mean of $\hat{\beta}_2$ (s.e.)</th>
<th>Mean s.e. of $\hat{\beta}_2$</th>
<th>MSE of $\hat{\beta}_2$</th>
<th>No. of 95% CIs for $\hat{\beta}_2$ including $\beta_2$ (%)</th>
<th>Mean of $\tau_1^2$ (no. of $\tau_1^2 = 0$)</th>
<th>Mean of $\tau_2^2$ (no. of $\tau_2^2 = 0$)</th>
<th>Mean of $\hat{\rho}_B$</th>
<th>No. of $\hat{\rho}_B = -1$ (%)</th>
<th>No. of $\hat{\rho}_B = 1$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>URMA</td>
<td>0.8</td>
<td>0.8</td>
<td>1000</td>
<td>-0.0027 (0.0689)</td>
<td>0.0708</td>
<td>0.0047</td>
<td>953 (95.3%)</td>
<td>1.999 (0.0974)</td>
<td>0.0996</td>
<td>0.0095</td>
<td>965 (96.5%)</td>
<td>0.244</td>
<td>0.246</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>Model A</td>
<td>0.8</td>
<td>0.8</td>
<td>1000</td>
<td>-0.0040 (0.0706)</td>
<td>0.0708</td>
<td>0.0050</td>
<td>953 (95.4%)</td>
<td>2.000 (0.0822)</td>
<td>0.0824</td>
<td>0.0067</td>
<td>953 (95.8%)</td>
<td>0.244</td>
<td>0.246</td>
<td>0.798</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>50</td>
<td>URMA</td>
<td>0</td>
<td>0.8</td>
<td>1000</td>
<td>-0.0040 (0.0705)</td>
<td>0.0708</td>
<td>0.0050</td>
<td>949 (94.9%)</td>
<td>2.000 (0.101)</td>
<td>0.0993</td>
<td>0.0101</td>
<td>950 (95.0%)</td>
<td>0.244</td>
<td>0.246</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>Model A</td>
<td>0</td>
<td>0.8</td>
<td>1000</td>
<td>-0.0040 (0.0705)</td>
<td>0.0708</td>
<td>0.0050</td>
<td>949 (94.9%)</td>
<td>2.000 (0.0825)</td>
<td>0.0823</td>
<td>0.0068</td>
<td>952 (95.2%)</td>
<td>0.244</td>
<td>0.246</td>
<td>0.799</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>URMA</td>
<td>0.8</td>
<td>0.8</td>
<td>1000</td>
<td>0.00029 (0.155)</td>
<td>0.155</td>
<td>0.0239</td>
<td>951 (95.1%)</td>
<td>2.003 (0.226)</td>
<td>0.215</td>
<td>0.0509</td>
<td>952 (95.2%)</td>
<td>0.248</td>
<td>0.258</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Model A</td>
<td>0.8</td>
<td>0.8</td>
<td>1000</td>
<td>0.0003 (0.155)</td>
<td>0.155</td>
<td>0.0239</td>
<td>952 (95.2%)</td>
<td>2.008 (0.194)</td>
<td>0.176</td>
<td>0.0375</td>
<td>951 (95.1%)</td>
<td>0.248</td>
<td>0.263</td>
<td>0.763</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>URMA</td>
<td>0</td>
<td>0.8</td>
<td>1000</td>
<td>-0.0022 (0.167)</td>
<td>0.154</td>
<td>0.0279</td>
<td>941 (94.1%)</td>
<td>1.997 (0.240)</td>
<td>0.209</td>
<td>0.0576</td>
<td>937 (93.7%)</td>
<td>0.244</td>
<td>0.246</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Model A</td>
<td>0</td>
<td>0.8</td>
<td>1000</td>
<td>-0.0022 (0.167)</td>
<td>0.154</td>
<td>0.0279</td>
<td>941 (94.1%)</td>
<td>1.999 (0.207)</td>
<td>0.174</td>
<td>0.0427</td>
<td>933 (93.3%)</td>
<td>0.244</td>
<td>0.256</td>
<td>0.760</td>
<td>39</td>
<td></td>
</tr>
</tbody>
</table>

N.B. $\hat{\rho}_B$ was not applicable when one or both of $\tau_1^2$ and $\tau_2^2$ was zero. MSE = mean-square-error, CIs = confidence intervals, s.e. = standard error. The 95% CI for $\hat{\beta}_j$ was calculated using a t-distribution with $(n_j - 1)$ degrees of freedom, where $n_j$ is the number of studies for which outcome $j$ was available. So for example, for $n = 50$ the degrees of freedom for $\hat{\beta}_1$ was 49, whereas the degrees of freedom for $\hat{\beta}_2$ was 24.
6.4.2 Data is missing at random for outcome \( j = 1 \)

For the MAR simulations I again considered simulations for \( n = 50 \) and \( n = 10 \) studies for each of the Settings (i) to (iv) as in Chapter 6.4.1 above, however this time I made \( \tilde{Y}_{i2} \) available for all \( n \) studies but \( \tilde{Y}_{i1} \) only available when its sign was positive. Importantly, as I generated the data, I knew that the \( \tilde{Y}_{i2} \)s were related to the missing \( \tilde{Y}_{i1} \)s and therefore MAR rather than NMAR is the correct mechanism for the missing data (see Section 6.2). This is undoubtedly an extreme example of MAR, but it may occur to a certain degree if authors report fewer outcomes when non-positive results are obtained. The key interest from these simulations is how effective is Model A in obtaining a more appropriate (e.g. least biased) estimate of \( \beta_1 \) than is possible from the URMA.

Averaging across all the simulations, the results show that, \textit{for all the settings assessed}, \( \tilde{\beta}_1 \) from the URMA is well above the true value of zero because negative \( \tilde{Y}_{i1} \) results were not available (Table 6.4); for the same reason, the URMA estimate of \( r_1^2 \) is far too small. The coverage of the 95\% confidence intervals for \( \tilde{\beta}_1 \) from the URMA is therefore also too small, and the intervals do not include the true value of zero most of the time (Table 6.4). Model A however gives far more reliable (i.e. less biased) results as it is able to ‘borrow strength’ from the negative \( \tilde{Y}_{i2} \) results and shrink the value of \( \tilde{\beta}_1 \) from the URMA down toward the true answer of zero whilst increasing \( \tilde{r}_1^2 \). The coverage for \( \tilde{\beta}_1 \) is therefore vastly improved, with a much larger number of intervals containing the true value of zero.
Table 6.4: Simulation results from the univariate (URMA) and bivariate (Model A) random-effects meta-analysis models for 50 and 10 studies, where outcome \( j = 2 \) was available from all studies but outcome \( j = 1 \) was missing at random (MAR) if \( y_{ni} \) was negative. The true parameter values to compare the results to are \( \beta_1 = 0, \beta_2 = 2, \tau_1^2 = 0.25, \tau_2^2 = 0.25 \). In Model A the within-study correlation \( (\rho_w) \) was known and the same for each study, whilst the between-study correlation \( (\rho_B) \) was estimated. The values of \( \rho_w \) and \( \rho_B \) in each simulation are shown in the table. Two settings were investigated for different sizes of the \( s_j^2 \) in relation to the \( \tau_j^2 \).

(i) On average the \( s_j^2 \) s are similar in size to the \( \tau_j^2 \).

| No. of studies (n) | Meta-analysis model | \( \rho_w \) | \( \rho_B \) | Converged out of 1000 | Mean of \( \hat{\beta}_1 \) (s.e. of mean) | MSE of \( \hat{\beta}_1 \) | No. of 95% CIs for \( \hat{\beta}_1 \) including \( \beta_1 \) (%) | Mean of \( \hat{\beta}_2 \) (s.e.) | MSE of \( \hat{\beta}_2 \) | No. of 95% CIs for \( \hat{\beta}_2 \) including \( \beta_2 \) (%) | Mean of \( \tau_1^2 \) (no. of \( \tau_1^2 = 0 \)) | Mean of \( \tau_2^2 \) (no. of \( \tau_2^2 = 0 \)) | Mean of \( \hat{\rho}_B \) | No. of \( \hat{\rho}_B = -1 \) (%) | No. of \( \hat{\rho}_B = 1 \) (%) |
|--------------------|---------------------|--------------|-------------|----------------------|---------------------------------------------|----------------|-----------------------------------------------|---------------------------------------------|----------------|-----------------------------------------------|------------------------|-------------------------------|------------------------|------------------------|
| 50                 | URMA                | 0.8          | 0.8         | 998                  | 0.505 (0.105)                              | 0.0928         | 0.266 (1%)                                     | 1.993 (0.107)                              | 0.106                      | 0.011 (94.5%)                                | 943 (9.0%)               | 0.0425 (0.117)               | 0.247 (0%)               | -                      |
| 50                 | Model A             | 0.8          | 0.8         | 983*                 | 0.314 (0.096)                              | 0.0840         | 0.108 (90.8%)                                 | 1.979 (0.103)                              | 0.101                      | 0.011 (93.8%)                                | 922 (9.4%)               | 0.0759 (0)                   | 0.245 (0%)               | 0.504 (9.7%)               | 96 (22.5%)               |
| 50                 | URMA                | 0            | 0.8         | 996                  | 0.505 (0.109)                              | 0.0929         | 0.267 (0%)                                     | 2.000 (0.107)                              | 0.106                      | 0.0114 (94.4%)                              | 940 (9.4%)               | 0.0438 (0.117)               | 0.246 (0%)               | -                      |
| 50                 | Model A             | 0            | 0.8         | 975*                 | 0.480 (0.102)                              | 0.0932         | 0.241 (0%)                                     | 2.001 (0.106)                              | 0.106                      | 0.0113 (94.9%)                              | 925 (9.4%)               | 0.0527 (0)                   | 0.249 (0%)               | 0.554 (8.8%)               | 86 (47.5%)               |

* Convergence is quite difficult, even though there are 50 studies, because \( \tau_1^2 \) is too small (due to the missing data) and with the \( s_j^2 \) much larger, \( \tau_j^2 \) and \( \hat{\rho}_B \) are very poorly defined and struggle to converge. Hence the simulations were not performed for \( n = 10 \).

The 95% CI for \( \hat{\beta}_j \) was calculated using a t-distribution with \((n_j - 1)\) degrees of freedom, where \( n_j \) is the number of studies for which outcome \( j \) was available.
On average the $s_{ij}^2$ s are much smaller in size to the $\tau_i^2$

| No. of studies (n) | Meta-analysis model | $\rho_W$ | $\rho_B$ | Converged out of 1000 | Mean of $\hat{\beta}_i$ (s.e. of mean) | Mean s.e. of $\hat{\beta}_i$, MSE of $\hat{\beta}_i$ | No. of 95% CIs for $\hat{\beta}_i$ including $\beta_i$ (%) | Mean of $\hat{\beta}_2$ (s.e.) | Mean s.e. of $\hat{\beta}_2$, MSE of $\hat{\beta}_2$ | No. of 95% CIs for $\hat{\beta}_2$ including $\beta_2$ (%) | Mean $\tilde{\tau}_1^2$ (no. of $\tilde{\tau}_1^2 = 0$) | Mean $\tilde{\tau}_2^2$ (no. of $\tilde{\tau}_2^2 = 0$) | Mean $\tilde{\rho}_B = -1$ (%) | Mean $\tilde{\rho}_B = 1$ (%) | No. of $\tilde{\rho}_B = 1$ (%) |
|---------------------|---------------------|---------|---------|----------------------|-----------------------------|---------------------------------|-----------------------------|-----------------------------|---------------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|
| 50                  | URMA                | 0.8     | 0.8     | 1000                | 0.397 (0.061)               | 0.060 0.161 0 (0%)          | 1.999 (0.0688) 0.0709 0.0047 964 (96.4%) | 0.084 0.245                  | -                              | -                              | -                              | -                              | -                              | -                              |
| 50                  | Model A             | 0.8     | 0.8     | 999                 | 0.243 (0.062)               | 0.058 0.063 37 (3.7%)       | 1.999 (0.0689) 0.0709 0.0047 960 (96.1%) | 0.111 0.245                  | 0.703 0 (0%)                  | 0 (0%)                        | 0 (0%)                        | -                              | -                              |
| 50                  | URMA                | 0       | 0.8     | 1000                | 0.400 (0.064)               | 0.061 0.164 0 (0%)          | 2.000 (0.0699) 0.0706 0.0049 954 (95.4%) | 0.087 0.242                  | -                              | -                              | -                              | -                              | -                              |
| 50                  | Model A             | 0       | 0.8     | 1000                | 0.248 (0.063)               | 0.059 0.065 45 (4.5%)       | 2.000 (0.0699) 0.0706 0.0049 955 (95.5%) | 0.112 0.243                  | 0.719 0 (0%)                  | 0 (0%)                        | 0 (0%)                        | -                              | -                              |
| 10                  | URMA                | 0.8     | 0.8     | 991                 | 0.396 (0.145)               | 0.124 0.178 383 (38.6%)     | 2.005 (0.152) 0.155 0.0230 951 (96.0%) | 0.082 0.249                  | -                              | -                              | -                              | -                              | -                              |
| 10                  | Model A             | 0.8     | 0.8     | 990                 | 0.233 (0.255)               | 0.125 0.119 571 (57.7%)     | 2.004 (0.152) 0.155 0.0230 946 (95.6%) | 0.163 0.248                  | 0.591 31 (3.1%)               | 122 (12.3%)                   | -                              | -                              |
| 10                  | URMA                | 0       | 0.8     | 985                 | 0.402 (0.151)               | 0.126 0.184 404 (41.0%)     | 2.007 (0.166) 0.154 0.0274 924 (93.8%) | 0.084 0.244                  | -                              | -                              | -                              | -                              | -                              |
| 10                  | Model A             | 0       | 0.8     | 981                 | 0.259 (0.317)               | 0.126 0.167 598 (61.0%)     | 2.007 (0.166) 0.154 0.0275 919 (93.7%) | 0.176 0.244                  | 0.625 40 (4.1%)               | 161 (16.4%)                   | -                              | -                              |

N.B. $\tilde{\rho}_B$ was not applicable when one or both of $\tilde{\tau}_1^2$ and $\tilde{\tau}_2^2$ was zero. MSE = mean-square-error, CIs = confidence intervals, s.e. = standard error.

The 95% CI for $\hat{\beta}_j$ was calculated using a t-distribution with $(n_j - 1)$ degrees of freedom, where $n_j$ is the number of studies for which outcome $j$ was available.
I showed through the analytic solutions of Model A that larger correlations enable Model A to yield more benefit (see Point (2) in Section 4.5). This is again emphasised by these simulation results, as \( \hat{\beta}_1 \) does not shrink as far in those settings where the within-study correlation is zero compared to those settings where it is 0.8. This is still true but less evident where \( \tau^2_i \) and \( \tau^2_j \) are much larger than the \( s^2_0 \) s, because the \( s^2_0 \) s and \( \rho_w \) have smaller impact due to the relatively larger \( \tau^2_i \) and \( \tau^2_j \) being the most influential, together with \( \tilde{\rho}_b \) (Table 6.4(ii)). These findings also illustrate that both within-study and between-study correlations can be influential on the pooled estimates, with the within-study correlation more important where the \( s^2_0 \) s are larger than the \( \tau^2_j \) s and the between-study correlation more important where the \( \tau^2_j \) s are larger than the \( s^2_0 \) s.

It would make sense if the shrinkage of \( \hat{\beta}_1 \) from Model A is greater in the \( n = 50 \) results than the \( n = 10 \) results, as the former incorporates more information. However, this is not witnessed in the results for \( \rho_w = 0, \rho_B = 0.8 \) in Table 6.4(ii) because the shrinkage in \( \hat{\beta}_1 \) is slightly more for \( n = 10 \) than for \( n = 50 \). One possible explanation for this is that in the \( n = 10 \) simulations Model A more commonly estimates \( \rho_B \) as 1 or -1 than in the \( n = 50 \) simulations. This problem causes \( \tilde{\tau}^2_i \) and \( \tilde{\tau}^2_j \) to be on average upwardly biased (see Chapter 5.7) and so this inflation, combined with the large value of \( \tilde{\rho}_b \), may well be enabling \( \hat{\beta}_1 \) to shrink further away from the URMA solution and closer toward zero. However, even though this leads to a \( \hat{\beta}_1 \) closer to the true answer and an improved coverage for this particular instance, in reality one must be cautious about using Model A results here because they are based on an upwardly biased \( \tilde{\tau}^2_i \) and \( \tilde{\tau}^2_j \), which can equally lead to misleading pooled estimates and conclusions. This was discussed in Chapter 5.6.4, where the dangers of using Model A when \( \rho_B \) is estimated as 1 or -1 were stressed and
this advice stands equally for missing data as it did for complete-case data. Further
evidence of the dangers of using Model A given biased $\hat{r}_j$'s will be shown in Sections
7.5.3, 8.3.5 and 8.3.6.

6.5 Main conclusions about using Model A to estimate $\beta_j$ given missing data

When there are missing summary statistics across studies a URMA is not ideal because it
does not incorporate those studies for which a summary statistic for the outcome of interest
is missing, and this can potentially lead to imprecise or, even worse, biased pooled results.
However, as Model A does utilise those studies for which only one outcome was missing,
Model A can thereby achieve substantial benefits over URMA for estimating $\beta_j$, assuming
the missing summary statistics are either MCAR or MAR. Firstly, when data is MCAR, the
simulations show that Model A produces a smaller variance and smaller MSE of $\hat{\beta}_j$ than
two independent URMAs on average, and secondly when data is MAR Model A produces
a smaller MSE and a more suitable coverage (i.e. closer to 95%) of $\hat{\beta}_j$ than two
independent URMAs on average. Furthermore, the larger the within- and between-study
correlations the larger the benefits of Model A over URMA. Hence, compared to a URMA,
Model A can potentially produce more reliable (i.e. least biased with more suitable
standard error, MSE and therefore coverage) evidence-based results about $\hat{\beta}_j$ when
missing data is MCAR or MAR. This is a very important finding, and should help
researchers to understand why bivariate meta-analysis may be useful in practice (see
original aims in Section 3.7). It is important to note that these benefits remain even when
considering just the subset of occasions where $\hat{\rho}_p$ did not equal 1 or $-1$ in Model A (Table
6.5). I am aware that these results still assume the within-study correlations are known,
even if in practice this may often be unrealistic (see Section 3.6.2). However, the benefits
for missing data can remain high even when $\rho_{wi} = 0$ (e.g. see Table 6.4), which can be a
sensible assumption in some situations [148;161].
Table 6.5: Comparison of univariate (URMA) versus bivariate (Model A) random-effects meta-analysis simulation results where the between-study correlation ($\rho_B$) was not equal to either 1 or -1 in Model A.

(i) Simulations results just for the 354 simulations where $\hat{\rho}_B$ was not equal to either 1 or -1 in Model A for the missing completely at random (MCAR) data for $n = 10$ where $\rho_w = \rho_w = 0.8$ and on average the $s_i^2$s were similar in size to the $\tau^2_1$ (for full results see Table 6.3(i))

<table>
<thead>
<tr>
<th>Meta-analysis model</th>
<th>Mean of $\beta_1$ (s.e. of mean)</th>
<th>Mean s.e. of $\beta_1$</th>
<th>Mean of $\beta_2$ (s.e. of mean)</th>
<th>Mean s.e. of $\beta_2$</th>
<th>Mean $\tau^2_1$</th>
<th>Mean $\tau^2_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>URMA</td>
<td>-0.0151 (0.0122)</td>
<td>0.230</td>
<td>1.963 (0.016)</td>
<td>0.301</td>
<td>0.296</td>
<td>0.339</td>
</tr>
<tr>
<td>Model A</td>
<td>-0.0149 (0.0120)</td>
<td>0.227</td>
<td>1.972 (0.014)</td>
<td>0.265</td>
<td>0.299</td>
<td>0.341</td>
</tr>
</tbody>
</table>

(ii) Simulations results just for the 837 simulations where $\hat{\rho}_B$ was not equal to either 1 or -1 for the missing at random (MAR) data for $n = 10$ where $\rho_w = \rho_w = 0.8$ and on average the $s_i^2$s were much smaller in size to the $\tau^2_1$ (for full results see Table 6.4(ii))

<table>
<thead>
<tr>
<th>Meta-analysis model</th>
<th>Mean of $\beta_1$ (s.e. of mean)</th>
<th>Mean s.e. of $\beta_1$</th>
<th>Mean of $\beta_2$ (s.e. of mean)</th>
<th>Mean s.e. of $\beta_2$</th>
<th>Mean $\tau^2_1$</th>
<th>Mean $\tau^2_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>URMA</td>
<td>0.399 (0.0048)</td>
<td>0.123</td>
<td>2.023 (0.0051)</td>
<td>0.156</td>
<td>0.0840</td>
<td>0.250</td>
</tr>
<tr>
<td>Model A</td>
<td>0.261 (0.0063)</td>
<td>0.124</td>
<td>2.023 (0.0051)</td>
<td>0.155</td>
<td>0.138</td>
<td>0.249</td>
</tr>
</tbody>
</table>

It is, however, important to re-emphasise the caveats to using Model A when there is missing data. Firstly, Model A is only applicable when the missing summary statistic is either MCAR or MAR in those studies only providing one outcome; if one cannot rule out that such data is NMAR (for example where there is the potential for dissemination bias) one should consider Model A only as a sensitivity analysis in order to assess the robustness of the URMA results under the assumption of MCAR or MAR in those studies (see Section 6.3.2); furthermore one should also consider the potential impact on URMA results if one assumes all missing data is NMAR (see Chapter 9). If one can assume MCAR or MAR in those studies providing only one outcome, then another issue for Model A is that it does not consider missing data from those studies for which neither outcome were available. Even if MCAR or MAR is plausible for those studies providing one outcome, it may not be plausible to assume MCAR or MAR in other studies. There is therefore a need to assess the threat of dissemination bias and that data is NMAR even after fitting Model
A; this issue will be considered in Chapter 9. It is important to emphasise again here that, aside from the problem of unknown within-study correlations, Model A is only suitable as a sensitivity analysis for the OS and DFS HRs from the neuroblastoma review (see Section 6.3), and ideally the potential impact of dissemination bias and data NMAR should also be made in relation to this dataset (see Chapter 9).

Another concern for whenever Model A is applied to missing data is that $\rho_B$ is still often estimated as 1 or $-1$, and one must still be extremely cautious about using Model A results in this situation because inappropriate (i.e. biased) evidence-based conclusions may be made (see Section 5.6.4). However, importantly $\rho_B$ is not always estimated as 1 or $-1$ when Model A is applied to missing data, and there are no problems of bias on these occasions (for example, see the $n = 50$ simulation results in Table 6.3(ii)). Finally, all the benefits established in this chapter about using Model A when there is missing data have assumed the within-study correlations (i.e. the $\rho_w$ s) are known. As mentioned previously, this may not be the case in reality (e.g. for OS and DFS HRs from prognostic marker studies) and so, even though Model A may be highly desirable, it may not be easily applicable in practice. This issue will be tackled in Chapter 8.

6.6 The clear benefit of Model A for estimating $(\beta_1 - \beta_2)$ and $\text{var}(\hat{\beta}_1 - \hat{\beta}_2)$ for both complete-case data and missing data

So far I have only considered Model A for estimating $\beta_j$ and not $(\beta_1 - \beta_2)$. However, the difference in the pooled estimates may be the main motivation for the use of Model A in some circumstances; for example, when $\hat{\beta}_j$ is a log-odds for treatment $j$, then $(\hat{\beta}_1 - \hat{\beta}_2)$ is the log-odds ratio which may be the most important statistic for making preferences about the $j$ treatments. The research in Chapter 5 and Section 6.1 to 6.5 has helped establish that Model A is only beneficial over a URMA for estimating $\beta_j$ when there is missing data.
(and it is MCAR or MAR). However, in terms of \( (\hat{\beta}_1 - \hat{\beta}_2) \), Model A is clearly always preferred (in complete-case, MCAR and MAR situations) because, unlike in a URMA, it allows the \( \text{corr}(\hat{\beta}_1, \hat{\beta}_2) \) to be estimated and therefore a more appropriate estimate of \( \text{var}(\hat{\beta}_1 - \hat{\beta}_2) \) is obtained (see Remark 6 in Section 4.8.6). To demonstrate this I obtained the \( (\hat{\beta}_1 - \hat{\beta}_2) \) results from a variety of the complete-case simulations of Chapter 5 and the missing data simulations in Section 6.4 (Tables 6.6, 6.7 and 6.8). The results show that the benefits of Model A are considerable, for example in terms of gain in precision and reduction in MSE for \( (\hat{\beta}_1 - \hat{\beta}_2) \) over and above URMA. These benefits are true for both the complete-case and missing data situations, and they are far greater than those benefits for the individual \( \hat{\beta}_j \)s themselves. The coverage for \( (\hat{\beta}_1 - \hat{\beta}_2) \) from Model A is around 95%, also a vast improvement over the high coverage close to 100% from the URMA.

The estimate \( (\hat{\beta}_1 - \hat{\beta}_2) \) from Model A is clearly related to the individual pooled estimates themselves, and behaves similarly. For example, for the complete-case data and MCAR simulations, \( (\hat{\beta}_1 - \hat{\beta}_2) \) is unbiased on average (as the average estimate is very close to the true answer of -2 across simulations). Furthermore, for the extreme MAR simulations in Table 6.8, Model A gives a \( (\hat{\beta}_1 - \hat{\beta}_2) \) much closer to the true value than the URMA. In this situation, Model A shrinks \( (\hat{\beta}_1 - \hat{\beta}_2) \) from the URMA toward its true value, and it can do this more the larger the within- and between-study correlations, and the greater the number of studies (see Section 6.4.2). It has been most common in the published literature for \( (\hat{\beta}_1 - \hat{\beta}_2) \) to be of interest from Model A when the within-study correlations are assumed zero [148], and importantly the simulation results show that where this assumption is true Model A is still beneficial (e.g. in terms of gain in precision, reduction in MSE and improved coverage) over two independent URMAs (Tables 6.7 and 6.8).
Table 6.6: Results for $\hat{\beta}_i - \hat{\beta}_j$ from some of the complete-case simulations of Tables 5.1, 5.2 and 5.4 of Chapter 5 for both a univariate (URMA) and a bivariate (Model A) random-effects meta-analysis.

(i) $\tau_1^2 = \tau_2^2 = 0.25$ (similar in size to the $s_i^2$ s on average - see results from Tables 5.1 and 5.2)

<table>
<thead>
<tr>
<th>No. of studies (n)</th>
<th>$p_W$</th>
<th>$p_B$</th>
<th>Meta-analysis model</th>
<th>No. simulations converged</th>
<th>Mean $(\hat{\beta}_1 - \hat{\beta}_2)$ (s.e. of mean)</th>
<th>Mean s.e of $\left(\hat{\beta}_1 - \hat{\beta}_2\right)$</th>
<th>Mean Corr $\left(\hat{\beta}_1, \hat{\beta}_2\right)$</th>
<th>MSE of $\left(\hat{\beta}_1 - \hat{\beta}_2\right)$</th>
<th>No. of 95% CIs for $\left(\hat{\beta}_1 - \hat{\beta}_2\right)$ including true value (%)</th>
<th>No. of $\hat{\rho}_B = -1$ (%)</th>
<th>No. of $\hat{\rho}_B = 1$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.8</td>
<td>0.8</td>
<td>URMA</td>
<td>1000</td>
<td>-1.997 (0.0841)</td>
<td>0.147</td>
<td>-</td>
<td>0.0071</td>
<td>992 (99.4%)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>50</td>
<td>0.8</td>
<td>0.8</td>
<td>Model A</td>
<td>1000</td>
<td>-1.998 (0.0749)</td>
<td>0.073</td>
<td>0.710 (0.075)</td>
<td>0.0056</td>
<td>940 (94.0%)</td>
<td>0 (0%)</td>
<td>20 (2%)</td>
</tr>
<tr>
<td>5</td>
<td>0.8</td>
<td>0.8</td>
<td>URMA</td>
<td>1000</td>
<td>-1.992 (0.215)</td>
<td>0.382</td>
<td>-</td>
<td>0.0459</td>
<td>999 (99.9%)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>0.8</td>
<td>0.8</td>
<td>Model A</td>
<td>998</td>
<td>-1.994 (0.204)</td>
<td>0.187</td>
<td>0.730 (0.193)</td>
<td>0.0415</td>
<td>966 (96.7%)</td>
<td>124 (12.4%)</td>
<td>535 (53.6%)</td>
</tr>
</tbody>
</table>

(ii) $\tau_1^2 = \tau_2^2 = 0.0025$ (much smaller than the $s_i^2$ s on average - see results from Table 5.4(i))

<table>
<thead>
<tr>
<th>No. of studies (n)</th>
<th>$p_W$</th>
<th>$p_B$</th>
<th>Meta-analysis model</th>
<th>No. simulations converged</th>
<th>Mean $(\hat{\beta}_1 - \hat{\beta}_2)$ (s.e. of mean)</th>
<th>Mean s.e of $\left(\hat{\beta}_1 - \hat{\beta}_2\right)$</th>
<th>Mean Corr $\left(\hat{\beta}_1, \hat{\beta}_2\right)$</th>
<th>MSE of $\left(\hat{\beta}_1 - \hat{\beta}_2\right)$</th>
<th>No. of 95% CIs for $\left(\hat{\beta}_1 - \hat{\beta}_2\right)$ including true value (%)</th>
<th>No. of $\hat{\rho}_B = -1$ (%)</th>
<th>No. of $\hat{\rho}_B = 1$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.8</td>
<td>0.8</td>
<td>URMA</td>
<td>998</td>
<td>-2.002 (0.0535)</td>
<td>0.0815</td>
<td>-</td>
<td>0.00286</td>
<td>992 (99.4%)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>50</td>
<td>0.8</td>
<td>0.8</td>
<td>Model A</td>
<td>951</td>
<td>-2.001 (0.0399)</td>
<td>0.0421</td>
<td>0.535 (0.055)</td>
<td>0.00160</td>
<td>920 (96.7%)</td>
<td>208 (21.9%)</td>
<td>534 (56.2%)</td>
</tr>
</tbody>
</table>

(iii) $\tau_1^2 = \tau_2^2 = 1.5$ (much larger than the $s_i^2$ s on average - see results from Table 5.4(iii))

<table>
<thead>
<tr>
<th>No. of studies (n)</th>
<th>$p_W$</th>
<th>$p_B$</th>
<th>Meta-analysis model</th>
<th>No. simulations converged</th>
<th>Mean $(\hat{\beta}_1 - \hat{\beta}_2)$ (s.e. of mean)</th>
<th>Mean s.e of $\left(\hat{\beta}_1 - \hat{\beta}_2\right)$</th>
<th>Mean Corr $\left(\hat{\beta}_1, \hat{\beta}_2\right)$</th>
<th>MSE of $\left(\hat{\beta}_1 - \hat{\beta}_2\right)$</th>
<th>No. of 95% CIs for $\left(\hat{\beta}_1 - \hat{\beta}_2\right)$ including true value (%)</th>
<th>No. of $\hat{\rho}_B = -1$ (%)</th>
<th>No. of $\hat{\rho}_B = 1$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.8</td>
<td>0.8</td>
<td>URMA</td>
<td>1000</td>
<td>-1.995 (0.146)</td>
<td>0.282</td>
<td>-</td>
<td>0.0212</td>
<td>1000 (100%)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>50</td>
<td>0.8</td>
<td>0.8</td>
<td>Model A</td>
<td>1000</td>
<td>-1.994 (0.140)</td>
<td>0.134</td>
<td>0.760 (0.063)</td>
<td>0.0196</td>
<td>932 (93.2%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>5</td>
<td>0.8</td>
<td>0.8</td>
<td>URMA</td>
<td>1000</td>
<td>-1.992 (0.382)</td>
<td>0.784</td>
<td>-</td>
<td>0.145</td>
<td>999 (99.9%)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>0.8</td>
<td>0.8</td>
<td>Model A</td>
<td>1000</td>
<td>-1.994 (0.379)</td>
<td>0.347</td>
<td>0.757 (0.264)</td>
<td>0.144</td>
<td>950 (95.0%)</td>
<td>23 (23.0%)</td>
<td>190 (19.0%)</td>
</tr>
</tbody>
</table>
Table 6.7: Results for \((\tilde{\beta}_1 - \tilde{\beta}_2)\) from the missing completely at random (MCAR) simulations in Table 6.3(i) for both a univariate (URMA) and a bivariate (Model A) random-effects meta-analysis, where on average the \(s_{ij}^2\)s are similar in size to the \(\tau^2\).

<table>
<thead>
<tr>
<th>No. of studies ((n))</th>
<th>(\rho_w)</th>
<th>(\rho_B)</th>
<th>Meta-analysis model</th>
<th>No. simulations converged</th>
<th>Mean ((\tilde{\beta}_1 - \tilde{\beta}_2)) (\text{(s.e. of mean)})</th>
<th>Mean s.e of ((\tilde{\beta}_1 - \tilde{\beta}_2))</th>
<th>Mean (\text{Cov} (\tilde{\beta}_1, \tilde{\beta}_2))</th>
<th>MSE of ((\tilde{\beta}_1 - \tilde{\beta}_2)) including true value ((%))</th>
<th>No. of 95% CIs for ((\tilde{\beta}_1 - \tilde{\beta}_2)) including true value ((%))</th>
<th>No. of (\tilde{\rho}_B = 1) ((%))</th>
<th>No. of (\tilde{\rho}_B = -1) ((%))</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0</td>
<td>0.8</td>
<td>URMA</td>
<td>1000</td>
<td>-2.003 (0.154)</td>
<td>0.178</td>
<td>-</td>
<td>0.0237</td>
<td>979 (97.9%)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>0.8</td>
<td>Model A</td>
<td>1000</td>
<td>-2.001 (0.145)</td>
<td>0.141</td>
<td>0.330 (0.139)</td>
<td>0.0210</td>
<td>959 (95.9%)</td>
<td>2 (0.2%)</td>
<td>95 (9.5)</td>
</tr>
<tr>
<td>50</td>
<td>0.8</td>
<td>0.8</td>
<td>URMA</td>
<td>1000</td>
<td>-2.000 (0.122)</td>
<td>0.178</td>
<td>-</td>
<td>0.0149</td>
<td>995 (99.5%)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>50</td>
<td>0.8</td>
<td>0.8</td>
<td>Model A</td>
<td>1000</td>
<td>-1.998 (0.100)</td>
<td>0.094</td>
<td>0.643 (0.105)</td>
<td>0.0101</td>
<td>940 (94.0%)</td>
<td>1 (0.1%)</td>
<td>356 (35.6%)</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0.8</td>
<td>URMA</td>
<td>1000</td>
<td>-1.996 (0.295)</td>
<td>0.349</td>
<td>-</td>
<td>0.0871</td>
<td>996 (99.6%)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0.8</td>
<td>Model A</td>
<td>999</td>
<td>-1.994 (0.289)</td>
<td>0.280</td>
<td>0.298 (0.283)</td>
<td>0.0835</td>
<td>993 (99.4%)</td>
<td>105 (10.5%)</td>
<td>540 (54.1%)</td>
</tr>
<tr>
<td>10</td>
<td>0.8</td>
<td>0.8</td>
<td>URMA</td>
<td>1000</td>
<td>-1.991 (0.253)</td>
<td>0.346</td>
<td>-</td>
<td>0.0640</td>
<td>997 (99.7%)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>10</td>
<td>0.8</td>
<td>0.8</td>
<td>Model A</td>
<td>997</td>
<td>-1.994 (0.213)</td>
<td>0.178</td>
<td>0.674 (0.203)</td>
<td>0.0453</td>
<td>959 (96.2%)</td>
<td>101 (10.1%)</td>
<td>600 (60.2%)</td>
</tr>
</tbody>
</table>

Table 6.8: Results for \((\tilde{\beta}_1 - \tilde{\beta}_2)\) from the missing at random (MAR) simulations in Table 6.4(i) for both a univariate (URMA) and bivariate (Model A) random-effects meta-analysis, where on average the \(s_{ij}^2\)s are similar in size to the \(\tau^2\).

<table>
<thead>
<tr>
<th>No. of studies ((n))</th>
<th>(\rho_w)</th>
<th>(\rho_B)</th>
<th>Meta-analysis model</th>
<th>No. simulations converged</th>
<th>Mean ((\tilde{\beta}_1 - \tilde{\beta}_2)) (\text{(s.e. of mean)})</th>
<th>Mean s.e of ((\tilde{\beta}_1 - \tilde{\beta}_2))</th>
<th>Mean (\text{Cov} (\tilde{\beta}_1, \tilde{\beta}_2))</th>
<th>MSE of ((\tilde{\beta}_1 - \tilde{\beta}_2)) including true value ((%))</th>
<th>No. of 95% CIs for ((\tilde{\beta}_1 - \tilde{\beta}_2)) including true value ((%))</th>
<th>No. of (\tilde{\rho}_B = -1) ((%))</th>
<th>No. of (\tilde{\rho}_B = 1) ((%))</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.8</td>
<td>0.8</td>
<td>URMA</td>
<td>998</td>
<td>-1.487 (0.130)</td>
<td>0.142</td>
<td>-</td>
<td>0.161</td>
<td>99 (9.9%)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>50</td>
<td>0.8</td>
<td>0.8</td>
<td>Model A</td>
<td>983*</td>
<td>-1.665 (0.113)</td>
<td>0.096</td>
<td>0.481 (0.155)</td>
<td>0.053</td>
<td>340 (34.6%)</td>
<td>96 (9.8%)</td>
<td>221 (26.6%)</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>0.8</td>
<td>URMA</td>
<td>996</td>
<td>-1.495 (0.142)</td>
<td>0.142</td>
<td>-</td>
<td>0.167</td>
<td>86 (86.9%)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>0.8</td>
<td>Model A</td>
<td>975*</td>
<td>-1.521 (0.137)</td>
<td>0.132</td>
<td>0.129 (0.134)</td>
<td>0.134</td>
<td>99 (10.2%)</td>
<td>86 (8.8%)</td>
<td>463 (47.5%)</td>
</tr>
</tbody>
</table>

* Convergence is quite difficult, even though there are 50 studies, because \(\tilde{T}_1^2\) is too small (due to the missing data) and with the \(S_{ij}^2\) being much larger, \(\tau^2\) and \(\tilde{\rho}_B\) are very poorly defined and struggle to converge. Hence the simulations were not performed for \(n = 10\).
Consider now the simulation results for \( (\hat{\beta}_1 - \hat{\beta}_2) \) when there is the common problem of \( \hat{\rho}_B \) being equal to either 1 or -1 (e.g. see Table 6.6(i) for \( n = 5 \)). As was true for \( \hat{\beta}_i \) and \( \hat{\beta}_2 \), on average \( (\hat{\beta}_1 - \hat{\beta}_2) \) remains unbiased across simulations even when this problem is predominant. However, this is misleading because by averaging across simulations one ignores the problem of biased \( \hat{\tau}^2 \)'s from any single occasion where \( \rho_B \) is estimated as 1 or -1. Bias in \( \hat{\tau}^2 \) can change the weighting of each study toward the overall pooled estimates, and can therefore lead to misleading pooled estimates and evidence-based conclusions (Section 5.6.4). The bias averages itself out across all simulations (which is why the average simulation results for \( (\hat{\beta}_1 - \hat{\beta}_2) \) are unbiased) but for one single Model A result the bias may be influential. This issue is a similar concern for the \( \hat{\beta}_i \)'s themselves (Section 5.7.4), however there is the additional concern for \( (\hat{\beta}_1 - \hat{\beta}_2) \) as to how a \( \hat{\rho}_B \) of 1 or -1 affects the estimate of \( \text{corr}(\hat{\beta}_1, \hat{\beta}_2) \), which is a vital component in the estimation of \( \text{var} (\hat{\beta}_1 - \hat{\beta}_2) \). This is difficult to ascertain, but one possible explanation is that \( \text{corr}(\hat{\beta}_1, \hat{\beta}_2) \) will on average be underestimated when this problem occurs, because \( \rho_B \) is itself underestimated on average across simulations (for example see \( n = 10 \) simulation results in Table 6.3(i)).

The main message is that caution is clearly needed before using the results from Model A when \( \hat{\rho}_B \) is equal to either 1 or -1. However, looking at the simulation results in Tables 6.6 to 6.8 one could argue that Model A still gives more appropriate results (e.g. in terms of coverage and MSE of the pooled estimates) than the URMA even when \( \rho_B \) is predominantly estimated as 1 or -1. For example, Model A gives a more appropriate coverage on average than the URMA for the \( n = 5 \) results in Tables 6.6(i), where a \( \hat{\rho}_B \) equal to 1 or -1 was highly prevalent. In response to this I would agree that the URMA is
far from appropriate, given the unsuitable estimate of $\text{var}(\hat{\beta}_1 - \hat{\beta}_2)$ due to $\text{corr}(\hat{\beta}_1, \hat{\beta}_2)$ being unknown, but I would also say that this does not mean a potentially biased Model A should be used instead. Indeed, the most sensible conclusion is that one must be extremely cautious in the use of either the $(\hat{\beta}_1 - \hat{\beta}_2)$ results from a URMA or the $(\hat{\beta}_1 - \hat{\beta}_2)$ results from Model A when $\hat{\rho}_b$ equals 1 or -1. Fortunately, there are many occasions where Model A does not produce a $\hat{\rho}_b$ equal to 1 or -1 (for example see the $n = 50$ simulation results in Table 6.6(iii)). In these situations Model A does not produce biased answers and is far more beneficial than a URMA for estimating $(\beta_1 - \beta_2)$ for either complete-case or missing data.

## 6.7 Extensions of Model A to other situations

I have established some important findings about using Model A to estimate $\beta_j$ and $(\beta_1 - \beta_2)$, both for complete-case and missing data (these will clearly summarised in Section 8.7). It is important to now consider how these extend to other important scenarios not covered so far. These extensions continue to assume that the within-study correlations are known (this issue will be considered in detail in Chapter 8).

### 6.7.1 Negative correlation

All my simulations have been based on non-negative within- and between-study correlations as I felt this to be the most likely scenario one would encounter in reality. In Table 6.9 I now present the results from simulations that used negative correlation parameters and one can see that they show consistent findings with those from the positive correlation simulations. Firstly, there is no benefit of Model A over URMA for estimating $\beta_j$ for complete-case data (Table 6.9(i)). Secondly, Model A is still beneficial for MCAR or MAR situations for estimating $\beta_j$, with gain in precision and MSE for $\hat{\beta}_j$ still evident (Table 6.9(ii) and (iii)). Furthermore, for all simulations one still obtains a more
appropriate estimate of $\text{var}(\hat{\beta}_1 - \hat{\beta}_2)$ than the URMA due to the incorporation of $\text{corr}(\hat{\beta}_1, \hat{\beta}_2)$. However, unlike for positive correlation simulations (see Section 6.6), this more appropriate $\text{var}(\hat{\beta}_1 - \hat{\beta}_2)$ could now be larger than the URMA estimate because $\text{corr}(\hat{\beta}_1, \hat{\beta}_2)$ may be negative. This is very important because in this situation the URMA could produce results which are too precise, and which could then lead to strong but misleading conclusions. Finally, it is also important to note that $\rho_B$ being estimated as 1 or -1 is still a problem for the negative correlation simulations (except $\bar{\rho}_B = -1$ is now more prevalent than $\bar{\rho}_B = 1$) and this still causes an upward bias in $\bar{\tau}^2_j$ on average.

6.7.2 Relation of findings to bivariate fixed-effects meta-analysis

I have so far compared a bivariate random-effects model (Model A) to a univariate random-effects model (URMA). However, the benefits for estimating $\beta_j$ and $(\beta_1 - \beta_2)$ from a bivariate fixed-effect meta-analysis model (i.e. where $\tau^2_j = 0$ and $\tau_{12} = 0$ in Model A, see equation (3.3) in Section 3.6.1) over a univariate fixed-effects meta-analysis will be the same as those identified for Model A over a URMA. For example, the main benefits for $\bar{\beta}_j$ will arise when some data is MCAR or MAR and there are large within-study correlations. Indeed, bivariate fixed-effects results will be very similar to those presented for Model A where the $s^2_j$'s are much larger than the $\bar{\tau}^2_j$'s on average (e.g. see Table 5.4(i) in Section 5.4.2). However, in a bivariate fixed-effects meta-analysis there is not the additional problem of estimating the between-study parameters, and thus there is no concern over $\bar{\rho}_B$ because it does not exist. A bivariate fixed-effects model only requires $\beta_1$ and $\beta_2$ to be estimated because the within-study parameters are still assumed known, as they were in Model A. Of course, the problem of unavailable within-study correlations will undoubtedly remain a major methodological hurdle for the approach (see Chapter 8).
Table 6.9: Simulation results from the univariate (URMA) and bivariate (Model A) random-effects meta-analysis models for some negative correlation situations. The true parameter values to compare the results to are $\beta_1 = 0$, $\beta_2 = 2$, $\tau_1^2 = 0.25$, $\tau_2^2 = 0.25$. In Model A the within-study correlation ($\rho_w$) was known and the same for each study, whilst the between-study correlation ($\rho_B$) was estimated. The values of $\rho_w$ and $\rho_B$ in each simulation are as shown, and on average the $s^2_j$ s were similar in size to the $\tau^2_j$.

(i) complete-case data from $n = 5$ studies

<table>
<thead>
<tr>
<th>Meta-analysis model</th>
<th>$\rho_w$</th>
<th>$\rho_B$</th>
<th>Converged out of 1000</th>
<th>Mean of $\hat{\beta}_1$ (s.e. of mean)</th>
<th>Mean s.e. of $\hat{\beta}_1$</th>
<th>MSE $\hat{\beta}_1$</th>
<th>Mean of $\hat{\beta}_2$ (s.e. of mean)</th>
<th>Mean s.e. of $\hat{\beta}_2$</th>
<th>MSE $\hat{\beta}_2$</th>
<th>Mean of $\hat{\tau}_1^2$ (no. of $\hat{\tau}_1^2 = 0$)</th>
<th>Mean of $\hat{\tau}_2^2$ (no. of $\hat{\tau}_2^2 = 0$)</th>
<th>Mean of $\hat{\rho}_B$</th>
<th>No. of $\hat{\rho}_B = -1$ (%)</th>
<th>No. of $\hat{\rho}_B = 1$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>URMA -0.8</td>
<td>0</td>
<td>1000</td>
<td>0.0059 (0.283)</td>
<td>0.267 (96.4%)</td>
<td>0.0801 (99%)</td>
<td>964 (96%)</td>
<td>2.008 (289)</td>
<td>0.266 (298)</td>
<td>0.0888 (96%)</td>
<td>937 (93.7%)</td>
<td>0.246 (89)</td>
<td>0.263 (67)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Model A -0.8</td>
<td>0</td>
<td>998</td>
<td>0.0046 (0.282)</td>
<td>0.269 (98.7%)</td>
<td>0.0793 (99%)</td>
<td>985 (96%)</td>
<td>2.012 (304)</td>
<td>0.283 (304)</td>
<td>0.0927 (96%)</td>
<td>962 (96.4%)</td>
<td>0.267 (5)</td>
<td>0.277 (0)</td>
<td>0.052 (20.4%)</td>
<td>342 (34.5%)</td>
</tr>
<tr>
<td>URMA -0.8</td>
<td>-0.8</td>
<td>1000</td>
<td>-0.0087 (0.294)</td>
<td>0.269 (96.4%)</td>
<td>0.086 (99%)</td>
<td>964 (96%)</td>
<td>2.012 (300)</td>
<td>0.263 (300)</td>
<td>0.0900 (93%)</td>
<td>937 (93.7%)</td>
<td>0.249 (92)</td>
<td>0.253 (99)</td>
<td>0.658 (52.9%)</td>
<td>528 (14.6%)</td>
</tr>
<tr>
<td>Model A -0.8</td>
<td>-0.8</td>
<td>998</td>
<td>-0.0051 (0.285)</td>
<td>0.261 (97.9%)</td>
<td>0.0813 (99%)</td>
<td>977 (96%)</td>
<td>2.011 (292)</td>
<td>0.253 (292)</td>
<td>0.0854 (96.1%)</td>
<td>959 (15)</td>
<td>0.256 (0)</td>
<td>0.255 (-0.658)</td>
<td>528 (52.9%)</td>
<td>145 (14.6%)</td>
</tr>
</tbody>
</table>

Additional results for ($\hat{\beta}_1 - \hat{\beta}_2$):

<table>
<thead>
<tr>
<th>Meta-analysis model</th>
<th>$\rho_w$</th>
<th>$\rho_B$</th>
<th>Mean ($\hat{\beta}_1 - \hat{\beta}_2$)</th>
<th>Mean s.e. of ($\hat{\beta}_1 - \hat{\beta}_2$)</th>
<th>Mean corr ($\hat{\beta}_1, \hat{\beta}_2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>URMA -0.8</td>
<td>0</td>
<td>-2.002</td>
<td>0.464</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Model A -0.8</td>
<td>0</td>
<td>-2.007</td>
<td>0.450</td>
<td>-0.354</td>
<td></td>
</tr>
<tr>
<td>URMA -0.8</td>
<td>-0.8</td>
<td>-2.021</td>
<td>0.382</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Model A -0.8</td>
<td>-0.8</td>
<td>-2.017</td>
<td>0.483</td>
<td>-0.734</td>
<td></td>
</tr>
</tbody>
</table>
(ii) n = 50 studies for which 25 gave complete-case data for both j = 1 and j = 2, but another 25 studies gave complete-case data only for j = 1, whereas the j = 2 data was missing completely at random (MCAR)

| Meta-analysis model | $\hat{\rho}_W$ | $\hat{\rho}_B$ | Converged out of 1000 | Mean of $\hat{\beta}_1$ (s.e. of mean) | Mean s.e. of $\hat{\beta}_1$ | MSE of $\hat{\beta}_1$ | No. of 95% CIs for $\hat{\beta}_1$ including $\beta_1$ (%) | Mean of $\hat{\beta}_2$ (s.e.) | MSE of $\hat{\beta}_2$ | No. of 95% CIs for $\hat{\beta}_2$ including $\beta_2$ (%) | Mean $\hat{\tau}_1^2$ (no. of $\hat{\tau}_1^2 = 0$) | Mean $\hat{\tau}_2^2$ (no. of $\hat{\tau}_2^2 = 0$) | Mean of $\hat{\rho}_B$ | No. of $\hat{\rho}_B = -1$ (%) | No. of $\hat{\rho}_B = 1$ (%) |
|--------------------|---------------|---------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|
| URMA               | -0.8          | -0.8          | 1000                 | -0.0069 (0.106)       | 0.102 (0.0014)      | 928 (93.8%)          | 2.010 (0.146)        | 0.0231               | 937 (93.7%)          | 0.250 (0)             | -                  | -                  |
| Model A            | -0.8          | -0.8          | 1000                 | -0.0075 (0.103)       | 0.0991 (0.0107)     | 941 (94.1%)          | 2.009 (0.119)        | 0.0159               | 947 (94.7%)          | 0.252 (0)             | -0.790 (88)         | 2                  |

Additional results for ($\hat{\beta}_1 - \hat{\beta}_2$):

<table>
<thead>
<tr>
<th>Meta-analysis model</th>
<th>$\rho_W$</th>
<th>$\rho_B$</th>
<th>Mean ($\hat{\beta}_1 - \hat{\beta}_2$)</th>
<th>Mean s.e. of ($\hat{\beta}_1 - \hat{\beta}_2$)</th>
<th>Mean corr($\hat{\beta}_1, \hat{\beta}_2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>URMA</td>
<td>-0.8</td>
<td>-0.8</td>
<td>-2.017 (0.179)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Model A</td>
<td>-0.8</td>
<td>-0.8</td>
<td>-2.016 (0.199)</td>
<td>-0.646</td>
<td>-</td>
</tr>
</tbody>
</table>

(iii) n = 50 studies for which 25 gave complete-case data for j = 1 and j = 2, but another 25 studies gave complete-case data only for j = 2, whereas the j = 2 data was missing at random

| Meta-analysis model | $\rho_W$ | $\rho_B$ | Converged out of 1000 | Mean of $\hat{\beta}_1$ (s.e. of mean) | Mean s.e. of $\hat{\beta}_1$ | MSE of $\hat{\beta}_1$ | No. of 95% CIs for $\hat{\beta}_1$ including $\beta_1$ (%) | Mean of $\hat{\beta}_2$ (s.e.) | MSE of $\hat{\beta}_2$ | No. of 95% CIs for $\hat{\beta}_2$ including $\beta_2$ (%) | Mean $\hat{\tau}_1^2$ (no. of $\hat{\tau}_1^2 = 0$) | Mean $\hat{\tau}_2^2$ (no. of $\hat{\tau}_2^2 = 0$) | Mean of $\hat{\rho}_B$ | No. of $\hat{\rho}_B = -1$ (%) | No. of $\hat{\rho}_B = 1$ (%) |
|--------------------|----------|----------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|
| URMA               | -0.8     | -0.8     | 999                  | 0.507 (0.113)        | 0.0925 (0.0269)     | 999 (100%)           | 2.006 (0.106)        | 0.0125               | 930 (93.1%)          | 0.0422               | 0.249 (115)         | -                  | -                  |
| Model A            | -0.8     | -0.8     | 975                  | 0.312 (0.0100)       | 0.0829 (0.107)      | 866 (88.8%)          | 2.019 (0.101)        | 0.0120               | 922 (94.6%)          | 0.0714               | 0.246 (3)           | -0.492 (235)        | 50                  |

Additional results for ($\hat{\beta}_1 - \hat{\beta}_2$):

<table>
<thead>
<tr>
<th>Meta-analysis model</th>
<th>$\rho_W$</th>
<th>$\rho_B$</th>
<th>Mean ($\hat{\beta}_1 - \hat{\beta}_2$)</th>
<th>Mean s.e. of ($\hat{\beta}_1 - \hat{\beta}_2$)</th>
<th>Mean corr($\hat{\beta}_1, \hat{\beta}_2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>URMA</td>
<td>-0.8</td>
<td>-0.8</td>
<td>-1.500 (0.142)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Model A</td>
<td>-0.8</td>
<td>-0.8</td>
<td>-1.707 (0.159)</td>
<td>-0.485</td>
<td>-</td>
</tr>
</tbody>
</table>

R.D. Riley, Ph.D. Thesis Chapter 6
6.7.3 Trivariate and other higher order multivariate meta-analyses

Although I have focused on two correlated summary statistics, the benefits and limitations of Model A can also be extended to higher order random-effects models, where three or more correlated outcomes are to be synthesised; for example, a trivariate model has been used elsewhere to assess summary statistics from three correlated outcomes [161], and other higher order multivariate meta-analysis models have been applied [150]. As for Model A, the main benefits of such models will arise when there is some data either MCAR or MAR, although the problem of \( \rho_b \) being estimated as 1 or -1 will remain; indeed, other things being equal, this problem may actually be observed more often than in the bivariate setting due to the greater number of between-study correlations that would need estimating. Also, for higher order multivariate models a greater number of within-study correlations would also need to be known, and this would again be a particular methodological issue in many real situations in practice.

6.7.4 Extension of Model A to bivariate random-effects meta-regression

Meta-regression is the term used to denote a meta-analysis model that seeks to explain the between-study heterogeneity by incorporating additional covariates alongside \( \beta_j \), and a univariate meta-regression was introduced in Section 1.7.3. Model A is also quite capable of handling extra covariates if one desires to take such an approach. It is important that I demonstrate this here because where heterogeneity exists it is highly advisable to, wherever possible, explain what is causing it [40].

Berkey et al. extend Model A to a ‘bivariate meta-regression’ for the periodontal disease dataset of Table 4.1, which was originally assessed using Model A in Section 4.7 [151]. The authors consider whether the ‘year of publication minus 1983’ \((X_i)\) of each study may explain some of the heterogeneity of the probing depth and attachment level values.
between studies. The year of publication covariate is centred at 1983 to help convergence and allow the parameter estimate for this covariate to be more interpretable. Model A in this situation is extended to a bivariate meta-regression as follows:

**Model A with regression covariate (‘Model A-reg’)**

\[
\hat{Y}_{ij} \sim N(\theta_{ij}, \delta_i) \\
\delta_i = \begin{pmatrix}
    s^2_{i1} & \lambda_i \\
    \lambda_i & s^2_{i2}
\end{pmatrix}
\]

\[
\theta_{ij} \sim N(\beta_j + \xi_j X_i, \Omega_2) \\
\Omega_2 = \begin{pmatrix}
    \tau^2_1 & \tau_{12} \\
    \tau_{12} & \tau^2_2
\end{pmatrix}
\]

(6.3)

where

\[
i = 1 \text{ to } n \text{ studies} \\
j = \begin{cases}
    1 & \text{for outcome 1} \\
    2 & \text{for outcome 2}
\end{cases}
\]

The term \( \xi_j \) in equation (6.3) is the average change in \( \beta_j \) between two studies published one year apart. Model A-reg can easily be fitted in SAS Proc Mixed (see Appendix B1).

When applying Model A-reg to the Berkey dataset, the ‘year of publication’ covariate did not explain a statistically significant amount of the between-study heterogeneity (Table 6.10). Surprisingly, \( \hat{\tau}_j^2 \) from the meta-regression is actually larger than \( \hat{\tau}_j^2 \) from Model A, which does not include the ‘year of publication’ covariate (Table 6.10). This is somewhat strange because, even though ‘year of publication’ appears to be unimportant, meta-regression should not be increasing the value of \( \hat{\tau}_j^2 \). However, the reason for this increase is due to the use of restricted iterative generalised least squares (RIGLS), which adjusts for the sample size and number of estimated model parameters (i.e. the degrees of freedom) to remove bias in \( \hat{\tau}_j^2 \) (see Section 4.1). The inclusion of an additional covariate reduces by one the degrees of freedom in the RIGLS estimation of \( \tau^2_j \), and as there is a small sample size \( n = 5 \) this has a large impact on the value of \( \hat{\tau}_j^2 \). It is analogous to dividing by 3 instead of 4 in the calculation of \( \hat{\tau}_j^2 \), which makes a large difference compared to \( n = 50 \),
say, where one would be dividing by 48 instead of 49. When I refitted Model A and the 
bivariate meta-regression using non-restricted IGLS estimation, \( \hat{\tau}_j^2 \) was practically 
identical in all the models. This is not surprising given the lack of importance of the 
covariate. I must add here that RIGLS estimation is still a more suitable estimation 
procedure than the non-restricted approach, especially when the sample size is small and 
despite the potential for increasing \( \hat{\tau}_j^2 \) when performing meta-regression, because the 
underlying mathematical theory suggests it produces unbiased variance parameters 
whereas the non-restricted approach does not [51].

Table 6.10: Results from univariate (URMA) and bivariate (Model A) random-effects meta-analysis 
models applied to the Berkey dataset, both with and without the additional 'year of publication' 
regression (reg) covariate

<table>
<thead>
<tr>
<th>Model</th>
<th>( \hat{\beta}_1 ) (s.e.)</th>
<th>( \hat{\tau}_1^2 ) (s.e.)</th>
<th>( \hat{\xi}_1 ) (s.e.)</th>
<th>( \hat{\beta}_2 ) (s.e.)</th>
<th>( \hat{\tau}_2^2 ) (s.e.)</th>
<th>( \hat{\xi}_2 ) (s.e.)</th>
<th>( \bar{\tau}_{12} )</th>
<th>( \hat{\rho}_B )</th>
</tr>
</thead>
<tbody>
<tr>
<td>URMA</td>
<td>0.361</td>
<td>0.012</td>
<td>-</td>
<td>-0.346</td>
<td>0.033</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>(0.059)</td>
<td>(0.089)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>URMA with reg</td>
<td>0.353</td>
<td>0.019</td>
<td>0.0057</td>
<td>-0.329</td>
<td>0.042</td>
<td>-0.0058</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>covariate</td>
<td>(0.072)</td>
<td>(0.023)</td>
<td>(0.101)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model A</td>
<td>0.353</td>
<td>0.012</td>
<td>-</td>
<td>-0.339</td>
<td>0.033</td>
<td>-</td>
<td>0.012</td>
<td>0.609</td>
</tr>
<tr>
<td></td>
<td>(0.059)</td>
<td>(0.088)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model A-reg</td>
<td>0.359</td>
<td>0.021</td>
<td>0.00486</td>
<td>-0.336</td>
<td>0.041</td>
<td>-0.0115</td>
<td>0.013</td>
<td>0.561</td>
</tr>
<tr>
<td></td>
<td>(0.074)</td>
<td>(0.022)</td>
<td>(0.098)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

N.B. My results differ slightly to those in the Berkey paper because I used RIGLS and not IGLS.

The key benefits and principles of Model A that I have identified for \( \hat{\beta}_1, \hat{\beta}_2 \) and \( (\hat{\beta}_1 - \hat{\beta}_2) \) 
will remain when extended to Model A-reg (e.g. gain in precision and MSE over a 
univariate meta-regression when some data is MAR etc.). However, as the addition of 
regression covariates should (the above RIGLS issue aside) reduce the \( \hat{\tau}_j^2 \)s, the impact of 
the between-study parameters during the estimation procedure will also decrease because 
the within-study parameters will become more and more dominant. Unfortunately, this will
make convergence of a bivariate meta-regression increasingly difficult (see Section 5.4.2) and, even when it does converge, there will also be an increased chance of $\rho_b$ being estimated as 1 or -1, because flat likelihoods will become more prevalent for $\rho_b$ (see Section 5.6.2, 'Reason (II)').

To demonstrate this, I replicated the bivariate meta-regression simulations performed by Berkey et al., which included the 'year of publication - 1983' covariate [151]. Firstly, I generated 5000 datasets using Model A-reg, each one involving 5 studies for which $Y_1$ and $Y_2$ were both available. The additional regression covariate in Model A-reg again denoted 'year of publication - 1983', which was (0, -1, -4, 4, 5) for studies 1 to 5 respectively for every simulation. The true parameter values used to simulate data from Model A-reg were the same as those from the Berkey paper, i.e. $\beta_1 = 0.359$, $\beta_2 = -0.336$, $\tau^2 = 0.022$, $\tau^2 = 0.028$, $\rho_b = 0.524$, $\xi_1 = 0.005$, $\xi_2 = -0.011$ and $\rho_w$ as specified in Table 4.1. I then applied and compared four models that were fitted to the simulated datasets using RIGLS. These models were: (i) a separate URMA for each outcome, not including the 'year of publication' covariate; (ii) Model A, which does not include the 'year of publication' covariate; (iii) a separate univariate meta-regression for each outcome, i.e. a URMA including the 'year of publication' covariate; and (iv) Model A-reg, a bivariate meta-regression including the 'year of publication' covariate.

The univariate meta-regression simulation results show that the 'year of publication' covariate slightly reduces the between-study variance for both outcomes on average (Table 6.11). The results from Model A and Model A-reg are very similar for the attachment level outcome ($j = 2$), however for the probing depth outcome ($j = 1$) $\tau^2$ is very slightly larger on average in Model A-reg, which is likely to be caused predominantly by the RIGLS issue as discussed above.
Table 6.11: A replication of the univariate and bivariate random-effects meta-analysis simulations from Berkey et al. where there is complete-case data for \( n = 5 \) and the true parameter values to compare the results to are \( \beta_1 = 0.359, \beta_2 = -0.336, \tau_1^2 = 0.022, \tau_2^2 = 0.028, \rho_B = 0.524 \) [151]. The within-study correlation is different but known for each study (see Table 4.1). The datasets were generated from Model A-reg using true ('year of publication') covariate values of \( \xi_1 = 0.005 \) and \( \xi_2 = -0.011 \). Univariate and bivariate models with (URMA-reg, Model A-reg) and without (URMA, Model A) the regression covariate were fitted to each dataset.

(i) Univariate simulation results

<table>
<thead>
<tr>
<th>Meta-analysis model</th>
<th>Converged out of 5000</th>
<th>Mean of ( \hat{\beta}_1 ) (s.e. of mean)</th>
<th>Mean s.e. of ( \hat{\beta}_1 )</th>
<th>MSE of ( \hat{\beta}_1 ) for ( \beta_1 ) including ( \beta_1 ) (%)</th>
<th>No. of 95% CIs including ( \beta_1 ) (%)</th>
<th>Mean of ( \hat{\beta}_2 ) (s.e. of mean)</th>
<th>Mean s.e. of ( \hat{\beta}_2 )</th>
<th>MSE of ( \hat{\beta}_2 ) for ( \beta_2 ) including ( \beta_2 ) (%)</th>
<th>No. of 95% CIs including ( \beta_2 ) (%)</th>
<th>Mean of ( \hat{\tau}_1^2 ) (no. of ( \hat{\tau}_1^2 ) = 0)</th>
<th>Mean of ( \hat{\tau}_2^2 ) (no. of ( \hat{\tau}_2^2 ) = 0)</th>
<th>Mean of ( \hat{\rho}_B ) = ( \hat{\rho}_B ) = ( \hat{\rho}_B ) = ( \hat{\rho}_B ) = 1 (%)</th>
<th>No. of ( \hat{\rho}_B ) = ( \hat{\rho}_B ) = ( \hat{\rho}_B ) = ( \hat{\rho}_B ) = 1 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>URMA</td>
<td>5000</td>
<td>0.362 (0.0757)</td>
<td>0.0704</td>
<td>0.00573</td>
<td>4868 (97.4%)</td>
<td>-0.341 (0.0827)</td>
<td>0.0780</td>
<td>0.00686</td>
<td>4762 (95.2%)</td>
<td>0.0227 (164)</td>
<td>0.0299 (50)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>URMA-reg</td>
<td>5000</td>
<td>0.360 (0.0760)</td>
<td>0.0696</td>
<td>0.00583</td>
<td>4643 (92.9%)</td>
<td>-0.337 (0.0826)</td>
<td>0.0748</td>
<td>0.00682</td>
<td>4493 (89.9%)</td>
<td>0.0224 (306)</td>
<td>0.0287 (178)</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Additional covariate information:

<table>
<thead>
<tr>
<th>Meta-analysis model</th>
<th>Mean of ( \xi_1 ) (s.e. of mean)</th>
<th>Mean s.e. of ( \xi_1 )</th>
<th>Mean of ( \xi_2 ) (s.e. of mean)</th>
<th>Mean s.e. of ( \xi_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>URMA</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>URMA-reg</td>
<td>0.0050 (0.0228)</td>
<td>0.0206</td>
<td>-0.0112 (0.0260)</td>
<td>0.0231</td>
</tr>
</tbody>
</table>

(ii) Bivariate simulation results

<table>
<thead>
<tr>
<th>Meta-analysis model</th>
<th>Converged out of 5000</th>
<th>Mean of ( \hat{\beta}_1 ) (s.e. of mean)</th>
<th>Mean s.e. of ( \hat{\beta}_1 )</th>
<th>MSE of ( \hat{\beta}_1 ) for ( \beta_1 ) including ( \beta_1 ) (%)</th>
<th>No. of 95% CIs including ( \beta_1 ) (%)</th>
<th>Mean of ( \hat{\beta}_2 ) (s.e. of mean)</th>
<th>Mean s.e. of ( \hat{\beta}_2 )</th>
<th>MSE of ( \hat{\beta}_2 ) for ( \beta_2 ) including ( \beta_2 ) (%)</th>
<th>No. of 95% CIs including ( \beta_2 ) (%)</th>
<th>Mean of ( \hat{\tau}_1^2 ) (no. of ( \hat{\tau}_1^2 ) = 0)</th>
<th>Mean of ( \hat{\tau}_2^2 ) (no. of ( \hat{\tau}_2^2 ) = 0)</th>
<th>Mean of ( \hat{\rho}_B ) = ( \hat{\rho}_B ) = ( \hat{\rho}_B ) = ( \hat{\rho}_B ) = 1 (%)</th>
<th>No. of ( \hat{\rho}_B ) = ( \hat{\rho}_B ) = ( \hat{\rho}_B ) = ( \hat{\rho}_B ) = 1 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model A</td>
<td>4989</td>
<td>0.363 (0.0756)</td>
<td>0.0704</td>
<td>0.00577</td>
<td>4743 (95.1%)</td>
<td>-0.340 (0.0825)</td>
<td>0.0776</td>
<td>0.00688</td>
<td>4670 (93.6%)</td>
<td>0.0228 (2)</td>
<td>0.0301 (0)</td>
<td>0.387 (10.3%)</td>
<td>512 (32.2%)</td>
</tr>
<tr>
<td>Model A-reg</td>
<td>4978</td>
<td>0.360 (0.0763)</td>
<td>0.0706</td>
<td>0.00583</td>
<td>4779 (96.2%)</td>
<td>-0.337 (0.0828)</td>
<td>0.0747</td>
<td>0.00685</td>
<td>4624 (92.9%)</td>
<td>0.0230 (20)</td>
<td>0.0298 (0)</td>
<td>0.424 (773)</td>
<td>1936 (15.9%)</td>
</tr>
</tbody>
</table>

Additional covariate information:

<table>
<thead>
<tr>
<th>Meta-analysis model</th>
<th>Mean of ( \xi_1 ) (s.e. of mean)</th>
<th>Mean s.e. of ( \xi_1 )</th>
<th>Mean of ( \xi_2 ) (s.e. of mean)</th>
<th>Mean s.e. of ( \xi_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model A</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Model A-reg</td>
<td>0.005 (0.0229)</td>
<td>0.0209</td>
<td>-0.0112 (0.0262)</td>
<td>0.0229</td>
</tr>
</tbody>
</table>
As expected the results in Table 6.11 show that adding the additional covariate to form Model A-reg greatly increases the number of occasions where $\hat{\rho}_B$ is equal to either 1 or -1 (54% of the simulations obtained one of these values in Model A-reg compared to 42% in Model A). Although the average bias this introduces across simulations appears small, it is still an undesirable aspect given previous discussions in Sections 5.6.4 and 6.5. Although this example demonstrates that bivariate meta-regression is possible, most importantly it also emphasises that using bivariate meta-regression for small datasets is not advisable. The same is true for univariate meta-regression, because there is very little statistical power to obtain meaningful results [54]. The added problem of $\hat{\rho}_B$ being equal to 1 or -1 when the number of studies is small makes the matter even worse for bivariate meta-regression.

6.8 Summary and rationale for subsequent chapters

This chapter has assessed the benefit of Model A for $(\hat{\beta}_1 - \hat{\beta}_2)$, missing data, and for various other extensions to that considered in Chapter 5. Clearly the major benefit of Model A over URMA is in the estimate $(\hat{\beta}_1 - \hat{\beta}_2)$, the difference between the two pooled values, because (for either complete-case data, data MCAR or data MAR) one can obtain a much more appropriate estimate of $\text{var}(\hat{\beta}_1 - \hat{\beta}_2)$ in Model A due to the incorporation of $\text{corr}(\hat{\beta}_1, \hat{\beta}_2)$, which cannot be calculated using URMA. The use of Model A for $(\hat{\beta}_1 - \hat{\beta}_2)$ in complete-case data settings has been the main focus of the previous multivariate meta-analysis literature (see Section 3.6.3 and 3.7). However, my findings in this chapter show that Model A is even more worthy of consideration over URMA when there is data MCAR or MAR, and the benefits (e.g. reduction in MSE) are not just for $(\hat{\beta}_1 - \hat{\beta}_2)$ but also for $\hat{\beta}_1$ and $\hat{\beta}_2$ themselves. Unavailable summary statistics and missing outcomes are common problems for meta-analysts, particularly those synthesising prognostic marker studies (see Section 3.1). Assuming that the missing summary statistic is either MCAR or
MAR in those studies providing only one outcome, Model A provides a suitable way to
‘borrow strength’ from the known summary statistic for the related outcome, and it allows
one to obtain a $\hat{\beta}_1$ and $\hat{\beta}_2$ with smaller MSE, greater precision and improved coverage
(i.e. closer to 95%) than otherwise possible from a URMA on average. Thus, although
Chapter 5 showed that URMA was a suitable evidence synthesis model for estimating $\beta_j$
when there is complete-case data, this chapter has shown that Model A is preferable to a
URMA for estimating $\beta_j$ when there is missing data, conditional on: (i) the missing
summary statistic being either MCAR or MAR in those studies providing only one of the
two outcomes, and (ii) that $\rho_\theta$ is not equal to either 1 or $-1$.

There are two main issues that still need to be addressed if Model A is to be more
commonly used in practice. The first issue is that the within-study correlations will often
not be available from the published literature, yet these are needed to apply Model A.
Although in some situations it may be possible to assume the within-study correlations are
zero [148;169], this problem forms a major methodological hurdle for using Model A
within evidence-based reviews of prognostic marker studies. The second major issue
remaining is that, even when the within-study correlations are known, it is common for $\rho_\theta$
to be estimated as 1 or $-1$, with potentially misleading results arising from this situation. In
the forthcoming Chapters 7 and 8 I will specifically consider possible alternative methods
and statistical models to help overcome or limit these two problems. Of course, my
ultimate aim remains as specified in Sections 3.1 and 3.7: to facilitate the most suitable
application of BRMA to the OS and DFS HR estimates from the neuroblastoma review
(which will be in the form of a sensitivity analysis – see Section 6.3.2), and to thereby
identify if this approach can help limit and reduce any problems (such as biased pooled
estimates) caused by missing summary statistics and missing outcomes, two issues that
commonly affect meta-analysis of prognostic marker studies (see Sections 3.1 and 3.2).
Chapter overview

A Bayesian framework for a univariate random-effects meta-analysis (URMA) was introduced and contrasted to a frequentist URMA in Section 1.8. In this chapter I will extend the Bayesian approach to a bivariate random-effects meta-analysis (BRMA), and in particular I will consider whether it can suitably limit some of the frequentist estimation problems associated with Model A. I will also draw attention to the issues associated with specifying prior distributions in the Bayesian bivariate framework, and I will discuss how these affect the usefulness of the Bayesian approach in practice.

7.1 Including all parameter uncertainty and prior knowledge about \( \rho_B \)

In terms of Model A, a Bayesian approach has a number of potential advantages over the frequentist approach. In particular, it will take into account all the uncertainty in all parameters during the parameter estimation, and therefore the uncertainty in \( \Omega \), (i.e. the between-study parameters, including \( \hat{\rho}_B \) - see equation (4.1)) will be taken into account when estimating \( \beta \). This is not the case for the frequentist estimation as described in Section 4.1, where the \( \tilde{\tau}_1^2, \tilde{\tau}_2^2 \), and \( \hat{\rho}_B \) from RIGLS are assumed known at each iteration. Thus, the \( \text{var}(\hat{\beta}_j) \) will be more suitable, and generally slightly larger, from a Bayesian approach to Model A. It is worth noting that it may also be possible to acknowledge the uncertainty in \( \tilde{\tau}_1^2, \tilde{\tau}_2^2 \), and \( \hat{\rho}_B \) in a frequentist estimation approach different to that described in Section 4.1. Hardy and Thompson use a profile likelihood approach that takes
into account the uncertainty in $\hat{\tau}_1^2$ and $\hat{\tau}_2^2$ in a URMA [45], and it would make interesting further research to extend this approach to Model A and to allow the uncertainty in $\hat{\rho}_B$ to be acknowledged alongside the uncertainty in $\hat{\tau}_1^2$ and $\hat{\tau}_2^2$.

One of the main issues arising from fitting Model A classically is that $\hat{\rho}_B$ is often equal to either 1 or -1, i.e. at the extreme end of the possible range for this parameter. In particular, the data sometimes provides such sparse information about the value of $\rho_B$ that 1 or -1 is obtained simply because the profile likelihood is so flat (this relates to the proposed Reason (II) in Section 5.6.2). What would be desirable is to incorporate, where available, further information about $\rho_B$ alongside the data to enable $\hat{\rho}_B$ to be better defined.

External information about $\rho_B$ may be available from previous studies, expert opinion or from related analyses. In this situation a Bayesian approach is particularly suitable as it can easily incorporate the related information alongside the data in the form of a prior distribution for $\rho_B$. For example, expert knowledge may indicate that $\rho_B$ is positive and so a uniform(0,1) prior distribution could be appropriately specified for $\rho_B$.

7.2 Formulating the Bayesian approach to Model A in WinBUGS

To fit Model A within a Bayesian framework it is necessary to specify the likelihood for the data and the prior distributions for the unknown parameters to be estimated (see Section 1.8). For a Bayesian approach to Model A, the main problem may be the specification of a prior distribution for the covariance matrix $\mathbf{\Omega}_2$. One possible prior distribution for $\mathbf{\Omega}_2^{-1}$ is the Wishart distribution, a multivariate extension of the $\chi^2$-distribution, and this is commonly used for covariance matrices in the examples within the WinBUGS manual and elsewhere [71;184;185]. However, this is a very unintuitive distribution to understand, and I find it particularly hard to translate informative prior
beliefs into parameter values within this distribution. Another alternative option is to use an approach called the product normal formulation, which re-parameterises a bivariate normal model and allows one to place prior distributions on the individual parameters within $\Omega_2$, including the between-study correlation parameter [71]. However, the model can also be formulated as a variance components model and this enables univariate prior distributions to be specified. For example, to keep things simple, note that there is no need to specify the $\theta_j$s within Model A if one does not need to estimate them, and Model A can be rewritten in a Bayesian framework as:

**Model A-Bayes**

\[
Y_{ij} \sim N(\beta, \mathbf{V}_i) \\
where \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \quad \mathbf{V}_i = \begin{pmatrix} s^2_{i1} + \tau_1^2 & \lambda_i + \tau_1 \tau_2 \rho_B \\ \lambda_i + \tau_1 \tau_2 \rho_B & s^2_{i2} + \tau_1^2 \end{pmatrix}
\]

\[
i = 1 \text{ to } n \text{ studies} \\
j = \begin{cases} 1 & \text{for outcome } 1 \\ 2 & \text{for outcome } 2 \end{cases}
\]

\[
\tau_j^{-2} \sim \text{Gamma} (0.001,0.001) \\
\rho_B \sim \text{Uniform} (-1,1) \\
\beta_j \sim N(0,1000000)
\]

This formulation allows one to specify separate ‘vague’ univariate prior distributions on each of the unknown parameters in $\mathbf{V}_i$, rather than one prior distribution for the matrix $\mathbf{V}_i$ as a whole. I acknowledge that this approach is open to debate because, although separate prior distributions are specified, the parameters are not independent of each other and so the specification of a prior on one parameter will have an impact on those specified for the other parameters. Hence, this makes it particularly important to perform sensitivity analyses on Model A-Bayes for the specification of these ‘vague’ prior distributions. It is worth noting that Nam et al. use the same specification and ‘vague’ priors as in Model A-Bayes for their published article about a Bayesian approach to BRMA [147], with the
exception that, unlike in Model A-Bayes, they do not assume that the within-study correlation is known and rather also put a vague prior on this parameter (this approach will be considered in detail in Section 8.1.4). The alternative use of informative rather than vague prior distributions is also possible in equation 7.1 when some external information is available alongside the data. The specification of Model A-Bayes in WinBUGS is shown in Appendix C1. One further point to note is that Model A-Bayes still assumes that the $s_i^2$ s and $s_i^2$ s are known, and still assumes the within-study correlations (i.e. $\rho_{wi} = \lambda_i / s_i s_{i2}$) are known. A Bayesian approach could of course also take into account the uncertainty in these parameters if necessary by extending the Model A-Bayes framework.

7.3 The influence of ‘vague’ prior distributions in Model A-Bayes

7.3.1 Application of Model A-Bayes to the Berkey dataset

Using the ‘vague’ prior distributions as specified in equation (7.1), I applied Model A-Bayes to the original Berkey dataset, which was introduced in Table 4.1 of Section 4.7. I also repeated my analyses using other prior distributions for $\rho_B$. In all my analyses I used a burn-in of 10000 samples and then took another 10000 samples in order to describe the posterior distribution of the parameters of interest (e.g. mean and variance of $\tilde{\beta}_j$).

However, there was not clear evidence that the Gibbs sampler has converged for the burn-in used (Figure 7.1), so for safety a range of longer burn-ins and different starting values were also assessed; reassuringly, this investigation produced similar posterior results each time (Figure 7.2), and there did not appear to be any strong autocorrelation between samples, for any of the analyses (Figure 7.3).
Figure 7.1: The last 200 of the 10000 samples taken from the MCMC chain after the 10000 burn-in when applying Model A-Bayes to the Berkey data, using the uniform(0,1) prior distribution for $\rho_B$ (denoted 'rhob' below) and other prior distributions as in equation (7.1). Convergence is a slight concern given some of the extreme values of $\tau^2_j$ (denoted 'tausq[j]') being sampled.

Figure 7.2: Posterior distributions for the parameters in Model A-Bayes formed from the 10000 samples taken from the MCMC chain after the 10000 burn-in when applying Model A-Bayes to the Berkey data, using the uniform(0,1) prior distribution for $\rho_B$ (denoted 'rhob' below) and other prior distributions as in equation (7.1); $\tau^2_j$ is denoted 'tausq[j]'; y-axis = density, x-axis = parameter value

Figure 7.3: Autocorrelation (y-axis) between successive parameter samples for different lags (x-axis) during the 10000 samples taken from the MCMC chain after a 10000 burn-in was used when applying Model A-Bayes to the Berkey data, using the uniform(0,1) prior distribution for $\rho_B$ (denoted 'rhob' below) and other prior distributions as in equation (7.1).
The results from Model A-Bayes, for each of the various $\rho_b$ prior distributions, are shown in Table 7.1 alongside the frequentist Model A results. The mean of the posterior distribution for $\tilde{\beta}_1$ and $\tilde{\beta}_2$ from Model A-Bayes are both very similar to the RIGLS estimates from the frequentist Model A. The uncertainty of $\tilde{\beta}_1$ and $\tilde{\beta}_2$ is slightly larger in the Bayesian analysis, reflecting the incorporation of all parameter uncertainty, and the fact that $\tilde{\tau}_1^2$ and $\tilde{\tau}_2^2$ are slightly higher in the Bayesian model. The increase in $\tilde{\tau}_1^2$ and $\tilde{\tau}_2^2$ is most likely due to the influence of the prior distribution for these parameters, even though it is a seemingly 'vague' distribution. As the values of $\tilde{\tau}_1^2$ and $\tilde{\tau}_2^2$ from the likelihood are small (i.e. the frequentist RIGLS estimates are small) and the sample size only five, the upper ends of the prior distributions for $\tau_1^2$ and $\tau_2^2$ (or the lower ends of the prior distributions for $1/\tau_1^2$ and $1/\tau_2^2$) are allowed to have an influence and therefore push the posterior mean of $\tilde{\tau}_1^2$ and $\tilde{\tau}_2^2$ slightly upwards. I also tried a uniform(0,10) prior on $\tau_j$ which has commonly been used elsewhere but this had an even stronger influence on the posterior results [67].

The prior distribution for $\rho_b$ is also extremely influential on the posterior mean of $\tilde{\rho}_b$. For example, one can see the flatness of the posterior distribution for $\tilde{\rho}_b$ between 0 and 1 where a uniform(0,1) prior was used (Figure 7.2). Even the vague uniform(-1,1) prior has a large influence and appears to pull down the posterior mean of $\tilde{\rho}_b$ toward zero (Table 7.1). This is perhaps not surprising given that only five studies were involved in the meta-analysis.
Table 7.1: Frequentist and Bayesian univariate (URMA) and bivariate (Model A, Model A-Bayes) random-effects meta-analysis results for the original Berkey data (see Table 4.1). In the Bayesian analysis a Gamma(0.001, 0.001) prior for $\tau_1^2$ and $\tau_2^2$, a N(0, 1000000) prior for $\beta_1$ and $\beta_2$, and a range of prior distributions for $\rho_B$ were used as indicated. A 10000 burn-in was used and a 10000 sample taken; convergence was assessed and considered adequate in each case (see Section 7.3.1).

<table>
<thead>
<tr>
<th>Frequentist (F) or Bayesian (B) approach</th>
<th>Model (prior for $\rho_B$)</th>
<th>Probing Depth</th>
<th>Attachment Level</th>
<th>$\rho_B$</th>
<th>$\Delta^\beta$ $\Delta^\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>URMA</td>
<td>0.361 (0.059)</td>
<td>0.0119</td>
<td>-0.346 (0.089)</td>
<td>0.0331</td>
</tr>
<tr>
<td>F</td>
<td>Model A</td>
<td>0.353 (0.059)</td>
<td>0.0117</td>
<td>-0.339 (0.088)</td>
<td>0.0327</td>
</tr>
<tr>
<td>B</td>
<td>URMA</td>
<td>0.361 (0.084)</td>
<td>0.0123</td>
<td>-0.346 (0.124)</td>
<td>0.0408</td>
</tr>
<tr>
<td>B</td>
<td>Model A-Bayes $\rho_B$ - Uniform(-1,1)</td>
<td>0.356 (0.079)</td>
<td>0.0124</td>
<td>-0.347 (0.123)</td>
<td>0.0426</td>
</tr>
<tr>
<td>B</td>
<td>Model A-Bayes $\rho_B$ - Uniform(0,1)</td>
<td>0.354 (0.075)</td>
<td>0.0117</td>
<td>-0.345 (0.118)</td>
<td>0.0399</td>
</tr>
<tr>
<td>B</td>
<td>Model A-Bayes $\rho_B$ - Uniform(0.5,1)</td>
<td>0.350 (0.082)</td>
<td>0.0114</td>
<td>-0.341 (0.122)</td>
<td>0.0435</td>
</tr>
</tbody>
</table>

N.B. the frequentist estimates are from RIGLS; the Bayesian estimates are the mean of the posterior distribution for $\beta_1$ and $\beta_2$, and the median of the posterior distribution for $\tau_1^2$, $\tau_2^2$ and $\rho_B$. 

R.D. Riley, Ph.D. Thesis  Chapter 7 228
These results therefore clearly indicate that one must be extremely cautious about using ‘vague’ prior distributions for the between-study parameters in Model A-Bayes as they may actually have a large influence on the posterior results. In particular the prior for \( \rho_s \) appears very influential, which is not surprising given how poorly defined the RIGLS estimate of \( \rho_s \) was in small samples (see Table 5.2 in Section 5.3.1) and how flat its profile likelihood can be (see Figure 5.4 in Section 5.6.2). The posterior distribution of \( \tilde{\tau}_j^2 \) and therefore the precision of \( \tilde{\beta}_j \) are clearly affected by the choice of prior for \( \rho_s \).

Furthermore, it has a particularly strong impact on \( \text{corr}(\tilde{\beta}_1, \tilde{\beta}_2) \) and therefore also on the precision of \( (\tilde{\beta}_1 - \tilde{\beta}_2) \), which can be seen to change markedly for the different prior distributions used (Table 7.1).

### 7.3.2 Application of Model A-Bayes to the modified Berkey dataset

As for the frequentist investigation of Model A, I also applied Model A-Bayes to the ‘modified’ Berkey data, a hypothetical dataset introduced in Table 4.1 of Section 4.7 which contains values of 0.1 for all the \( s_n \)s and zero within-study correlation (\( \rho_{w_s} \)) in each study.

For application of Model A to this dataset, the frequentist RIGLS analytic solutions showed that the \( j = 1 \) parameter estimates will be independent of the \( j = 2 \) parameters, and therefore the Model A results will be identical to the URMA results for \( j = 1 \) (see results in Table 7.2, and Remark 4 in Section 4.8.4). However, when Model A-Bayes was applied to this dataset, the \( j = 1 \) results are similar but are certainly not identical to the Bayesian URMA results and they are dependent on the prior for \( \rho_s \) (Table 7.2). This is again a consequence of the prior distributions used. The prior distributions for the between-study variance parameters were ‘vague’ and specified independently but they are still inherently related due to there being a prior for \( \rho_s \). This prior dependency causes the posterior estimates of the \( j = 1 \) parameters to be dependent on the \( j = 2 \) parameters, even though the
likelihood (from which the frequentist RIGLS estimates are derived) supports the notion that they are independent. This is another example of how the specification of seemingly 'vague' prior distributions can have much more of an influence on the parameter estimates than one may think. This is especially true when there is relatively little data combined with a complex hierarchical model structure [67].

**Table 7.2:** Frequentist and Bayesian univariate (URMA) and bivariate (Model A, Model A-Bayes) random-effects meta-analysis results for the modified Berkey data (see Table 4.1). In the Bayesian analyses a Gamma(0.001,0.001) prior was used for $\tau_{i1}^2$ and $\tau_{i2}^2$, a N(0, 1000000) prior for $\beta_i$ and $\beta_j$, and a range of prior distributions for $\rho_B$ as indicated. A 10000 burn-in was used and a 10000 sample taken; convergence was assessed and considered adequate in each case.

<table>
<thead>
<tr>
<th>Frequentist (F) or Bayesian (B) approach</th>
<th>Model (prior for $\rho_B$)</th>
<th>Probing Depth</th>
<th>Attachment Level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\beta}_1$ (s.e.)</td>
<td>$\tau_{i1}^2$</td>
<td>$\hat{\beta}_2$ (s.e.)</td>
</tr>
<tr>
<td><strong>F</strong> URMA</td>
<td>0.378 (0.066)</td>
<td>0.0119</td>
<td>-0.346 (0.089)</td>
</tr>
<tr>
<td><strong>F</strong> Model A</td>
<td>0.378 (0.066)</td>
<td>0.0119</td>
<td>-0.325 (0.088)</td>
</tr>
<tr>
<td><strong>B</strong> URMA</td>
<td>0.379 (0.085)</td>
<td>0.0106</td>
<td>-0.350 (0.120)</td>
</tr>
<tr>
<td><strong>B</strong> Model A-Bayes ($\rho_B$ ~Uniform(-1,1))</td>
<td>0.376 (0.083)</td>
<td>0.0107</td>
<td>-0.339 (0.124)</td>
</tr>
<tr>
<td><strong>B</strong> Model A-Bayes ($\rho_B$ ~Uniform(0,1))</td>
<td>0.377 (0.084)</td>
<td>0.0111</td>
<td>-0.334 (0.120)</td>
</tr>
<tr>
<td><strong>B</strong> Model A-Bayes ($\rho_B$ ~Uniform(0.5,1))</td>
<td>0.377 (0.085)</td>
<td>0.0115</td>
<td>-0.331 (0.124)</td>
</tr>
</tbody>
</table>

N.B. the frequentist estimates are from RIGLS; the Bayesian estimates are the mean of the posterior distribution for $\hat{\beta}_j$, and the median of the posterior distribution for $\tau_{i1}^2$, $\tau_{i2}^2$ and $\hat{\rho}_B$.

### 7.3.3 The dangers of using 'vague' prior distributions in Model A-Bayes

A Bayesian approach was recently highlighted by Nam et al. [147], and in their paper the authors used a vague uniform(-1,1) prior distribution for $\rho_B$. They also assess the impact of the $\rho_B$ prior distribution by using a range of other different prior distributions, similar to my approach in Sections 7.3.1 and 7.3.2. They conclude that the impact of the choice of prior distribution on the pooled estimates is negligible for their dataset, but they also state...
that this finding may not be generalisable. Indeed, given the findings of Chapter 4, this is clearly not a generalisable result because the RIGLS analytic solutions from Section 4.3 clearly show that $\hat{\beta}_j$ and $\text{var}(\hat{\beta}_j)$ are dependent on $\hat{\rho}_B$; for example, the gain in precision for $\hat{\beta}_j$ by using Model A instead of a URMA is dependent on the size of $\hat{\rho}_B$ (see Point (2) in Section 4.5). The relationships between parameters that were exposed in the frequentist analytic solutions also exist within the Model A-Bayes framework because the Bayesian approach also incorporates the likelihood, from which the frequentist RIGLS solutions are derived. Hence, prior distributions which influence the value of $\hat{\rho}_B$ could clearly influence the pooled estimates and their precision. This can be seen for the Berkey results in Table 7.1 where, even though I found the individual pooled estimates were similar for all my analyses, the standard error of $\hat{\beta}_j$ and also $(\hat{\beta}_1 - \hat{\beta}_2)$ are clearly dependent on the prior distribution for $\rho_B$. The impact of prior distributions may be even more marked when some data is MAR because in this situation the size of $\hat{\rho}_B$ dictates how far $\hat{\beta}_j$ shrinks away from the URMA analysis (see Sections 6.1 and 6.4.2).

Caution is therefore clearly needed when performing a multivariate Bayesian meta-analysis because the prior distributions can easily dictate the final conclusions. Hence, I strongly recommend that if Model A-Bayes is used, and one does not have any strong prior knowledge about the between-study parameters, a sensitivity analysis using a range of different prior distributions is essential to assess how the results change.

### 7.4 The tension between accounting for uncertainty and the influence of prior distributions

An appealing property of Model A-Bayes is that it inherently incorporates the uncertainty in all the parameter estimates, including the between-study parameters. This is particularly appealing because the between-study parameter estimates are often very poorly defined
Unfortunately, it is also in this situation where the prior distributions on these parameters will most likely be influential, even those considered 'vague' as in equation (7.1) (this was seen in the Berkey examples of Section 7.3). Hence, a tension exists when considering a Bayesian approach between wanting to acknowledge the overall uncertainty and not wanting to influence the results through the choice of prior distribution. Ironically, where the prior distributions have smallest impact is also where the between-study parameters are better defined and it is less important to acknowledge the uncertainty in this situation.

7.5 The benefit of incorporating external information in Model A-Bayes

For the analyses of the Berkey datasets in Section 7.3 I showed that all the prior distributions for $\rho_B$ were at least slightly informative. Although this is unwelcome where one desires ‘vague’ and non-informative prior distributions (Section 7.4), one advantage of this added information is that it prevents the MCMC sampler from drawing a value of $\tilde{\rho}_B$ equal to 1 or -1; for example, none of the 10000 samples drawn during any of the Berkey analyses gave a $\tilde{\rho}_B$ equal to 1 or -1. The prior distributions for $\tilde{\rho}_B$ give zero probability of $\tilde{\rho}_B$ being outside the range -1 to 1, and this means that, whenever the likelihood suggests a value of 1 or -1 for $\hat{\rho}_B$, the prior distribution influences the posterior value of $\tilde{\rho}_B$ away from 1 or -1 toward zero. This is clearly advantageous given the problems of bias that were shown to arise when $\rho_B$ is estimated as 1 or -1 using RIGLS in the frequentist Model A (see Section 5.6).

Therefore, although I would advise caution about using Model A-Bayes where one has little or no prior information, where one has strong prior knowledge about $\rho_B$ (or indeed $\tau_j^*$ or $\bar{\beta}_j$), Model A-Bayes is very appealing. Indeed, when prior information is available it is clearly very important to incorporate this alongside the data in order to obtain pooled
estimates and evidence-based conclusions which are appropriately based on all the
evidence available. To demonstrate the benefit of prior information in Model A-Bayes I
will now consider the following four datasets:

(i) One of the \( n = 50 \) simulated datasets (id no. = 33) for complete-case data from
Section 5.2 for which Model A estimated \( \rho_B \) as 1.

(ii) One of the \( n = 50 \) simulated missing completely at random (MCAR) datasets from
Section 6.4.1 where \( Y_{il} \) was available for all studies but \( Y_{l2} \) was MCAR for 25
studies, and for which Model A estimated \( \rho_B \) as 1 (id no. = 221).

(iii) One of the \( n = 50 \) simulated missing at random (MAR) datasets (id no. = 48) in
Section 6.4.2 where \( Y_{l2} \) was available for all studies but \( Y_{il} \) was only available
when it was positively signed, and for which Model A estimated \( \rho_B \) as 1.

(iv) One of the \( n = 50 \) simulated MAR datasets (id no. = 7) in Chapter 6.4.2 where \( Y_{l2} \)
was available for all studies but \( Y_{il} \) was only available when it was positively
signed, and for which Model A estimated \( \rho_B \) as -1.

All of these datasets were generated from Model A using true parameter values of \( \beta_1 = 0, \beta_2 = 2, \tau_1^2 = 0.25, \tau_2^2 = 0.25, \rho_B = 0.8 \) and all the \( \rho_{wi}, s = 0.8 \) (for details on exactly how
they were generated see Section 5.2.1). Both a Bayesian URMA (as described in Section
1.8.3) and Model A-Bayes were applied to all of these datasets using the prior distributions
specified as in equation (1.13) and equation (7.1) respectively, except a uniform(0.6, 0.9)
prior was used for \( \rho_B \) in Model A-Bayes because (hypothetical) prior information
indicates that \( \rho_B \) is between 0.6 and 0.9. All the datasets involved 50 studies and this
allowed (even in the datasets containing some missing data) the ‘vague’ prior distributions
for \( \tau_j^2 \) to truly have a minimal influence on the posterior results. Indeed, for all my
analyses I tried changing the starting values and the 'vague' prior distributions for $\tau_j^2$ but this made little difference to the Bayesian URMA and Model A-Bayes results in each case. This is important as it enables the impact of the informative prior distribution for $\rho_B$ to be assessed more clearly from the results. Before discussing the results for each dataset, one general finding is that the Bayesian analyses generally produce slightly larger standard errors of the parameter estimates than the equivalent frequentist analyses; this is due to the Bayesian approach incorporating the uncertainty in $\tilde{\tau}_j^2$ and $\hat{\rho}_B$ when estimating $\beta_j$, and vice-versa, something that does not occur in the frequentist RIGLS estimation of Model A.

7.5.1 Results for the complete-case dataset

Model A was shown to be beneficial for complete-case data when estimating $(\beta_i - \beta_3)$ but for this particular complete-case dataset (dataset (i)) it is perhaps inadvisable to use the Model A results because $\rho_B$ is estimated as 1 (Table 7.3). However, the Model A-Bayes results are not subject to this problem because the prior information restricts the posterior estimate of $\rho_B$ to within the range 0.6 to 0.9, and a posterior mean for $\hat{\rho}_B$ of 0.842 is obtained, which is closer to the true value of $\rho_B$ (Table 7.3). As is true for the frequentist approach, one can see that Model A-Bayes is clearly beneficial over a URMA Bayesian analysis for estimating $(\beta_i - \beta_3)$ as $\text{var}(\hat{\beta}_i - \hat{\beta}_3)$ is appropriately smaller in Model A-Bayes, as it incorporates $\text{corr}(\hat{\beta}_i, \hat{\beta}_3)$. The posterior median value of $\tilde{\tau}_j^2$ is also much smaller and closer to the true value of 0.25 than that from the Bayesian URMA, which is a major reason why Model A-Bayes also reduces $\text{var}(\hat{\beta}_j)$. Model A-Bayes has therefore allowed pooled estimates to be formed that are likely to be more appropriate than those from a URMA (because $\text{corr}(\hat{\beta}_i, \hat{\beta}_3)$ is incorporated) and also Model A (because the prior information has removed the problem of $\hat{\rho}_B$ being equal to 1).
Table 7.3: Frequentist and Bayesian univariate (URMA) and bivariate (Model A, Model A-Bayes) random-effects meta-analysis results for the three \( n = 50 \) datasets described in Section 7.5, for which the true parameter values are \( \beta_1 = 0, \beta_2 = 2, \tau_1^2 = 0.25, \tau_2^2 = 0.25, \rho_B = 0.8 \) and all the \( \rho_{wi} = 0.8 \). In the Model A-Bayes analyses a Gamma(0.001,0.001) prior was used for \( \tau_1^2 \) and \( \tau_2^2 \), a N(0, 1000000) prior for \( \beta_1 \) and \( \beta_2 \), and a uniform(0.6,0.9) prior was used for \( \rho_B \). A 50000 burn-in was used and a 10000 sample taken; convergence was assessed and considered adequate in each case. N.B. the frequentist estimates are from RIGLS.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Frequentist (F) or Bayesian (B)</th>
<th>Model</th>
<th>( \hat{\beta}_1 ) (s.e.)</th>
<th>( \tau_1^2 )</th>
<th>( \hat{\beta}_2 ) (s.e.)</th>
<th>( \tau_2^2 )</th>
<th>( \hat{\rho}_B ) (s.e.)</th>
<th>( \hat{\beta}_1 - \hat{\beta}_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i)</td>
<td>F URMA</td>
<td>( 0.0128 ) (0.120)</td>
<td>( 0.401 ) (0.129)</td>
<td>1.996 (0.451)</td>
<td>-</td>
<td>-1.883 (0.176)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Model A</td>
<td>( 0.0177 ) (0.0942)</td>
<td>( 0.269 ) (0.106)</td>
<td>1.947 (0.357)</td>
<td>1</td>
<td>-1.929 (0.0545)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>URMA</td>
<td>( 0.0120 ) (0.120)</td>
<td>( 0.393 ) (0.131)</td>
<td>1.895 (0.441)</td>
<td>-</td>
<td>-1.883 (0.178)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Model A-Bayes</td>
<td>( 0.0243 ) (0.0983)</td>
<td>( 0.235 ) (0.105)</td>
<td>1.928 (0.267)</td>
<td>0.842 (0.0725)</td>
<td>-1.904</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(ii)</td>
<td>F URMA</td>
<td>( -0.127 ) (0.0883)</td>
<td>( 0.145 ) (0.129)</td>
<td>1.878 (0.145)</td>
<td>-</td>
<td>-2.077 (0.137)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Model A</td>
<td>( -0.113 ) (0.0856)</td>
<td>( 0.162 ) (0.0979)</td>
<td>1.978 (0.152)</td>
<td>1</td>
<td>-2.091 (0.0653)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>URMA</td>
<td>( -0.122 ) (0.0901)</td>
<td>( 0.134 ) (0.124)</td>
<td>1.874 (0.100)</td>
<td>-</td>
<td>-1.996 (0.152)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Model A-Bayes</td>
<td>( -0.131 ) (0.0902)</td>
<td>( 0.215 ) (0.105)</td>
<td>1.944 (0.141)</td>
<td>0.804 (0.0903)</td>
<td>-2.075</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(iii)</td>
<td>F URMA</td>
<td>( 0.562 ) (0.092)</td>
<td>( 0.145 ) (0.119)</td>
<td>1.988 (0.379)</td>
<td>-</td>
<td>-1.426 (0.150)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Model A</td>
<td>( 0.216 ) (0.120)</td>
<td>( 0.211 ) (0.120)</td>
<td>2.006 (0.430)</td>
<td>1</td>
<td>-1.790 (0.124)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>URMA</td>
<td>( 0.572 ) (0.109)</td>
<td>( 0.0174 ) (0.128)</td>
<td>1.985 (0.409)</td>
<td>-</td>
<td>-1.413 (0.168)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Model A-Bayes</td>
<td>( 0.345 ) (0.129)</td>
<td>( 0.0690 ) (0.128)</td>
<td>1.854 (0.574)</td>
<td>0.774 (0.151)</td>
<td>-1.509</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(iv)</td>
<td>F URMA</td>
<td>( 0.374 ) (0.0685)</td>
<td>( 0.0510 ) (0.110)</td>
<td>1.933 (0.296)</td>
<td>-</td>
<td>-1.559 (0.130)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Model A</td>
<td>( 0.229 ) (0.0656)</td>
<td>( 0.0101 ) (0.102)</td>
<td>1.893 (0.243)</td>
<td>-1</td>
<td>-1.664 (0.121)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>URMA</td>
<td>( 0.322 ) (0.106)</td>
<td>( 0.0253 ) (0.114)</td>
<td>1.931 (0.285)</td>
<td>-</td>
<td>-1.609 (0.156)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Model A-Bayes</td>
<td>( 0.0891 ) (0.0874)</td>
<td>( 0.0533 ) (0.111)</td>
<td>1.822 (0.360)</td>
<td>0.714 (0.106)</td>
<td>-1.733</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

N.B. the Bayesian estimates are the mean of the posterior distribution for \( \hat{\beta}_1 \) and \( \hat{\beta}_2 \), and the median of the posterior distribution for \( \tau_1^2 \), \( \tau_2^2 \) and \( \hat{\rho}_B \).
7.5.2 Results for the missing completely at random dataset

For the MCAR dataset (ii), the frequentist Model A results may be misleading because $\hat{\rho}_B$ is equal to 1. However, Model A-Bayes is again not subject to this problem because the prior information restricts the posterior estimate of $\rho_B$ to within the range 0.6 to 0.9, and a posterior mean for $\hat{\rho}_B$ of 0.804 is obtained, which is closer to the true value of $\rho_B$ (Table 7.3). Of key interest is how the precision of $\hat{\beta}_2$ changes from the Bayesian URMA to Model A-Bayes because outcome $j = 2$ is missing for 25 studies. The results show Model A-Bayes makes a considerable gain in precision for $\hat{\beta}_2$, and also $(\beta_i - \beta_j)$. Such gains are observed despite the posterior median value of $\hat{\tau}_j^2$ in Model A-Bayes being much larger and closer to the true value of 0.25 than the equivalent Bayesian URMA estimates (Table 7.3). Model A-Bayes has therefore allowed pooled estimates to be formed that are likely to be more appropriate than those from a URMA (because $\text{corr}(\hat{\beta}_1, \hat{\beta}_2)$ is incorporated and MCAR is known to be true) and also Model A (because the prior information has removed the problem of $\hat{\rho}_B$ being equal to 1).

7.5.3 Results for the missing at random dataset

For the analyses of the MAR dataset (iii), of key interest is the value of $\hat{\beta}_1$ in Model A and Model A-Bayes because the frequentist and Bayesian URMA results will be very misleading due to the extreme data MAR in 27 studies for outcome $j = 1$. The frequentist Model A shrinks the URMA $\hat{\beta}_1$ value (of 0.562) to 0.216, much closer to the true value of zero for $\hat{\beta}_2$ (Table 7.3). However, these results are perhaps inadvisable to use because $\hat{\rho}_B$ is equal to 1 and therefore the results are potentially subject to bias. However, Model A-Bayes is not subject to this problem because the prior information restricts the posterior estimate of $\rho_B$ to within the range 0.6 to 0.9, and a posterior mean for $\hat{\rho}_B$ of 0.774 is...
obtained, which is closer to the true value of $\rho_B$. The Model A-Bayes results also shrink the Bayesian URMA value of $\beta_1$ down toward zero, as the mean posterior estimate from Model A-Bayes is equal to 0.345 compared to 0.572 from the Bayesian URMA.

Interestingly the shrinkage from Model A-Bayes is not as far as Model A, which may partly be due to the smaller and more sensible $\hat{\rho}_B$ in Model A-Bayes having less influence than the $\hat{\rho}_B$ value of 1 in Model A. Similarly, the substantially smaller $\hat{\tau}_i^2$ from Model A-Bayes may also be preventing as much shrinkage.

Interestingly in this particular example, apart from $\hat{\rho}_B$, Model A-Bayes does not provide parameter estimates as close to the true values as those from Model A. To assess if this was due to prior distributions I performed a number of analyses using Model A-Bayes with different ‘vague’ prior distributions for $\tau^2_i$ and various starting values; however, they all produced very similar results to those presented in Table 7.3. Hence it appears that, for this particular dataset, the bias in Model A when $\hat{\rho}_B$ equals 1 is actually causing parameter estimates to move closer to the true parameter values than in Model A-Bayes, where $\hat{\rho}_B$ equal to 1 is not a concern. This is perhaps surprising but does not always occur, and indeed in other datasets the bias in Model A will cause more misleading answers than Model A-Bayes. To demonstrate this I present the results in Table 7.4 for a similar MAR dataset (dataset (iv)) but for which Model A estimates $\rho_B$ as -1. This time Model A-Bayes produces a mean posterior value of 0.0891 for $\beta_1$, far closer to the true value of zero than the RIGLS $\beta_1$ value of 0.229 from Model A.

The key message from these two analyses is that bias in the frequentist Model A influences the parameter estimates, for example in how much they are shrunk away from the frequentist URMA results. This point was discussed in Section 5.6.4. However, Model A-
Bayes is not subject to this problem as the prior information about $\rho_B$ moves the posterior distribution of $\tilde{\rho}_B$ well away from 1 or -1. These two MAR datasets (dataset (iii) and (iv)) therefore illustrate that sometimes the bias in Model A produces estimates closer to the true parameter values than those in Model A-Bayes, but also sometimes they are further away from the true values than Model A-Bayes. Importantly, in real life situations one would not know what the true parameter values are, and thus in reality it would be perhaps inadvisable to use any biased Model A results in this situation. This section has shown that Model A-Bayes incorporating strong prior information, with an assessment of any 'vague' prior distributions as necessary, has the potential to be more suitable than the frequentist Model A in some situations, especially where prior information is available for $\rho_B$.

7.5.4 The difficulty of eliciting strong external information

Although prior information is clearly desirable, it may be practically very difficult to elicit strong external information about $\rho_B$ or $\tau_j^2$. For example, where similar datasets or analyses exist it may be difficult to differentiate the between-study correlation from the overall correlation[186], which is a weighted average of the within- and between-study correlations. Indeed, it is potentially easier to elicit a measure of the overall correlation between the $\tilde{Y}_{ij}$s and the $\tilde{Y}_{ij}$s (the observed summary statistic estimates) than it is to elicit the correlation between the $\theta_{ij}$s and $\theta_{ij}$s (the true underlying summary statistics), which are unknown and would perhaps first need to be estimated by using the $\tilde{Y}_y$ and the $s_y$ (see Section 4.6). For similar reasons, asking experts in the field for their opinions of the likely value of the between-study correlation may also be difficult [187;188].
7.6 Summary and rationale for subsequent chapters

This chapter has introduced Model A-Bayes, a Bayesian approach to BRMA that has the potential to overcome some of the estimation problems associated with the frequentist Model A. In particular, inclusion of prior information for $\rho_B$ prevents $\hat{\rho}_B$ being equal to either 1 or -1 in Model A-Bayes, which is a common problem in Model A that subsequently introduces bias in $\hat{\tau}_j^2$ and also potentially $\hat{\beta}_j$. However, Model A-Bayes will often not be straightforward because elicitation of prior information may be extremely difficult, and one needs to be cautious about using 'vague' prior distributions, especially for $\rho_B$ (see Section 7.6). Where no strong prior information is available and Model A-Bayes is used, I recommend sensitivity analyses to assess the robustness of the results to changes in the 'vague' prior distributions specified, something that of course should be an essential part of any Bayesian analysis [67].

Although I have shown that the Bayesian approach, through the incorporation of prior information, can help address the estimation problem of $\rho_B$, there still remains a major methodological problem for both Model A-Bayes and Model A: how does one proceed with these approaches when the within-study correlations (i.e. the $\rho_{w_i}$s) are unavailable? The next chapter addresses this question directly and considers possible methods to solve the problem, including both Bayesian and frequentist approaches. Indeed, as unavailable $\rho_{w_i}$s will be the common situation for most meta-analysts in practice, the need to address this issue is a pressing concern as otherwise the benefits of Model A and Model A-Bayes demonstrated in this thesis may often remain unattainable. This is of particular importance if BRMA is to be beneficial for evidence syntheses of prognostic marker studies, and indeed if the approach is to be made applicable to the OS and DFS HRs from the neuroblastoma review of Chapter 2.
Chapter 8

BIVARIATE RANDOM-EFFECTS META-ANALYSIS WHEN THE WITHIN-STUDY CORRELATIONS ARE UNKNOWN

Chapter overview

In this chapter I will firstly review and apply the main approaches suggested in the current meta-analysis literature for dealing with unknown within-study correlations in Model A and Model A-Bayes. Secondly, I will introduce a new alternative parameterisation for bivariate random-effects meta-analysis (BRMA), which does not require the within-study correlations ($\rho_w$), and evaluate this against Model A and a univariate random-effects meta-analysis (URMA). To help apply and compare the various methods I introduce and discuss in this chapter, I will often use the MYCN dataset from the neuroblastoma review (see Table 3.1 of Section 3.2). Model A is particularly attractive for the MYCN dataset because it would allow the relatively large positive correlation between OS and DFS to be utilised (see Figure 3.3 in Section 3.2). However, Model A has not yet been applicable for MYCN because the within-study correlations (i.e. the $\rho_{ws}$) between the OS and DFS HRs are unknown and are highly unlikely to be zero due to the summary statistics being structurally related and estimated from the same set of patients (see Section 3.6.2). I aim to specifically address this problem in this chapter and thereby facilitate the use of BRMA within future evidence syntheses of prognostic marker studies. This is particularly important because, given the findings about Model A in previous chapters (e.g. the benefits for estimating $\beta_j$ given missing data; see Section 6.5), the BRMA approach may help to limit the problems of missing outcomes and missing summary statistics that commonly affects evidence synthesis for prognostic markers (see Section 3.2 and Figure 3.2).
8.1 Options suggested by the current literature for applying Model A or Model A-Bayes when the within-study correlations are unknown

8.1.1 Option 1: apply Model A assuming $\rho_{wi} = 0$ for all studies

There are some situations where one may plausibly assume the $\rho_{wi}$s are zero (see Section 3.6.2) and this assumption removes the need to obtain this statistic, which is why it has been commonly used in the current literature where appropriate [148;161]. However, what are the implications of assuming all the $\rho_{wi}$s are zero when this is not true, such as in the MYCN dataset? In particular, is Model A still beneficial (e.g. in terms of precision and MSE) over a URMA if one takes this approach? To understand this, I repeated the $n = 50$ and $n = 5$ simulations for complete-case data (which were described in Sections 5.2 and 5.3), and also the $n = 50$ and $n = 10$ simulations for missing data (see Sections 6.4.1 and 6.4.2) for each of the following settings:

Setting (i): $\rho_{wi} = 0.8$ for all studies; $\rho_{B} = 0.8$; $\tau_j^2 = 1.5$, on average larger than the $s_j^2$s

Setting (ii): $\rho_{wi} = 0.8$ for all studies; $\rho_{B} = 0.8$; $\tau_j^* = 0.25$, on average similar to the $s_j^2$s

Setting (iii): $\rho_{wi} = 0.8$ for all studies; $\rho_{B} = 0.8$; $\tau_j^* = 0.0025$ on average smaller than the $s_j^2$s

However, this time, alongside application of (a) two independent URMAs, and (b) Model A, I also applied (c) Model A again but where $\rho_{wi}$ was assumed zero in each study (denoted ‘Model A-zero’). The suitability of Model A-zero was then assessed by comparing the results from (c) with those from (a) and (b) in each of (i) to (iii).

Complete-case data simulation results for Model A-zero

In terms of estimating $\beta_1$ and $\beta_2$, I showed in Chapter 5 that the benefit of Model-A over and above two independent URMAs is negligible given complete-case data, even when the $\rho_{wi}$s are known. Model A-zero most often reduces this negligible benefit entirely (Table...
8.1 and Appendix C2). For example, consider \( n = 50 \) and \( n = 5 \) in Setting (ii), where \( \tau_j^2 \) is on average similar in size to the \( s_{ij}^2 \) s (Table 8.1). Model A-zero increases the problem observed in Model A of \( \rho_w \) being estimated as 1, because the model is trying to compensate for the understated \( \rho_{wi} \) values (see Section 5.6.1). This causes the \( \hat{\tau}_j^2 \) s to again be biased upward on average (up to 20% above the true value of \( \tau_j^2 \)) and the coverage of \( \hat{\beta}_j \) to be well above 95%, with both these problems worse than in Model A.

These problems only improve slightly for 50 compared to 5 studies.

In terms of estimating \( (\beta_1 - \beta_2) \), I showed in Section 6.6 that there is considerable benefit of using Model A over two independent URMA s, even for complete-case data; the benefit of Model A-zero for estimating \( (\beta_1 - \beta_2) \) appears to fall in between these approaches. For example, the estimate of \( \text{corr}(\beta_1, \beta_2) \) from Model A-zero is between zero (univariate estimate) and the estimate from Model A, whilst the variance and MSE of \( (\beta_1 - \hat{\beta}_2) \) from Model A-zero are on average larger than those from Model A but smaller than those from a URMA (Table 8.1, Appendix C2).

The Model A-zero results are closest to the URMA results when \( \hat{\tau}_j^2 \) is on average smaller than the \( s_{ij}^2 \) s (the model will also often not converge in this situation, see Appendix C2), and they are closest to the Model A results where \( \tau_j^2 \) is on average larger in size to the \( s_{ij}^2 \) s. The reason for this is that when the \( s_{ij}^2 \) s are relatively small, mis-specifying \( \rho_{wi} \) has only a very small impact on the between-study parameter estimates (i.e. \( \hat{\rho}_B \) and \( \hat{\tau}_j^2 \)), which only need to increase slightly to accommodate the mis-specification of \( \rho_{wi} \).

R.D. Riley, Ph.D. Thesis  Chapter 8  242
Table 8.1: Simulation results from the univariate (URMA) and bivariate (Model A) random-effects meta-analysis for complete-case data of 50 and 5 studies for true parameter values of $\beta_1 = 0$, $\beta_2 = 2$, and $\tau_1^2 = \tau_2^2 = 0.25$ (on average similar in size to the $s_i^2$'s). The 95% confidence intervals (CIs) for the pooled estimates were calculated using a t-distribution with $(n_i - 1)$ degrees of freedom. The within-study correlation ($\rho_{wi}$) was known to be 0.8 and the same for each study, whilst the between-study correlation ($\rho_B$) was estimated. However, for Model A-zero $\rho_{wi}$ was wrongly set to zero to assess the impact of this. The true values of $\rho_{wi}$ and $\rho_B$ in each simulation are shown in the table.

<table>
<thead>
<tr>
<th>No. of studies (n)</th>
<th>Meta-analysis model</th>
<th>$\rho_{wi}$</th>
<th>$\rho_B$</th>
<th>Converged out of 1000</th>
<th>Mean of $\hat{\beta}_1$ (s.e. of mean)</th>
<th>Mean of $\hat{\beta}_2$ (s.e. of mean)</th>
<th>MSE of 95% CIs for $\hat{\beta}_1$ including $\beta_1$ (%)</th>
<th>MSE of 95% CIs for $\hat{\beta}_2$ including $\beta_2$ (%)</th>
<th>Mean of $\tau_1^2$ (no. = 0)</th>
<th>Mean of $\tau_2^2$ (no. = 0)</th>
<th>Mean of $\hat{\rho}_B$ = $\rho_B$-1 (%)</th>
<th>No. of $\hat{\rho}_B$ = $\rho_B$ = 1 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>URMA</td>
<td>0.8</td>
<td>0.8</td>
<td>1000</td>
<td>-0.0044 (0.101)</td>
<td>0.102/0.102</td>
<td>0.0102/0.0102 (94.5%)</td>
<td>1.993/1.016 (94.6%)</td>
<td>0.248/0.247 (0)</td>
<td>0.240 (0)</td>
<td>0.238 (0)</td>
<td>-</td>
</tr>
<tr>
<td>50</td>
<td>Model A</td>
<td>0.8</td>
<td>0.8</td>
<td>1000</td>
<td>-0.0031 (0.094)</td>
<td>0.095/0.095</td>
<td>0.0089/0.0095 (95.1%)</td>
<td>1.995/1.097 (93.9%)</td>
<td>0.250/0.248 (0)</td>
<td>0.240 (0)</td>
<td>0.235 (0)</td>
<td>0.786 (0%)</td>
</tr>
<tr>
<td>50</td>
<td>Model A-zero</td>
<td>0.8</td>
<td>0.8</td>
<td>1000</td>
<td>-0.0051 (0.0990)</td>
<td>0.104/0.102</td>
<td>0.0098/0.0105 (96.3%)</td>
<td>1.993/1.102 (96.5%)</td>
<td>0.299/0.304 (0)</td>
<td>0.304 (0)</td>
<td>0.999 (0)</td>
<td>0.976 (97.6%)</td>
</tr>
<tr>
<td>5</td>
<td>URMA</td>
<td>0.8</td>
<td>0.8</td>
<td>1000</td>
<td>-0.0083 (0.289)</td>
<td>0.271/0.257</td>
<td>0.0834/0.094 (97.9%)</td>
<td>1.984/1.303 (93.2%)</td>
<td>0.254/0.251 (0)</td>
<td>0.179 (84)</td>
<td>0.166 (106)</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>Model A</td>
<td>0.8</td>
<td>0.8</td>
<td>998</td>
<td>-0.0094 (0.284)</td>
<td>0.262/0.250</td>
<td>0.0805/0.0870 (97.7%)</td>
<td>1.984/1.295 (92.9%)</td>
<td>0.262/0.259 (0)</td>
<td>0.193 (8)</td>
<td>0.184 (0)</td>
<td>0.651 (12.4%)</td>
</tr>
<tr>
<td>5</td>
<td>Model A-zero</td>
<td>0.8</td>
<td>0.8</td>
<td>998</td>
<td>-0.0084 (0.287)</td>
<td>0.284/0.250</td>
<td>0.0823/0.0884 (96.7%)</td>
<td>1.984/1.297 (95.6%)</td>
<td>0.289/0.286 (0)</td>
<td>0.286 (0)</td>
<td>0.896 (4.0%)</td>
<td>850</td>
</tr>
</tbody>
</table>

N.B. $\hat{\rho}_B$ was not applicable when one or both of $\hat{\tau}_1^2$ and $\hat{\tau}_2^2$ was zero. MSE = mean-square-error, CIs = confidence intervals, s.e. = standard error

Additional results for $(\hat{\beta}_1 - \hat{\beta}_2)$:

<table>
<thead>
<tr>
<th>n</th>
<th>Mean of $(\hat{\beta}_1 - \hat{\beta}_2)$ (s.e.)</th>
<th>Mean s.e of $(\hat{\beta}_1 - \hat{\beta}_2)$</th>
<th>Mean MSE of $(\hat{\beta}_1 - \hat{\beta}_2)$</th>
<th>Mean corr$(\hat{\beta}_1, \hat{\beta}_2)$</th>
<th>n</th>
<th>Mean of $(\hat{\beta}_1 - \hat{\beta}_2)$ (s.e.)</th>
<th>Mean s.e of $(\hat{\beta}_1 - \hat{\beta}_2)$</th>
<th>Mean MSE of $(\hat{\beta}_1 - \hat{\beta}_2)$</th>
<th>Mean corr$(\hat{\beta}_1, \hat{\beta}_2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>-1.998 (0.147)</td>
<td>0.147</td>
<td>0.00707</td>
<td>-</td>
<td>URMA</td>
<td>5 -1.992 (0.382)</td>
<td>0.382</td>
<td>0.00462</td>
<td>-</td>
</tr>
<tr>
<td>50</td>
<td>-1.998 (0.0731)</td>
<td>0.0731</td>
<td>0.00562</td>
<td>0.710 (0.0753)</td>
<td>Model A</td>
<td>5 -1.994 (0.187)</td>
<td>0.187</td>
<td>0.00418</td>
<td>0.730 (0.193)</td>
</tr>
<tr>
<td>50</td>
<td>-1.998 (0.101)</td>
<td>0.101</td>
<td>0.00619</td>
<td>0.543 (0.0713)</td>
<td>Model A</td>
<td>5 -1.992 (0.262)</td>
<td>0.262</td>
<td>0.00428</td>
<td>0.488 (0.263)</td>
</tr>
</tbody>
</table>

R.D. Riley, Ph.D. Thesis Chapter 8 243
**Missing data simulation results for Model A-zero**

I showed in Section 6.4 that when the missing summary statistic is either MCAR or MAR in those studies only providing one outcome, Model A is particularly beneficial over two independent URMA$s for estimating \((\beta_1 - \beta_2)\) and also \(\beta_1\) and \(\beta_2\) themselves; in particular, there can be a considerable gain in MSE, standard error, and more suitable coverage. Considering Model A-zero now, the missing data simulation results are again comparable to Model A when the \(\tau_j^2\)s are on average larger in size to the \(s_j^2\)s (Appendix C3), but again when the \(s_j^2\)s are relatively larger than the \(\tau_j^2\)s Model A-zero will rarely converge (Appendix C3). However, when the \(s_j^2\)s are similar in size to the \(\tau_j^2\)s, Model A-zero converges much more often and its benefits fall in between those for Model A and a URMA (Table 8.2, Appendix C3). For example, for the MCAR simulations Model A-zero produces a \(\text{var}(\tilde{\beta}_2)\) and \(\text{var}(\tilde{\beta}_1 - \tilde{\beta}_2)\) larger than those from Model A but still smaller than those from the URMA on average (Table 8.2).

Even when Model A-zero does converge, the increasingly large number of occasions where \(\rho_8\) is estimated as 1 is again a major concern, and this would appear to be a major issue preventing Model A-zero from being a useful tool in practice as one needs to be cautious about using such results (see Section 5.6.4).
Table 8.2: Simulation results from the univariate (URMA) and bivariate (Model A) random-effects meta-analysis models for 50 and 10 studies, where outcome \( j = 1 \) was available from all studies but outcome \( j = 2 \) was missing completely at random (MCAR) from half the studies in each case. The true parameter values to compare the results to are \( \beta_1 = 0 \) and \( \beta_2 = 2 \), whilst \( \tau_1^2 = 0.25 \) and \( \tau_2^2 = 0.25 \) (on average similar in size to the \( s_j^2 \)s). In Model A the within-study correlation (\( \rho_{wi} \)) was known and the same for each study, whilst the between-study correlation (\( \rho_B \)) was estimated. In Model A-zero \( \rho_{wi} \) was wrongly assumed zero to assess the impact of this. The true values of \( \rho_{wi} \) and \( \rho_B \) in each simulation are shown in the table.

| No. of | Meta-analysis model | \( \rho_{wi} \) | \( \rho_B \) | Converged out of 1000 | Mean of \( \hat{\beta}_1 \) (s.e.) | Mean s.e. of \( \hat{\beta}_1 \) | No. of 95% CIs for \( \hat{\beta}_1 \) including \( \beta_1 \) (%) | Mean of \( \hat{\beta}_2 \) (s.e.) | Mean s.e. of \( \hat{\beta}_2 \) | No. of 95% CIs for \( \hat{\beta}_2 \) including \( \beta_2 \) (%) | Mean of \( \tilde{\tau}_1^2 \) (no. of \( \tilde{\tau}_1^2 = 0 \)) | Mean of \( \tilde{\tau}_2^2 \) (no. of \( \tilde{\tau}_2^2 = 0 \)) | Mean of \( \hat{\rho}_B \) = 1 (%) | No. of \( \hat{\rho}_B \) = 0 (%) |
|-------|---------------------|----------------|---------------|-----------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|----------------|----------------|----------------|
| 50    | URMA                | 0.8            | 0.8           | 1000                  | 1.995 (0.147)    | 0.146           | 0.0216          | 949 (94.9%)     | 0.253 (0)       | 0.248 (0)       | -               | -               | -               |
| 50    | Model A             | 0.8            | 0.8           | 1000                  | 1.993 (0.122)    | 0.119           | 0.0150          | 946 (94.6%)     | 0.249 (0)       | 0.252 (0)       | 0.794 (2)       | 95 (9.5%)       |
| 50    | Model A-zero        | 0.8            | 0.8           | 1000                  | 1.995 (0.128)    | 0.131           | 0.0163          | 957 (95.7%)     | 0.255 (0)       | 0.230 (0)       | 0.986 (0)       | 841 (84.1%)     |
| 10    | URMA                | 0.8            | 0.8           | 999                   | 1.994 (0.303)    | 0.262           | 0.0921          | 932 (93.2%)     | 0.250 (30)      | 0.251 (86)      | -               | -               |
| 10    | Model A             | 0.8            | 0.8           | 1000                  | 1.986 (0.226)    | 0.225           | 0.0708          | 932 (93.3%)     | 0.265 (3)       | 0.261 (0)       | 0.672 (105)     | 540 (54.1%)     |
| 10    | Model A-zero        | 0.8            | 0.8           | 1000                  | 1.986 (0.226)    | 0.251           | 0.0721          | 980 (98.0%)     | 0.273 (3)       | 0.275 (0)       | 0.902 (9.5%)    | 875 (87.5%)     |

Additional results for (\( \tilde{\beta}_1 - \tilde{\beta}_2 \)):

| n    | Meta-analysis model. | Mean of (\( \tilde{\beta}_1 - \tilde{\beta}_2 \)) | Mean s.e of (\( \tilde{\beta}_1 - \tilde{\beta}_2 \)) | Mean corr (\( \tilde{\beta}_1, \tilde{\beta}_2 \)) | n    | Meta-analysis model. | Mean of (\( \tilde{\beta}_1 - \tilde{\beta}_2 \)) | Mean s.e of (\( \tilde{\beta}_1 - \tilde{\beta}_2 \)) | Mean corr (\( \tilde{\beta}_1, \tilde{\beta}_2 \)) |
|------|----------------------|----------------|----------------|-----------------|------|----------------------|----------------|----------------|-----------------|------|----------------------|----------------|----------------|----------------|
| 50   | URMA                 | -1.999         | 0.179          | 0               | 10   | URMA                 | -1.991         | 0.346          | -               | 50   | Model A               | -1.998         | 0.0942         | 0.644          | 10   | Model A               | -1.994         | 0.178          | 0.674          | 50   | Model A-zero          | -2.000         | 0.128          | 0.471          | 10   | Model A-zero          | -1.993         | 0.248          | 0.422          |
Summary of the benefits of Model A-zero, with application to MYCN

The usefulness of Model A-zero varies from being reasonably good where $\tau_j^2$ is on average larger in size to the $s_j^2$'s to being quite poor when $\tau_j^2$ is on average smaller in size to the $s_j^2$'s. However, even in the former situation, an increase and greater impact of $\rho_b$ being estimated as 1 or -1 is a major problem. On the occasions where this does not happen, and for the situations Model A was shown to be worthwhile over URMA (e.g. when data is MAR), Model A-zero appears to still be beneficial over URMA (e.g. with reduction in MSE, gain in precision etc.), but not to the same extent as Model A. Indeed, Model A-zero will most often be a conservative alternative to Model A, for example with a slightly increased $\text{var}(\hat{\beta}_i)$; there are two caveats to this however:

(i) I have only considered situations where $\rho_{w_i}$ was truly positive. Where $\rho_{w_i}$ was truly negative, Model A-zero may not be a conservative model, in particular for $(\hat{\beta}_1 - \hat{\beta}_2)$ as $\text{var}(\hat{\beta}_1 - \hat{\beta}_2)$ may often be too small in this situation if $\text{corr}(\hat{\beta}_1, \hat{\beta}_2)$ is negative.

(ii) Neither Model A nor Model A-zero is ideal because they both have the potential to be biased when $\rho_b$ is estimated as 1 or -1. Furthermore, this problem occurs more often in Model A-zero than Model A, and one needs to be extremely cautious about using any results when this problem occurs (see Section 5.6.4).

Given these points, Model A-zero is perhaps most useful when: (i) one knows that setting $\rho_{w_i} = 0$ will underestimate the true $\rho_{w_i}$'s, as then one could be assured Model A-zero was most likely a conservative approach, and (ii) Model A-zero does not produce a $\hat{\rho}_b$ equal to 1 or -1. This scenario may well be rare in practice, however. For an illustrative example consider the MYCN dataset. I knew that it was highly likely that using Model A-zero will underestimate the true $\rho_{w_i}$'s because there is a positive correlation between OS and DFS.
HRs in neuroblastoma (see Section 3.2), and so Model A-zero should be conservative. However, when applying Model A-zero to the MYCN dataset $\hat{\rho}_B$ was equal to 1 (Table 8.3). This is because most of the $s^2_j$'s in this dataset are at least as large as the $\tau^2_j$'s and therefore the $\rho_{w,i}$'s are influential in the estimation procedure, meaning the under specification of $\rho_{w,i}$ has a large impact on $\hat{\rho}_B$. It would perhaps be ill-advised to use the potentially biased Model A-zero results here, even though they are comparable to those from the URMA. To make the most appropriate (i.e. least-biased) evidence-based conclusions one would most likely prefer to use the URMA results in this situation (e.g. for the 95% CI of $\hat{\beta}_j$), for which bias caused by $\hat{\rho}_B$ is not a concern.

Table 8.3: Univariate random-effects (URMA) and Model A-zero results for the MYCN dataset (see Table 3.1), where OS = overall survival, DFS = disease-free survival and s.e. = standard error.

<table>
<thead>
<tr>
<th>Model</th>
<th>DFS</th>
<th>OS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\beta}_1$ (s.e.)</td>
<td>$\tau^2_1$</td>
</tr>
<tr>
<td>URMA</td>
<td>1.478 (0.127)</td>
<td>0.386</td>
</tr>
<tr>
<td>Model A-zero</td>
<td>1.498 (0.122)</td>
<td>0.406</td>
</tr>
</tbody>
</table>

8.1.2 Option 2: sensitivity analysis using a range of $\rho_{w,i}$ values

Where one does not know whether $\rho_{w,i}$ is likely to be positive or negative, instead of assuming the $\rho_{w,i}$'s equal zero, one could assess Model A across a range of $\rho_{w,i}$ values. This type of sensitivity analysis has been used by Berkey et al. in a fixed-effects multivariate analysis of 44 RCTs which evaluated the effectiveness of injectable gold, auranofin and placebo on three treatment outcomes; this was discussed in the literature review in Section 3.6.3 [164]. No matter what values of $\rho_{w,i}$ were assumed, all the sensitivity analyses indicated that gold was significantly better than auranofin on all three outcomes, even though the individual trials reported no significant differences. Given the previous simulations results throughout the thesis, when applying such sensitivity analyses
to Model A where \( \tau_j^2 \) is much larger in size to the \( s_j^2 \)s one should obtain reasonably consistent results because the \( \rho_{ij} \)s will have little impact (e.g. see Section 5.4.4). As an example, consider the original Berkey dataset presented in Table 4.1, relating to the outcomes of probing depth and attachment level, and assume the \( \rho_{ij} \)s were unknown (even though they were available for this dataset). Relatively large \( \tau_j^2 \)s occur for this dataset and so, assuming \( \rho_{ij} \) was unknown in each study, I performed a sensitivity analysis by applying three separate Model A analyses, firstly for an assumed value of \( \rho_{ij} = -0.8 \) for all studies, and then similarly for \( \rho_{ij} = 0 \) and \( \rho_{ij} = 0.8 \) (Table 8.4). All the analyses showed reasonably consistent findings regardless of the \( \rho_{ij} \) value assumed, and the main conclusions did not change.

<p>| Table 8.4: RIGLS univariate (URMA) and bivariate random-effects (Model A) meta-analysis and sensitivity-analysis results for the original Berkey dataset (see Table 4.1) where PD = probing depth, AL = attachment level and s.e. = standard error [151]. |
|---------------------------------|--------|--------|--------|--------|--------|--------|</p>
<table>
<thead>
<tr>
<th>Model</th>
<th>( \tilde{\beta}_1 ) (s.e.)</th>
<th>( \tau_1^2 )</th>
<th>( \tilde{\beta}_2 ) (s.e.)</th>
<th>( \tau_2^2 )</th>
<th>( \hat{\rho}_B )</th>
<th>Corr (( \hat{\beta}_1 ), ( \hat{\beta}_2 ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>URMA</td>
<td>0.361 (0.0592)</td>
<td>0.0119</td>
<td>-0.346</td>
<td>0.0331</td>
<td>-</td>
<td>0.706 (0.107)</td>
</tr>
<tr>
<td>Model A (( \rho_{ij} )s known)</td>
<td>0.353 (0.0590)</td>
<td>0.0117</td>
<td>-0.339</td>
<td>0.0327</td>
<td>0.609</td>
<td>0.547 (0.0740)</td>
</tr>
</tbody>
</table>

**Sensitivity analysis results:**

<table>
<thead>
<tr>
<th>Model</th>
<th>( \rho_{ij} )</th>
<th>( \tilde{\beta}_1 ) (s.e.)</th>
<th>( \tau_1^2 )</th>
<th>( \tilde{\beta}_2 ) (s.e.)</th>
<th>( \tau_2^2 )</th>
<th>( \hat{\rho}_B )</th>
<th>Corr (( \hat{\beta}_1 ), ( \hat{\beta}_2 ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model A assuming ( \rho_{ij} = 0.8 )</td>
<td>0.367 (0.0646)</td>
<td>0.0152</td>
<td>-0.356</td>
<td>0.0334</td>
<td>0.409</td>
<td>0.467</td>
<td>0.723 (0.0817)</td>
</tr>
<tr>
<td>Model A-zero</td>
<td>0.345 (0.0546)</td>
<td>0.0097</td>
<td>-0.325</td>
<td>0.0330</td>
<td>0.775</td>
<td>0.609</td>
<td>0.670 (0.0698)</td>
</tr>
<tr>
<td>Model A assuming ( \rho_{ij} = -0.8 )</td>
<td>0.338 (0.0500)</td>
<td>0.0084</td>
<td>-0.298</td>
<td>0.0350</td>
<td>1</td>
<td>0.677</td>
<td>0.635 (0.0651)</td>
</tr>
</tbody>
</table>

Consistent findings from a sensitivity analysis can clearly strengthen the evidence-based conclusions that can be drawn. However, sensitivity analyses for Model A using a range of \( \rho_{ij} \) values may sometimes introduce more complexity. For example, when \( \tau_j^2 \) is on
average similar or smaller in size to the \( s^2 \) s, using a range of \( \rho_{wi} \) values is likely to cause Model A to estimate \( \rho_B \) as 1 or -1 on at least a few of the occasions, and it would perhaps be inadvisable to use the results from these occasions. This was evident in the Berkey sensitivity analysis results when \( \rho_{wi} = -0.8 \) was assumed (Table 8.4), and, although the findings were consistent with others, one needs to be extremely cautious about using the \( \rho_{wi} = -0.8 \) results because they may be biased (see Section 5.6.4). This problem can also be seen in a similar sensitivity analysis of the MYCN dataset both when \( \rho_{wi} = -0.8 \) and when \( \rho_{wi} = 0 \) was assumed (Table 8.5).

**Table 8.5:** RIGLS univariate (URMA) and bivariate random-effects (Model A) meta-analysis and sensitivity-analysis results for the MYCN dataset (see Table 3.1) where OS = overall survival, DFS = disease-free survival and s.e. = standard error.

<table>
<thead>
<tr>
<th>Model</th>
<th>OS</th>
<th>DFS</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \hat{\beta}_2 ) (s.e.)</td>
<td>( \tau_2^2 )</td>
<td>( \hat{\beta}_2 ) (s.e.)</td>
<td>( \tau_2^2 )</td>
<td>( \hat{\rho}_B )</td>
</tr>
<tr>
<td>URMA</td>
<td>1.627 (0.118)</td>
<td>0.374</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitivity analysis results:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model A assuming ( \rho_{wi} = 0.8 )</td>
<td>1.477 (0.111)</td>
<td>0.382</td>
<td>1.642 (0.108)</td>
<td>0.378</td>
<td>0.777</td>
</tr>
<tr>
<td>Model A-zero</td>
<td>1.498 (0.122)</td>
<td>0.406</td>
<td>1.657 (0.114)</td>
<td>0.427</td>
<td>1</td>
</tr>
<tr>
<td>Model A assuming ( \rho_{wi} = -0.8 )</td>
<td>1.522 (0.144)</td>
<td>0.654</td>
<td>1.706 (0.132)</td>
<td>0.636</td>
<td>1</td>
</tr>
</tbody>
</table>

It is worth adding here that if when assuming a certain \( \rho_{wi} \) value across studies and Model A estimates \( \rho_B \) to be 1, this *does not* necessarily mean that the true \( \rho_{wi} \) values are larger than those values assumed. For example, the simulations in Chapter 5 and Section 6.4 all used *known* \( \rho_{wi} \) s but even in this situation Model A often estimated \( \rho_B \) as 1. One other problem with the sensitivity analyses presented for the Berkey and MYCN datasets is that they assume the \( \rho_{wi} \) s are the same in each study, but of course this may not be appropriate. For example, if two of the largest studies in the meta-analysis have large but oppositely
signed correlations then this may produce very different Model A results than if both were assumed to have the same sign. However, if one decides to do an extended sensitivity analysis that looks at all the different permutations of positive and negative within-study correlation values across studies, then one is going to require a large number of Model A analyses and for which a $\rho_B$ estimated as 1 or -1 will inevitably occur on occasions. This may therefore make it very difficult to form any conclusions that are not potentially subject to bias, and one could therefore argue that a single URMA would provide much clearer and perhaps more helpful evidence-based results than this large-scale sensitivity approach using Model A.

8.1.3 Option 3: Utilising prior knowledge to narrow the range of possible $\rho_{wi}$ values

Model A-zero and the sensitivity analysis approach in Section 8.1.2 are clearly far from ideal, however the latter may be more appealing if one could narrow down the possible range of the $\rho_{wi}$ values. I now consider a method that has been proposed to do just that when two relative risks are the summary statistics of interest.

Berrington and Cox consider the ‘synthesis of correlated information’, which is a similar framework to multivariate meta-analysis but the multiple summary statistics are all assumed to be estimating the same underlying pooled value (i.e. the multiple summary statistics from a study can all be used to estimate the same single pooled value of interest) [189]. Consider the example used within their paper where a series of published case-control studies of oral contraceptive use and cervical cancer were synthesised, and only one pooled estimate was of interest ($\tilde{\beta}$), i.e. the log-relative risk of any form of cervical cancer for long-term oral contraceptive use (say > 5 years). In most of the $n$ publications to be synthesised authors had presented a single relative risk ($\tilde{Y}$) relating to oral contraception and any form of cervical cancer. However, in a few studies authors had
published two relative risks ($\tilde{Y}_{i1}$ and $\tilde{Y}_{i2}$), one for each of two sub-types of cervical cancer (invasive ($j = 1$) and in situ ($j = 2$) cancers). In these studies, as the same control group was used to calculate the relative risks for each of the two sub-groups, $\tilde{Y}_{i1}$ and $\tilde{Y}_{i2}$ are not independent but both could be used to inform $\tilde{\beta}$. However, if one does not acknowledge the within-study correlation between $\tilde{Y}_{i1}$ and $\tilde{Y}_{i2}$, these values will wrongly be treated as two independent pieces of information (as if coming from two separate studies) and therefore $\text{var}(\tilde{\beta})$ will be inappropriate and potentially misleading.

To apply their model, Berrington and Cox also need to overcome the problem of unknown $\rho_{wi}$ values in those studies presenting $\tilde{Y}_{i1}$ and $\tilde{Y}_{i2}$. To do this, the authors perform a sensitivity analysis similar to that discussed in Section 8.1.2. However, in addition they are also able to narrow the range of possible $\rho_{wi}$ values by calculating lower and upper bounds (assuming a negative and a positive $\rho_{wi}$ value respectively) from the $2 \times 2$ tables available in the published articles. The authors conclude that a sensitivity analysis using the upper bound of $\rho_{wi}$ is the most appropriate as it provides the most conservative estimate of $\text{var}(\tilde{\beta})$. The identification of a range of correlation values has similarly helped inform meta-analysis in other contexts [186].

The approach of Berrington and Cox for dealing with unknown within-study correlations could be useful for Model A, where now only studies reporting $\tilde{Y}_{i1}$ and $\tilde{Y}_{i2}$ would be desired (i.e. not $\tilde{Y}_{i}$ anymore) and two pooled estimates ($\tilde{\beta}_1$ and $\tilde{\beta}_2$) would be of interest (e.g. one for each of invasive and in situ cancers). For Model A one would also need to consider a range of values between the lower and upper bounds of $\rho_{wi}$, and not just the upper-bound as suggested, because extreme values of $\tilde{\beta}_1$, $\tilde{\beta}_2$ and their variances may be within this range.
Unfortunately, the method as described is only applicable to relative risk estimates, although it may be possible to derive similar ranges for other effect estimates, such as the log-odds ratio, and this requires further work if the approach can be applied more generally. As such I cannot directly apply the Berrington and Cox approach to the MYCN dataset. However, taking the same philosophical approach, prior knowledge about \( \rho_{wi} \) would suggest that it is highly likely to be non-negative for the MYCN dataset, because of the positive correlation between OS and DFS. It is therefore possible to limit the range of sensitivity analyses previously performed in Section 8.1.2 (Table 8.5) to values of \( \rho_{wi} \) greater than or equal to 0, and all these analyses show consistency with the URMA conclusions about MYCN. However, one still needs to be very cautious about interpreting the result for when the \( \rho_{wi}s \) were assumed zero because \( \hat{\rho}_B \) is equal to 1 for this analysis (Table 8.5). Furthermore, even for this reduced range of \( \rho_{wi} \), permutations of different \( \rho_{wi} \) values across studies should also be considered, which may again lead to further analyses where \( \rho_B \) is estimated as 1.

Hence, in conclusion, the application of the methods of Berrington and Cox are only likely to be an improvement on the sensitivity analyses in Section 8.1.3 when the range of possible \( \rho_{wi} \) values is found to be small relative to the range \(-1 \) to 1. In these circumstances, and when no results include a \( \hat{\rho}_B \) equal to 1 or \(-1 \), the method is highly appealing. Such situations may be few and far between, however the development of similar approaches to Berrington and Cox for other (non relative-risk) situations would make interesting further research. Indeed, where survival proportions at various follow-up times are of interest, Dear reports an iterative method for retrospectively estimating the within-study correlations when an individual study only reports the estimate of the proportion surviving and its standard error for each time-point [166]. However, the
approach requires additional modelling assumptions in order to retrospectively calculate
the within-study correlations, and this is perhaps one reason why similar approaches in
other settings have not yet been considered. One possible solution to the problem arises
where IPD is available for some studies in the meta-analysis, because for these one may be
able to estimate the within-study correlation directly and then perhaps use the average of
these values as a proxy for the missing within-study correlations in the other non-IPD
studies. Further discussion on the synthesis of studies providing IPD with those studies
providing only summary statistics can be found in Section 10.4.4.

8.1.4 Option 4: Placing a prior distribution on $\rho_{wi}$ within Model A-Bayes

Of course, where prior information about the $\rho_{wi}$s is available it may be more suitable to
apply Model A-Bayes rather than Model A, as the Bayesian framework naturally allows
prior information to be incorporated alongside the data (Chapter 7) [67]. However, unless
there is also some strong prior information about the between study parameters (e.g. $\rho_B$),
this approach may be even more problematic than those using the frequentist Model A as
described in Sections 8.1.1 to 8.1.3 above. ‘Vague’ prior distributions for the between-
study parameters can be influential toward the posterior parameter estimates (see Section
7.3), hence even if strong prior information existed for the $\rho_{wi}$s one would still need to
perform assessments of the ‘vague’ prior distributions used for the between-study
parameters. Thus, using Model A-Bayes in this situation may need even more sensitivity
assessments than if the frequentist Model A was used.

This is particularly true when one were to use Model A-Bayes when no prior information
exists for either the $\rho_{wi}$s or $\rho_B$. For example, Nam et al. apply Model A-Bayes where the
two summary statistics of interest are the log-odds ratio for developing asthma and the log-

odds ratio of developing lower respiratory disease, comparing children exposed and
unexposed to passive smoking [147]. No prior information about the \( \rho_w \)'s was available to inform their analysis, and similarly no prior information existed about the between-study parameters. To address this, the authors consider a range of different prior distributions for \( \rho_w \), assuming it was the same in all studies (i.e. \( \rho_w = \rho_w \)), alongside various different prior distributions for \( \rho_b \). This is akin to performing sensitivity analyses as in Section 8.1.2 but has the additional problem of having to also consider prior distributions for the between-study parameters.

Fortunately, Nam et al. report that the pooled estimates were consistent for all prior distributions used (although they do not report details on how the standard errors of the pooled estimates change), and the only estimates that altered were for the correlation parameters themselves. The reason for this is that for this particular dataset there does not appear to be much benefit of Model A or Model A-Bayes over two independent URMA's (see Appendix C4 for results which demonstrate this). In other situations where there are benefits over URMA, such straightforward conclusions may not be possible because both \( \rho_w \) and \( \tilde{\rho}_b \) can be influential in the estimation process. For example, the Model A-Bayes results from the Berkey data show the influence of even 'vague' prior distributions for \( \tilde{\rho}_b \) (see Table 7.1 in Section 7.3), whilst the prior distribution for \( \rho_w \) will always be influential because the data (i.e. \( \bar{Y}_1 \) and \( \bar{Y}_{12} \)) provides no information about \( \rho_w \) as Model A-Bayes starts at the summary statistics level (where \( \rho_w \) is assumed known and cannot be estimated from other information). It may also be inappropriate to make the \( \rho_w = \rho_w \) assumption that Nam et al. used, and where this is not plausible, one may be faced with an even larger scale sensitivity analysis, needing to assess all the permutations of different prior distributions for \( \tilde{\rho}_b \) and now each of the \( \rho_w \)'s.
Using Model A-Bayes when the $\rho_{wi}$s are unknown would therefore seem most suitable when there is strong prior information about the $\rho_{wi}$s and also $\rho_B$. Unfortunately strong prior information was not available for either the $\rho_{wi}$s or $\rho_B$ for my OS and DFS prognostic marker datasets and so I do not apply the Nam et al. approach to the MYCN dataset here.

8.1.5 Summary of current approaches for dealing with unknown $\rho_{wi}$s

All of the possible approaches discussed in Sections 8.1.1 to 8.1.4 impute either a single value or a range of values for each $\rho_{wi}$ in order to assess the sensitivity of Model A or Model A-Bayes to the unknown $\rho_{wi}$s. None of the approaches discussed are ideal but they do become more valuable when the range of possible $\rho_{wi}$ values can be narrowed down. In particular, where the true $\rho_{wi}$s are likely to be positive, Model A-zero is appealing because it will be a conservative alternative to Model A (i.e. produce slightly larger standard errors for $\tilde{\beta}_j$) yet often still be beneficial (i.e. smaller MSE and increased precision for $\tilde{\beta}_j$) over two independent URMAs when Model A itself would have been (see Section 8.1.1).

However, for either complete-case or missing data situations, Model A-zero and all the approaches can unfortunately introduce their own problems (such as producing and having to interpret results where $\tilde{\rho}_B$ is equal to either 1 or -1 in Model A, or requiring additional sensitivity assessments of the prior distributions for the between-study parameters in Model A-Bayes) and may be quite time-consuming, thus not making them immediately appealing over the simpler URMA approach. Indeed, unless I knew Model A-zero was conservative or that only a very small number of sensitivity analyses were required, I would actually prefer to use two independent URMAs for simplicity and interpretability, although of course a URMA is itself not without concerns (especially in MAR situations;
see Section 6.5). One possible exception would be when the $\tilde{r}_j^2$'s are much larger than the $s_{\tilde{y}}^2$'s, as in this situation all the above approaches could help show that $\rho_{w_l}$ has very little impact and $\rho_{g}$ is more unlikely to be estimated as 1 or $-1$ (for example see Section 5.4.4, Table 5.4(iii)). However, for the majority of situations it is clear that further research is required to develop additional methods for facilitating BRMA models when the within-study correlations are unknown. The following sections in this chapter will therefore investigate one possible new approach (called ‘Model B’) to meet this need, and the ultimate aim of this research is the appropriate application of BRMA in the evidence synthesis of prognostic marker studies, where within-study correlations are rarely known.

8.2 Introducing an alternative model for bivariate random-effects meta-analysis

8.2.1 Modelling directly the overall correlation between the $\tilde{Y}_{11}$'s and $\tilde{Y}_{12}$'s

For a BRMA the meta-analyst seeks $\tilde{Y}_{11}$ and $\tilde{Y}_{12}$ from each study, and all the $\tilde{Y}_{11}$'s and $\tilde{Y}_{12}$'s obtained indicate the overall variance of the $\tilde{Y}_{11}$'s and also the overall variance of the $\tilde{Y}_{12}$'s.

The specification of Model A partitions these overall variances into the sum of the within- and the between-study variances, i.e. $s_{\tilde{y}}^2 + \tau_j^2$ in $\mathbf{V}$ (see Section 4.1.1 and equation (4.2)).

As $s_{11}^2$ and $s_{12}^2$ are known, Model A can estimate $\tau_1^2$ and $\tau_2^2$ to obtain a measure of the overall variances indicated by the data. Similarly, Model A partitions the overall covariance between the $\tilde{Y}_{11}$'s and $\tilde{Y}_{12}$'s into the within- and between-study covariances, which can be seen by $\lambda_{11} + \tau_{12}$ in $\mathbf{V}$ (see Section 4.1.1). Given known $\lambda_{11}$'s (and therefore known within-study correlations, i.e. known $\rho_{w_l}$'s), Model A can estimate $\tau_{12}$ and therefore the between-study correlation, $\rho_{g}$.

Alternatively, rather than partitioning the overall covariance into within- and between-study parameters, what if a single parameter could be used to directly indicate the overall...
covariance between the $\bar{Y}_n$'s and $\bar{Y}_{12}$'s? Hence, rather than the $\lambda_i$ and $\tau_{12}$ parameters there would be a single covariance term, or equivalently rather than the $\rho_{w_i}$ and $\rho_B$ parameters there would be a single correlation parameter ($\rho$, say) that models directly the overall correlation between the $\bar{Y}_n$'s and $\bar{Y}_{12}$'s. This situation is specified in Model B below (equation (8.1)), which is essentially a modification of Model A in order to model directly the overall correlation ($\rho$):

**Model B**

$$Y_{ij} \sim N(\beta, \Phi)$$

(8.1)

where

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \quad \text{and} \quad \Phi = \begin{pmatrix} \psi_1^2 + s_{11}^2 & \rho \sqrt{(\psi_1^2 + s_{11}^2)(\psi_2^2 + s_{12}^2)} \\ \rho \sqrt{(\psi_1^2 + s_{11}^2)(\psi_2^2 + s_{12}^2)} & \psi_2^2 + s_{12}^2 \end{pmatrix} = \begin{pmatrix} \psi_1^2 + s_{11}^2 & \delta_i \\ \delta_i & \psi_2^2 + s_{12}^2 \end{pmatrix}$$

One can see that, unlike $\mathbf{V}$ in Model A (see equation (4.2)), the covariance matrix $\Phi$ for Model B is not decomposed into within- and between-study correlation as there is just one overall correlation parameter ($\rho$). However, importantly the within-study variances (i.e. the $s_{ij}^2$'s) are still specified (and are still assumed known) and this allows each study to retain their own individual weighting during the estimation procedure. There is another major change from Model A though, in that $\tau_{ij}^2$ has been replaced by $\psi_{ij}^2$, which explains the extra variation observed in addition to the $s_{ij}^2$'s. It is important to recognise that $\psi_{ij}^2$ is not directly equivalent to $\tau_{ij}^2$, the between-study variance from Model A, although it will be very closely related. The reason they are not the same entity is that by only specifying $\rho$ (and not $\rho_{w_i}$ or $\rho_B$) Model B is not naturally decomposed into within- and between-study components, and it therefore does not have a proper hierarchical structure [190]. Indeed, Model B is a cross between a standard (non meta-analysis) bivariate mixed model (where the overall correlation and also the overall variation are modelled directly and not
decomposed into within and between-study parameters) and Model A (where both the overall variances and the overall correlation are partitioned into within- and between-study parameters). For Model B to be suitable for evidence synthesis when there is missing data one needs to assume that the missing summary statistic is either MCAR or MAR in those studies only providing one outcome, as was the case for Model A (see Section 6.2). Hence, where this assumption cannot be ascertained, Model B should be used a sensitivity analysis to assess the robustness of the URMA conclusions when MCAR or MAR is assumed valid (see Section 6.3.2 and also Section 8.7.3 later). It is worth noting at this stage that Model B has not been previously been suggested in the meta-analysis literature, and is a different model than the reparameterised version of Model A used by Thompson et al. [169].

A clearly desirable property of Model B is that it does not require the \( \rho_m \)s to be available, as one is modelling directly the overall correlation and the \( \rho_m \)s are not a part of Model B. Indeed, the only information required to fit Model A is that which would have been required for two independent URMAs (i.e. \( \hat{Y}_{i1}, s_{i1}^2, \hat{Y}_{i2}, s_{i2}^2 \)). Hence, although Model B is not a hierarchical model, the key question is how beneficial is Model B in relation to both Model A and a URMA for estimating the pooled values \( \beta_1, \beta_2 \), and \( (\beta_1 - \beta_2) \)? For example, does Model B produce unbiased pooled estimates and do they have a smaller MSE and larger precision than the pooled estimates from two independent URMAs on average? Furthermore, does Model B have any problems estimating the correlation \( (\rho) \) similar to those for estimating \( \rho_{B} \) in Model A? I will now address such questions in the next few sections.

8.2.2 RIGLS analytic solutions for Model B, with comparison to Model A

To estimate the unknown parameters in Model B (i.e. \( \beta_1, \beta_2, \psi_1^2, \psi_2^2, \rho \)) one can use restrictive iterative generalised least squares (RIGLS) as for Model A, where
\[ \hat{\beta} = (X^T \Phi^{-1} X)^{-1} X^T \Phi^{-1} Y \]
and \(X\) is the design matrix for the fixed effects \(\beta\), \(Y\) is a vector of the \(\tilde{Y}_y\)s, and \(\Phi\) is an \(n\) by \(n\) matrix block diagonal in \(\Phi_1\) and zero elsewhere. The RIGLS estimation procedure iterates between estimating \(\beta\) and \(\Phi\) until the estimates for each unknown parameter in these matrices have converged to a pre-specified level (e.g. 6 decimal places). Let \(\tilde{\delta}_i = \tilde{\theta}_i \sqrt{\left(\tilde{\psi}_i^2 + s_i^2\right)(\tilde{\psi}_i^2 + s_i^2)}\), then the analytic solution for \(\tilde{\beta}_i\) from Model B at each iteration is:

\[
\hat{\beta}_i = \frac{\sum_{i=1}^{n} \left[ \tilde{Y}_{ii} \left( \left(\tilde{\psi}_i^2 + s_i^2\right)(\tilde{\psi}_i^2 + s_i^2) - \tilde{\delta}_i \, \tilde{\delta}_i \right) - \left( \left(\tilde{\psi}_i^2 + s_i^2\right)(\tilde{\psi}_i^2 + s_i^2) - \tilde{\delta}_i \, \tilde{\delta}_i \right) \right] \right] + \sum_{i=1}^{n} \left[ \tilde{Y}_{ii} \left( \left(\tilde{\psi}_i^2 + s_i^2\right)(\tilde{\psi}_i^2 + s_i^2) - \tilde{\delta}_i \, \tilde{\delta}_i \right) - \left( \left(\tilde{\psi}_i^2 + s_i^2\right)(\tilde{\psi}_i^2 + s_i^2) - \tilde{\delta}_i \, \tilde{\delta}_i \right) \right]}{\sum_{i=1}^{n} \left( \left(\tilde{\psi}_i^2 + s_i^2\right)(\tilde{\psi}_i^2 + s_i^2) - \tilde{\delta}_i \, \tilde{\delta}_i \right) - \left( \left(\tilde{\psi}_i^2 + s_i^2\right)(\tilde{\psi}_i^2 + s_i^2) - \tilde{\delta}_i \, \tilde{\delta}_i \right)}^2} \tag{8.2}
\]

and

\[
\text{var}(\hat{\beta}_i) = \frac{\sum_{i=1}^{n} \left( \left(\tilde{\psi}_i^2 + s_i^2\right)(\tilde{\psi}_i^2 + s_i^2) - \tilde{\delta}_i \, \tilde{\delta}_i \right) \right]}{\sum_{i=1}^{n} \left( \left(\tilde{\psi}_i^2 + s_i^2\right)(\tilde{\psi}_i^2 + s_i^2) - \tilde{\delta}_i \, \tilde{\delta}_i \right) - \left( \left(\tilde{\psi}_i^2 + s_i^2\right)(\tilde{\psi}_i^2 + s_i^2) - \tilde{\delta}_i \, \tilde{\delta}_i \right)}^2} \tag{8.3}
\]

, where \(k = 1, \ldots, \) \(n\) representing the \(n\) studies, and \(k\) is used to distinguish the summation from \(1\) to \(n\) within the summation for \(i = 1, \ldots, \) \(n\). Of course, \(\tilde{\psi}_j^2\) is also an estimate (from the previous iteration) but the analytic solution for \(\tilde{\psi}_j^2\) is complex, just as it was for \(\tilde{\gamma}_j^2\) (Appendix B3), and is thus not described here. One can also estimate the \(\text{cov}(\tilde{\beta}_i, \tilde{\beta}_2)\) at each iteration by:

\[
\text{cov}(\hat{\beta}_i, \hat{\beta}_2) = \frac{\sum_{i=1}^{n} \left( \left(\tilde{\psi}_i^2 + s_i^2\right)(\tilde{\psi}_i^2 + s_i^2) - \tilde{\delta}_i \, \tilde{\delta}_i \right) \right]}{\sum_{i=1}^{n} \left( \left(\tilde{\psi}_i^2 + s_i^2\right)(\tilde{\psi}_i^2 + s_i^2) - \tilde{\delta}_i \, \tilde{\delta}_i \right) - \left( \left(\tilde{\psi}_i^2 + s_i^2\right)(\tilde{\psi}_i^2 + s_i^2) - \tilde{\delta}_i \, \tilde{\delta}_i \right)}^2} \tag{8.4}
\]
The solutions in equations (8.2) to (8.4) are equivalent to those from Model A (see Section 4.1.3), but with \( \tilde{\tau}_{i2} + \lambda_i \) replaced by \( \tilde{\sigma}_i \), and \( \tilde{\tau}_{i}^2 \) replaced by \( \tilde{\psi}_i^2 \), and similarly one can obtain Model B solutions for \( \tilde{\beta}_2 \) and var(\( \tilde{\beta}_2 \)). If one sets \( \rho = 0 \) then \( \delta_i = 0 \) and this reduces Model B to two independent URMA s, where \( \tilde{\psi}_i^2 \) would then be equivalent to \( \tilde{\tau}_i^2 \).

Compare now the \( j = 1 \) analytic solutions from Model A and Model B. I showed in Section 4.5 that \( \tilde{\beta}_i \) from Model A does not utilise the \( \tilde{Y}_{i2} \) values when:

\[
\text{(i)} \quad \tilde{\tau}_{i2} + \lambda_i = 0 \quad \text{for all } i, \text{ or}
\]

\[
\text{(ii)} \quad s_{i1}^2 - s_{k1}^2 = 0 \quad \text{and} \quad \lambda_i = \lambda_k \quad \text{for all } i \text{ and } k.
\]

I also stated in Section 4.5 that, although (i) was sensible, perhaps point (ii) was not very intuitive, because when \( s_{i1}^2 - s_{k1}^2 = 0 \) and \( \lambda_i = \lambda_k \) for all \( i \) it may still seem sensible if Model A could ‘borrow strength’ from the related \( j = 2 \) data, for which the \( s_{i2}^2 \)s may be very different for each \( i \); however this is not the case. Interestingly in Model B this issue is redundant because \( \tilde{\beta}_i \) from Model B does not utilise the \( \tilde{Y}_{i2} \) values when:

\[
\text{(i)} \quad \tilde{\rho} = 0, \text{ or}
\]

\[
\text{(ii)} \quad s_{i1}^2 - s_{k1}^2 = 0 \quad \text{and} \quad \tilde{\delta}_i = \tilde{\delta}_k \quad \text{for all } i \text{ and } k.
\]

Again (i) is sensible, but now (ii) is also more intuitive as when (ii) is true this implies all the \( s_{i2}^2 \)s must also be equal otherwise \( \delta_i \neq \delta_k \). The reason for this difference between Model A and Model B is due to the latter having only one overall correlation term. Thus, unless \( \rho = 0 \), for Model B to revert to two independent URMAs, it is necessary for the URMA weights for \( j = 1 \) to be the same as each other (i.e. \( w_i = \left(s_{i1}^2 + \tau_i^2\right)^{-1} \) are the same for all \( i \) and so too the univariate weights for \( j = 2 \) (i.e. \( w_i = \left(s_{i2}^2 + \tau_2^2\right)^{-1} \) are the same for all \( i \)).
Now compare the covariance between \( \tilde{Y}_{i1} \) and \( \tilde{Y}_{i2} \) as parameterised in Model A to that from Model B, i.e. compare the off-diagonal components of \( V_i \) and \( \Phi_i \) respectively:

Model A: \[
\text{cov}(\tilde{Y}_{i1}, \tilde{Y}_{i2}) = \rho_B \sqrt{\frac{\tau_1^2 + \tau_2^2}{\tau_1^2 \tau_2^2}} + \rho_m \sqrt{s_{1i}^2 s_{2i}^2}
\]
Model B: \[
\text{cov}(\tilde{Y}_{i1}, \tilde{Y}_{i2}) = \rho \sqrt{\left(\psi_1^2 + s_{1i}^2\right)\left(\psi_2^2 + s_{2i}^2\right)}
\]

There are two solutions which will now be used to help compare and contrast Model B with Model A. Firstly, where the \( s_{ij}^2 \) s are very small relative to the between-study variation, the \( \text{cov}(\tilde{Y}_{i1}, \tilde{Y}_{i2}) \) from Model A will tend to \( \rho_B \sqrt{\frac{\tau_1^2 + \tau_2^2}{\tau_1^2 \tau_2^2}} \) and that from Model B will tend to \( \rho \sqrt{\left(\psi_1^2 + s_{1i}^2\right)\left(\psi_2^2 + s_{2i}^2\right)} \). These entities are likely to be similar in this situation because when the \( s_{ij}^2 \) s have little influence, the overall correlation, \( \rho \), will be based mainly on \( \rho_B \), the between-study correlation (see Section 7.5.4); similarly, \( \psi_j^2 \) is also likely to be very similar to \( \tau_j^2 \) in this situation (see Table 8.6 in Section 8.2.3 for evidence of this). Indeed, both Model A and Model B are tending to an ordinary (rather than weighted) least squares model, i.e. not a meta-analysis model, but one where each study has equal weighting and there is only variation between studies because all the studies have very similar weighting due to the \( s_{ij}^2 \) s being relatively small.

Secondly for equation (8.5), consider where the \( s_{ij}^2 \) s are relatively large so that the within-study variation dominates the overall variation; in this situation, the \( \text{cov}(\tilde{Y}_{i1}, \tilde{Y}_{i2}) \) from Model A will tend to \( \rho_m \sqrt{s_{1i}^2 s_{2i}^2} \) and that from Model B will tend to \( \rho \sqrt{s_{1i}^2 s_{2i}^2} \). These two entities may not be very similar, for two reasons:

(i) The \( \rho_m \) in Model B may vary considerably across the \( i \) studies, but Model B uses a common \( \rho \) for each study.
(ii) There is no theoretical justification for $\tilde{\rho}$ from Model B to be similar to the $\rho_{wi}$s from Model A because they are essentially different models. In Model B, $\tilde{\rho}$ relates to the correlation observed between all the $\tilde{Y}_{it}$s and the $\tilde{Y}_{t2}$s across studies, whereas $\rho_{wi}$ in Model A relates to the correlation within one particular study for a single $\tilde{Y}_{it}$ and $\tilde{Y}_{t2}$ calculated from the raw study data (i.e. the IPD, see Section 3.4.2).

It is worth noting here that Model A itself is not very worthwhile when the $s_{i}^{2}$s are large in comparison to the $r_{j}^{2}$s, as it often does not converge or it produces flat likelihoods for the between-study parameters (see Sections 5.4.2 and 5.6.2). Convergence may therefore also be a problem for Model B in this situation.

The above discussions have providing an insight into how and when Model B may differ from Model A. In particular, when the between-study variation dominates, Model B may produce similar results to Model A; however, when the within-study variation dominates, Model B may not be very similar to Model A. In between these two scenarios, for example where the within- and between-study variation are similarly sized, it is more difficult to ascertain the differences between Model B and Model A. Furthermore it remains to be seen whether the pooled estimates from Model B are unbiased, whether the benefits of Model A over URMA remain for Model B (e.g. gain in precision of $\hat{\beta}_{j}$ when some data is MCAR), and indeed whether there are any estimation difficulties when fitting Model B. A simulation study is therefore essential to properly assess the suitability of Model B for a BRMA. Clearly, if Model B is suitable then it would make the implementation of BRMA in practice much more feasible because the approach does not require the $\rho_{wi}$s to be available. This is of particular relevance to the OS and DFS HRs from the neuroblastoma review, and indicates the general potential of Model B to be beneficial in future evidence syntheses of prognostic marker studies.
8.2.3 Fitting Model B in STATA, with application to the Berkey data

In order to assess the performance of Model B in the simulation study, it is obviously important to be able to fit Model B within a statistical software package. However, it is not possible to use SAS Proc Mixed because Model B is not a fully hierarchical model. Hence, it is necessary to specify the likelihood directly and I chose to do this in STATA using the ‘maximize’ procedure and basing my program on the STATA code for Model A provided by Thompson et al. [169]. There are various ways to specify the likelihood in this STATA procedure, and the approach I used was to specify the individual contribution of each study to the total log-likelihood [191]. The STATA procedure then sums these contributions to obtain the total log-likelihood, which it then maximised using Newton-Raphson or another specified optimisation method.

The log-likelihood for a standard multivariate normal model is:

\[-\frac{1}{2} \left[ n \log(2\pi) + \log|V| + (Y - X\beta)^T V^{-1} (Y - X\beta) \right] \]  (8.6)

where \( X \) is a design matrix for the fixed effects \( \beta \) and \( V \) is the covariance matrix of \( Y \).

The contribution of each study to this log-likelihood is:

\[-\frac{1}{2} \left[ \log(2\pi) + \log|V_i| + (Y_i - \beta)^T V_i^{-1} (Y_i - \beta) \right] \]  (8.7)

For Model B one would replace \( V \) by \( \Phi \) in equation (8.6), and one could ignore \( \log(2\pi) \) when maximising this expression as it is simply a constant. Using this, I developed a program for maximising the log-likelihood for Model B using maximum likelihood (Appendix C5). However, to enable Model B results to be compared with those from Model A and a URMA, I needed to modify this first program to obtain restricted maximum likelihood (REML) estimates. REML is equivalent to the RIGLS estimation I used for the URMA and Model A, as for these models I have assumed a normal error structure (see Section 4.1). Using standard notation as in Chapter 4, the REML log-likelihood for a
multivariate normal model (of $k$ order) is:

\[- \frac{1}{2} \left[ (n - k) \log(2\pi) - \log|X^TX| + \log|V| + \log|X^TV^{-1}X| + (Y - X\beta)^T V^{-1}(Y - X\beta) \right] \tag{8.8}\]

This is the same as the log-likelihood in equation (8.6) but with added terms $\log(X^TX)$, $(n - k - 1)\log(2\pi)$, and $\log|X^TV^{-1}X|$. For a bivariate normal model $k = 2$, and for Model B one would again replace $V$ by $\Phi$. To specify the REML log-likelihood for Model B in STATA, I needed to again specify the separate contributions to the total REML log-likelihood of each study, where the sum of these components gave the total value required. Firstly, the terms $(n - k - 1)\log(2\pi)$ and $\log(X^TX)$ could be ignored because they are both constants. The $\log(X^TV^{-1}X)$ term did, however, need incorporating but this was non-trivial as the term itself was a sum (the sum of the individual covariance matrices) and more difficult to partition than a product (i.e. whereas $\log(ab) = \log(a) + \log(b)$, it is more difficult to partition $\log(a + b)$). Hence, to overcome this I divided $\log(X^TV^{-1}X)$ by $n$ and gave each individual study an $n$th of it in their individual log-likelihood component. Even though this meant each study’s individual log-likelihood component was not necessarily correct, the sum of the components would still produce the overall correct total REML log-likelihood required, and it is this total which is maximised by STATA. The syntax for the REML estimation of Model B in STATA is shown in Appendix C6.

To demonstrate the use of my STATA program I applied it to the original and modified Berkey datasets (see Table 4.1) once more, and then compared the results of Model B to those from a URMA, Model A and also Model A-zero (Table 8.6). Of course, for application of Model B and Model A-zero the within-study correlations were assumed unknown, even though in truth they were available and used for the Model A analysis. The
results obtained for Model B were comparable to those from Model A for both the original and modified datasets. This is perhaps not surprising given that Model A itself does not appear to be very different to a URMA in terms of the $\hat{\beta}_j$ s, whilst the $\rho_{w}$ s in Model A appear to have little impact because Model A-zero is also comparable to Model A (Table 8.6). In terms of $(\hat{\beta}_1 - \hat{\beta}_2)$, Model B and Model A-zero are again comparable to Model A, and far better than a URMA because the $\text{var}(\hat{\beta}_1 - \hat{\beta}_2)$ is appropriately smaller, taking into account $\text{corr}(\hat{\beta}_1, \hat{\beta}_2)$.

Table 8.6: REML Model B results alongside RIGLS estimates for Model A-zero and a univariate (URMA) and bivariate (Model A) random-effects meta-analysis for the original and modified Berkey datasets (see Table 4.1) where s.e. = standard error.

<table>
<thead>
<tr>
<th>Probing Depth</th>
<th>Attachment Level</th>
<th>Corr $(\hat{\beta}_1, \hat{\beta}_2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model</strong></td>
<td>$\hat{\beta}_1$</td>
<td>$\hat{\beta}_2$</td>
</tr>
<tr>
<td><strong>Original Data:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>URMA</td>
<td>0.361</td>
<td>-0.346</td>
</tr>
<tr>
<td>(0.0592)</td>
<td>(0.0885)</td>
<td></td>
</tr>
<tr>
<td>Model A</td>
<td>0.353</td>
<td>-0.339</td>
</tr>
<tr>
<td>(0.0590)</td>
<td>(0.0879)</td>
<td></td>
</tr>
<tr>
<td>Model A-zero</td>
<td>0.345</td>
<td>-0.325</td>
</tr>
<tr>
<td>(0.0546)</td>
<td>(0.0879)</td>
<td></td>
</tr>
<tr>
<td>Model B</td>
<td>0.358</td>
<td>-0.345</td>
</tr>
<tr>
<td>(0.0595)</td>
<td>(0.0867)</td>
<td></td>
</tr>
<tr>
<td><strong>Modified Data:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>URMA</td>
<td>0.378</td>
<td>-0.346</td>
</tr>
<tr>
<td>(0.0660)</td>
<td>(0.0885)</td>
<td></td>
</tr>
<tr>
<td>Model A</td>
<td>0.378</td>
<td>-0.325</td>
</tr>
<tr>
<td>(0.0660)</td>
<td>(0.0877)</td>
<td></td>
</tr>
<tr>
<td>Model A-zero</td>
<td>0.378</td>
<td>-0.325</td>
</tr>
<tr>
<td>(0.0660)</td>
<td>(0.0877)</td>
<td></td>
</tr>
<tr>
<td>Model B</td>
<td>0.381</td>
<td>-0.337</td>
</tr>
<tr>
<td>(0.0673)</td>
<td>(0.0885)</td>
<td></td>
</tr>
</tbody>
</table>

N.B. For the modified data, the Model A-zero results are identical to Model A because the true value of all the $\rho_{w}$ s is zero.
Perhaps the most important thing to observe from the Berkey analyses is that Model B appears to be conservative for \( (\hat{\beta}_1 - \hat{\beta}_2) \) in relation to Model A, in that the standard error of \( (\hat{\beta}_1 - \hat{\beta}_2) \) is slightly higher in this model. This is not true for Model A-zero, whose standard error of \( (\hat{\beta}_1 - \hat{\beta}_2) \) is less than or equal to that from Model A in these two particular examples. The conservatism of Model B arises because it is not a fully hierarchical model like Model A (see Section 8.2.2), which also incorporates the extra known information about the \( \rho_m \)s, and this conservatism could be a particularly important property, and may lead to more statistically appropriate results than the (perhaps strong) assumption in Model A-zero that the \( \rho_m \)s are zero.

8.3 A simulation study to assess the benefits and limitations of Model B

To assess the benefits and limitations of Model B, I used a selection of the datasets simulated from Model A in Chapter 5 (complete-case data, \( n = 50 \) and \( n = 5 \)) and Section 6.4 (missing data, \( n = 50 \) and \( n = 10 \)), and I compared the results from applying Model B to those from applying Model A, Model A-zero and two independent URMAs. Model A is a fully hierarchical model (see Section 3.6.1), whereas Model B is not (see Section 8.2.1), and so, rather than assessing datasets simulated from Model B, I considered it more appropriate to assess datasets simulated from Model A as this is likely to be more realistic for meta-analysis in practice. Hence, datasets were simulated from Model A using parameter values of \( \beta_1 = 0 \) and \( \beta_2 = 2 \), with the \( s_{ij}^2 \)s as described in Section 5.2.1 and each of the following three within- and between-study settings used:

Setting (i): \( \rho_{wi} = 0.8 \) for all studies; \( \rho_B = 0.8 \); \( \tau_j^2 = 1.5 \), on average larger than the \( s_{ij}^2 \)s

Setting (ii): \( \rho_{wi} = 0.8 \) for all studies; \( \rho_B = 0.8 \); \( \tau_j^2 = 0.25 \), on average similar to the \( s_{ij}^2 \)s

Setting (iii): \( \rho_{wi} = 0.8 \) for all studies; \( \rho_B = 0.8 \); \( \tau_j^2 = 0.0025 \), on average smaller than the \( s_{ij}^2 \)s
Unfortunately, I found that comparison of Model B to the other models was not feasible in Setting (iii) because Model B would most often not converge in this situation, which was a similar problem for Model A-zero. This is perhaps not surprising given the simulations for Model A would often not converge in this situation either, or would mostly obtain a \( \tilde{\rho}_B \) of 1 or -1 (see Section 5.4.2). For Model A (and Model A-zero) the problem in this situation is caused by the within-study variation (i.e. the \( s_f^2 \)s) being the major component of the overall variation, and thus the profile likelihoods for the between-study parameters are often very flat (see Section 5.6.2). For Model B I discussed in Section 8.2.2 that when the within-study variation dominates the total variation, the \( \text{cov}(\tilde{Y}_{1i}, \tilde{Y}_{12}) \) from Model B will tend to \( \tilde{\rho}_V \sqrt{s_f^2 s_{f2}^2} \). This evidently makes it increasingly difficult to estimate \( \rho \), especially as the \( s_f^2 \)s and \( s_{f2}^2 \)s are fixed. Of course, where the within-study variation (i.e. the \( s_f^2 \)s and \( s_{f2}^2 \)s) totally dominates is where there is no between-study variation, and this is the situation where a bivariate fixed-effects meta-analysis, rather than a BRMA, should be applied. Thus, where a bivariate fixed-effects meta-analysis is truly required, one will find that Model B, Model A-zero and, to a slightly lesser extent, Model A will have particular difficulties converging or producing sensible results. Model B is therefore not suitable when the \( s_f^2 \)s are relatively large, and in this situation one really does require the within-study correlations to be available in order to estimate the pooled values.

The following assessments of Model B therefore focus on the results for Settings (i) and (ii), where Model B converges more easily as the \( \tau_f^2 \)s become more dominant in relation to the \( s_f^2 \)s. As I wanted to assess Model B directly to Model A in these settings, I only compared the results from the simulated datasets that produced convergence in both Model B and Model A in order to ensure an appropriate comparison and thus avoid introducing any bias in the model comparisons themselves. When Model B converged it was extremely
rare for the equivalent analysis for Model A to not converge, and so virtually all the Model B simulations that converged are still reported in these comparisons. Of course, the full assessment of all the simulations that converged just for Model A has already been done (see Chapter 5 and Section 6.4, and also similarly Section 8.1.1 for all the Model A-zero simulation results that converged).

8.3.1 *Complete-case data simulation results for* $n = 50$ *studies in relation to* $\hat{\beta}_j$

Consider the complete-case $n = 50$ results first and just in relation to the $\hat{\beta}_j$ s themselves; most of the simulations converged in both Model A and Model B (Table 8.7). The most encouraging observation now is that the overall correlation ($\rho$) is never estimated as 1 or -1 for either Setting (i) or (ii). This is in contrast to the number of occasions where $\hat{\rho}_B$ was equal to either 1 or -1 in Model A (20 times for Setting (ii)) and Model A-zero (60 times for Setting (i), 976 times for Setting (ii)) (Table 8.7, Table 8.1, Appendix C2). The $\hat{\beta}_j$ s are also unbiased in all models and their standard error, MSE and coverage are very similar in Model A and Model B (Table 8.7, Figure 8.1 and Figure 8.2). Model B exhibits better properties than Model A-zero on average, in particular in terms of the precision and MSE, and keeps most of the (albeit negligible) benefits Model A has over two independent URMAs. Furthermore, when comparing just those results where Model A did *not* estimate $\rho_B$ as 1 or -1, the $\text{var}(\hat{\beta}_j)$ from Model B was *at least as large* as that from Model A on average, which again indicates the conservatism of Model B.

Consider now the results of Model B compared to those from Model A where $\rho_B$ was estimated as 1 or -1. For Setting (ii), there were 20 simulations where Model A estimated $\rho_B$ as 1 and the equivalent 20 Model B results gave a $\text{var}(\hat{\beta}_j)$ noticeably *smaller* on average, contradictory to the results when $\hat{\rho}_B$ was not 1 or -1, where $\text{var}(\hat{\beta}_j)$ was slightly
larger on average (Table 8.7). This suggests that, when the problem of \( \rho_B \) being estimated as 1 occurs in Model A, the equivalent simulation results in Model B are not subject to an upward bias in \( \bar{\psi}_j^2 \), unlike Model A which has an upward bias in \( \bar{\tau}_j^2 \) (and therefore an inflation in \( \text{var}(\hat{\rho}_j) \)) on average in this situation. The lack of any observed bias in the Model B results may be due to the fact that \( \hat{\rho} \) was always in the range 0.48 to 0.95, and thus no results are included where \( \hat{\rho} \) is 1 or -1, or even extremely close to 1 or -1.

Table 8.7: Results when applying the univariate (URMA) and various bivariate random-effects meta-analysis models to simulated data from Model A where the number of studies was 50 and the true values parameter values were \( \beta_1 = 0 \) and \( \beta_2 = 2 \). Two different settings were assessed relating to the how the data were simulated (see Section 8.3). The number of simulations ('no. sims') compared were only those that converged in both Model A and Model B.

(i) \( \bar{\tau}_j^2 \)'s were larger in size to the \( s_j^2 \)'s on average

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>Meta-analysis Model</th>
<th>No. sims</th>
<th>Mean Estimate</th>
<th>No. of 95% CIs including true value (%)</th>
<th>Mean-Square Error</th>
<th>Mean s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>URMA</td>
<td>999</td>
<td>-0.001 1.995</td>
<td>957 (95.8%) 948 (94.9%)</td>
<td>0.0380 0.0383</td>
<td>0.198 0.200</td>
</tr>
<tr>
<td>Model A</td>
<td>999</td>
<td>0.001 1.995</td>
<td>956 (95.7%) 954 (95.5%)</td>
<td>0.0356 0.0358</td>
<td>0.194 0.196</td>
<td></td>
</tr>
<tr>
<td>Model B</td>
<td>999</td>
<td>0.0007 1.996</td>
<td>951 (95.2%) 956 (95.7%)</td>
<td>0.0363 0.0362</td>
<td>0.192 0.194</td>
<td></td>
</tr>
<tr>
<td>Model A-zero</td>
<td>999</td>
<td>0.0011 1.995</td>
<td>950 (95.1%) 954 (95.5%)</td>
<td>0.0365 0.0372</td>
<td>0.196 0.198</td>
<td></td>
</tr>
</tbody>
</table>

(ii) \( \bar{\tau}_j^2 \)'s were similar in size to the \( s_j^2 \)'s on average

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>Meta-analysis Model</th>
<th>No. sims</th>
<th>Mean Estimate</th>
<th>No. of 95% CIs including true value (%)</th>
<th>Mean-Square Error</th>
<th>Mean s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>URMA</td>
<td>1000</td>
<td>-0.0044 1.993</td>
<td>946 (94.6%) 945 (94.5%)</td>
<td>0.0102 0.0114</td>
<td>0.102 0.106</td>
</tr>
<tr>
<td>Model A</td>
<td>1000</td>
<td>-0.00311 1.995</td>
<td>951 (95.1%) 939 (93.9%)</td>
<td>0.00929 0.009511</td>
<td>0.0943 0.0973</td>
<td></td>
</tr>
<tr>
<td>Model B</td>
<td>1000</td>
<td>-0.00294 1.996</td>
<td>945 (94.5%) 941 (94.1%)</td>
<td>0.00885 0.008396</td>
<td>0.0954 0.0972</td>
<td></td>
</tr>
<tr>
<td>Model A-zero</td>
<td>1000</td>
<td>-0.0051 1.993</td>
<td>963 (96.3%) 965 (96.5%)</td>
<td>0.00982 0.0105</td>
<td>0.104 0.108</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>Meta-analysis Model</th>
<th>No. sims</th>
<th>Mean Estimate</th>
<th>No. of 95% CIs including true value (%)</th>
<th>Mean-Square Error</th>
<th>Mean s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>URMA</td>
<td>980</td>
<td>-0.0046 1.993</td>
<td>926 (94.5%) 925 (94.4%)</td>
<td>0.0103 0.0115</td>
<td>0.102 0.106</td>
</tr>
<tr>
<td>Model A</td>
<td>980</td>
<td>-0.0034 1.995</td>
<td>932 (95.1%) 921 (94.0%)</td>
<td>0.00932 0.00949</td>
<td>0.0944 0.0973</td>
<td></td>
</tr>
<tr>
<td>Model B</td>
<td>980</td>
<td>-0.0036 1.994</td>
<td>926 (94.5%) 923 (94.2%)</td>
<td>0.00876 0.008976</td>
<td>0.0955 0.0972</td>
<td></td>
</tr>
<tr>
<td>Model A-zero</td>
<td>980</td>
<td>-0.0054 0.104</td>
<td>943 (96.2%) 945 (96.4%)</td>
<td>0.00985 0.0105</td>
<td>0.104 0.108</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>Meta-analysis Model</th>
<th>No. sims</th>
<th>Mean Estimate</th>
<th>No. of 95% CIs including true value (%)</th>
<th>Mean-Square Error</th>
<th>Mean s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>URMA</td>
<td>20</td>
<td>0.00644 2.002</td>
<td>20 (100%) 20 (100%)</td>
<td>0.00758 0.00977</td>
<td>0.101 0.109</td>
</tr>
<tr>
<td>Model A</td>
<td>20</td>
<td>0.214 2.026</td>
<td>19 (90%) 18 (90%)</td>
<td>0.00747 0.0075</td>
<td>0.0939 0.0988</td>
<td></td>
</tr>
<tr>
<td>Model B</td>
<td>20</td>
<td>0.216 2.024</td>
<td>19 (90%) 18 (90%)</td>
<td>0.00775 0.01066</td>
<td>0.0894 0.0965</td>
<td></td>
</tr>
<tr>
<td>Model A-zero</td>
<td>20</td>
<td>0.0121 2.010</td>
<td>20 (100%) 20 (100%)</td>
<td>0.00798 0.00955</td>
<td>0.106 0.113</td>
<td></td>
</tr>
</tbody>
</table>

R.D. Riley, Ph.D. Thesis Chapter 8 269
Figure 8.1: Model A versus Model B results from all the 999 simulations that converged in both models for the \( n = 50 \) simulations of Table 8.7, where the \( \tau_j^2 \)s were larger in size to the \( s_j^2 \)s on average (the solid line shows line of equality, dotted line shows linear regression line through the points).

Figure 8.2: Model A versus Model B results from simulations that converged in both models for the \( n = 50 \) simulations of Table 8.7 where the \( \tau_j^2 \)s were similar in size to the \( s_j^2 \)s on average (the solid line shows line of equality, dotted line shows linear regression line through the points).

(i) All where \( \hat{\rho}_B \neq 1 \) or \(-1\) in Model A (980 results to compare)

(ii) Just those where \( \hat{\rho}_B = 1 \) in Model A (20 results to compare)
8.3.2 Complete-case data results for \( n = 5 \) studies in relation to \( \hat{\beta}_j \)

Encouragingly for the \( n = 5 \) simulations, \( \rho \) is again never estimated as 1 or -1 for either Setting (i) or (ii), which is in contrast to the large number of occasions where \( \rho_B \) is estimated as 1 or -1 in Model A and Model A-zero (Figure 8.3). However, Figure 8.4(i) shows that Model B does not appear to be as similar to Model A when \( n = 5 \) as when \( n = 50 \) because there are a number of simulation results for Model B where \( \hat{\beta}_j \) and \( \text{var}(\hat{\beta}_j) \) are extremely different than Model A. Investigating this further I found that these extreme discrepancies were related to the set of simulations where \( \rho_B \) was estimated as 1 or -1 in Model A, because the simulations not estimating \( \rho_B \) as 1 or -1 in Model A gave comparable results in Model B (Figure 8.4(ii)).

For a dataset that leads to \( \rho_B \) being estimated as 1 or -1 in Model A I have shown that the Model A results are biased (see Section 5.5). However, for some of the datasets for which Model A gave a \( \hat{\rho}_B \) equal to 1 or -1, Figure 8.4(i) indicates that Model B is also potentially producing some biased results. For example, for one of the simulated datasets Model A gave \( \hat{\beta}_1 \) equal to 0.1 but Model B gave \( \hat{\beta}_1 \) equal to 1.7, indicating a relatively large discrepancy between models (Figure 8.4(i)). Furthermore, some of the \( \text{var}(\hat{\beta}_j) \) values from Model B were extremely close to zero (which does not seem sensible given the \( s_{ij}^2 \)s themselves have a median value of 0.25, see Section 5.2.1) and some others were extremely high (Figure 8.4(i)).
Figure 8.3: (a) Histogram of $\hat{\rho}_B$ values from the simulation results for Model A where $n = 5$ and the $\tau_j^2$'s were similar in size to the $s_y^2$'s on average (N.B. the values on the x-axis are the centre of each bar), and (b) Histogram of $\hat{\rho}$ values from the simulation results from Model B where $n = 5$ and the $\tau_j^2$'s were similar in size to the $s_y^2$'s on average (N.B. the values on the x-axis are the values at the end of each bar). In Model A $\hat{\rho}_B$ was often $-1$ or $1$ which is why there are bars centred at 1 and $-1$; however in Model B $\hat{\rho}$ was never 1 or $-1$, but sometimes $\hat{\rho}$ was very close to 1 which is why the final bar in (b) starts from about 0.95 and goes up to the value of 1.

Figure 8.4: Model A versus B results for those $n = 5$ simulations (for $\tau_j^2$ similar in size to $s_y^2$) that converged in both; the solid line shows line of equality, dotted line shows linear regression line through the points. Those large differences between models relate to when $\hat{\rho}$ was very close to 1 in Model B (see next page).

(i) All that converged in Model A and Model B

(ii) Just those from (i) above where $\hat{\rho}_B \neq 1$ or $-1$ in Model A
Investigating such extreme Model B results further I found they all involved a $\hat{\rho}$ very close to 1 (e.g. $\hat{\rho} = 0.99$), which is where the REML analytic solutions for $\hat{\beta}_j$ and var($\hat{\beta}_j$) start to be divided by a value very close to zero. The analytic solution for $\hat{\beta}_i$ from Model B is:

$$
\hat{\beta}_i = \frac{\left( \sum_{i=1}^{n} \left( \frac{\tilde{Y}_{i1}}{(\tilde{\psi}_i^2 + s_{i1}^2)(\tilde{\psi}_i^2 + s_{i2}^2)} - \hat{\delta}_i \right)^2 \right)}{\sum_{i=1}^{n} \frac{\tilde{Y}_{i2}}{(\tilde{\psi}_i^2 + s_{i1}^2)(\tilde{\psi}_i^2 + s_{i2}^2)} - \hat{\delta}_i} \left( \sum_{i=1}^{n} \frac{\tilde{Y}_{i1}^2 + s_{i1}^2}{(\tilde{\psi}_i^2 + s_{i1}^2)(\tilde{\psi}_i^2 + s_{i2}^2)} - \hat{\delta}_i \right)^2}
$$

(8.9)

As $\hat{\delta}_i = \sqrt{\frac{\tilde{Y}_{i1}^2 + s_{i1}^2}{(\tilde{\psi}_i^2 + s_{i1}^2)(\tilde{\psi}_i^2 + s_{i2}^2)}}$, when $\hat{\rho} = 1$ this will make the denominators of all the terms in equation (8.9) equal to zero. Hence, as $\hat{\rho}$ tends very close to 1 the denominators in equation (8.9) will tend toward zero which thus causes each term to tend toward infinity. It is this problem which causes Model B to produce potentially misleading values of $\hat{\beta}_j$ and var($\hat{\beta}_j$).

The problem of dividing by zero in equation (8.9) is emphasised further by considering the profile log-likelihood for one of the $n = 5$ simulations where Model B actually gave $\hat{\rho} = 0.999$ (Figure 8.5). One can see that the profile log-likelihood for $\hat{\rho}$ increases sharply as $\hat{\rho}$ gets close to 1. The values of $\hat{\beta}_j$ and $\tilde{\psi}_j^2$ also change rapidly as $\hat{\rho}$ approaches 1 (Figure 8.6), which makes me doubt whether Model B truly had reached convergence when $\hat{\rho}$ was extremely close to 1 or -1, even though the STATA 'maximize' procedure indicated convergence had been achieved at the $10^{-6}$ level.
The problem of \( \hat{\rho} \) being very close to 1 or -1 in Model B is potentially related to the proposed Reason (II) why Model A often estimates \( \rho_B \) to be 1 or -1 (see Section 5.6.2).

Essentially there is relatively little information about the correlation between the \( \hat{Y}_{1i} \)s and \( \hat{Y}_{12} \)s from the 5 studies that the likelihood for \( \hat{\rho} \) is poorly defined and gradually increases as \( \hat{\rho} \) approaches 1 or -1. No model can overcome this problem, and what is truly needed is more data or external information to allow the correlation to be better defined.
Importantly, the simulation results do show however that the problem of flat and poorly
defined profile likelihoods for the correlation parameters is more of a problem in Model A
than in Model B. For example, of the \( n = 50 \) simulations in Figure 8.2 Model A had 20
occasions where \( \rho_b \) was estimated as 1 but Model B never estimated \( \rho \) above 0.95, with
no signs of any strange values of \( \text{var}(\tilde{\beta}_j) \) for any of the Model B analyses. One possible
explanation for this is that the overall correlation between the \( \tilde{Y}_i \)'s and the \( \tilde{Y}_{i2} \)'s is better
defined than its components, i.e. the within and between-study correlations. Importantly,
this means that **Model B may be useful over and above Model A, even when the within-
study correlations are known**, because for some situations where Model A estimates \( \rho_b \)
as 1 or -1 the simulations show that Model B will actually allow a well-defined and
sensible value of \( \tilde{\rho} \) (see Section 8.3.5 for further evidence of this).

Given that a \( \tilde{\rho} \) very close to 1 or -1 in Model B may cause misleading answers, I further
limited my investigations of Model B to just those simulation results where \(-0.95 < \tilde{\rho} <
0.95\), and I assumed those results giving a \( \tilde{\rho} \) outside this range could not be trusted. The
\( \pm 0.95 \) limit was chosen because \( \tilde{\rho} \) values up to 0.95 did not appear to be causing any
problems of bias in the \( n = 50 \) simulations of Section 8.3.1 (further discussion on this issue
is given in Section 8.4.3). Using this restricted range, the \( n = 5 \) complete-case simulation
results are shown in Table 8.8 and also in Appendix C7. Furthermore, the Model B results
are plotted against the Model A results in Appendix C8. Although there were now an
incomplete set of the 1000 simulations to compare, the results indicate that the coverage of
\( \tilde{\beta}_j \) from Model B was comparable to both Model A and a URMA, and also far better than
Model A-zero. Furthermore, the standard error and MSE of \( \tilde{\beta}_j \) from Model B were again
slightly larger than those from Model A (Table 8.8), but slightly smaller than those from
Model A-zero on average.
8.3.3 Complete-case data results for $n = 5$ and $n = 50$ studies in relation to $(\hat{\beta}_1 - \hat{\beta}_2)$

I have shown that for complete-case data Model A is only beneficial over URMA for estimating $(\beta_1 - \beta_2)$ (see Section 5.5) and thus I will now focus on this for Model B. The model comparisons were restricted to the results from those datasets which:

(i) converged in both Model A and Model B, and

(ii) gave a $-0.95 < \tilde{\rho} < 0.95$ in Model B.

However, as long as the equivalent analysis for Model B met the above criteria, there was no restriction on incorporating Model A results where $\hat{\rho}_B$ was equal to either 1 or -1 (as I wanted to compare Model B to these occasions). Criteria (i) and (ii) will be used throughout my simulation comparisons from this point forward.

The results for $(\hat{\beta}_1 - \hat{\beta}_2)$ are shown in Table 8.8. Conditional on $-0.95 < \tilde{\rho} < 0.95$, Model B appears a suitable BRMA model as, for the simulations compared, it obtains unbiased estimates of $(\hat{\beta}_1 - \hat{\beta}_2)$ with standard error and MSE only very slightly larger than those from Model A on average. The coverage is also very close to 95% in most situations, far better than that from a URMA or from Model A-zero, and even on occasions better than that from Model A. The $\text{corr}(\hat{\beta}_1, \hat{\beta}_2)$ is also very close to that from Model A on average.
Table 8.8: Summary of all $\hat{\beta}_j$ and $(\hat{\beta}_1 - \hat{\beta}_2)$ from the complete-case data simulation results for all the simulations (no. sims) out of 1000 that passed the comparison criteria specified in Section 8.3.3. The true parameter values were $\beta_1 = 0$ and $\beta_2 = 2$.

(i) $n = 50$ simulation results, where the 1000 datasets were simulated from Model A for $\tau^2_j$'s larger in size to the $s^2_j$'s on average.

<table>
<thead>
<tr>
<th>Meta-analysis Model</th>
<th>No. sims</th>
<th>Mean estimate</th>
<th>No. of 95% CIs including true value (%)</th>
<th>Mean-Square Error</th>
<th>Mean s.e.</th>
<th>corr ($\hat{\beta}_1, \hat{\beta}_2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>URMA 999</td>
<td>-0.001</td>
<td>1.995</td>
<td>-1.996</td>
<td>957 (95.8%)</td>
<td>948 (94.9%)</td>
<td>999 (100%)</td>
</tr>
<tr>
<td>Model A 999</td>
<td>0.001</td>
<td>1.995</td>
<td>-1.994</td>
<td>966 (95.7%)</td>
<td>954 (95.5%)</td>
<td>931 (93.2%)</td>
</tr>
<tr>
<td>Model B 999</td>
<td>0.0007</td>
<td>1.996</td>
<td>-1.995</td>
<td>961 (95.2%)</td>
<td>956 (95.7%)</td>
<td>942 (94.3%)</td>
</tr>
<tr>
<td>Model A-zero 999</td>
<td>0.0011</td>
<td>1.995</td>
<td>-1.994</td>
<td>954 (95.5%)</td>
<td>927 (92.8%)</td>
<td>905 (90.6%)</td>
</tr>
</tbody>
</table>

(ii) $n = 50$ simulation results, where the 1000 datasets were simulated from Model A for $\tau^2_j$'s similar in size to the $s^2_j$'s on average.

<table>
<thead>
<tr>
<th>Meta-analysis Model</th>
<th>No. sims</th>
<th>Mean estimate</th>
<th>No. of 95% CIs including true value (%)</th>
<th>Mean-Square Error</th>
<th>Mean s.e.</th>
<th>corr ($\hat{\beta}_1, \hat{\beta}_2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>URMA 1000</td>
<td>-0.0044</td>
<td>1.993</td>
<td>-1.997</td>
<td>956 (95.6%)</td>
<td>945 (94.5%)</td>
<td>998 (99.8%)</td>
</tr>
<tr>
<td>Model A 1000</td>
<td>-0.0031</td>
<td>1.995</td>
<td>-1.992</td>
<td>951 (95.1%)</td>
<td>939 (93.9%)</td>
<td>940 (94.0%)</td>
</tr>
<tr>
<td>Model B 1000</td>
<td>-0.0029</td>
<td>1.996</td>
<td>-1.999</td>
<td>945 (94.5%)</td>
<td>941 (94.1%)</td>
<td>954 (95.4%)</td>
</tr>
<tr>
<td>Model A-zero 1000</td>
<td>-0.0051</td>
<td>1.993</td>
<td>-1.998</td>
<td>963 (96.3%)</td>
<td>965 (96.5%)</td>
<td>986 (98.6%)</td>
</tr>
</tbody>
</table>

(iii) $n = 5$ simulation results, where the 1000 datasets were simulated from Model A for $\tau^2_j$'s larger in size to the $s^2_j$'s on average.

<table>
<thead>
<tr>
<th>Meta-analysis Model</th>
<th>No. sims</th>
<th>Mean estimate</th>
<th>No. of 95% CIs including true value (%)</th>
<th>Mean-Square Error</th>
<th>Mean s.e.</th>
<th>corr ($\hat{\beta}_1, \hat{\beta}_2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>URMA 787</td>
<td>0.0162</td>
<td>2.0126</td>
<td>-1.996</td>
<td>743 (94.4%)</td>
<td>734 (93.3%)</td>
<td>786 (99.9%)</td>
</tr>
<tr>
<td>Model A 787</td>
<td>0.0182</td>
<td>2.0119</td>
<td>-1.994</td>
<td>744 (94.5%)</td>
<td>735 (93.4%)</td>
<td>750 (95.3%)</td>
</tr>
<tr>
<td>Model B 787</td>
<td>0.0173</td>
<td>2.0122</td>
<td>-1.995</td>
<td>748 (95.0%)</td>
<td>739 (93.9%)</td>
<td>750 (95.3%)</td>
</tr>
<tr>
<td>Model A-zero 787</td>
<td>0.0169</td>
<td>2.013</td>
<td>-1.996</td>
<td>747 (94.9%)</td>
<td>741 (94.2%)</td>
<td>755 (95.9%)</td>
</tr>
</tbody>
</table>

(iv) $n = 5$ simulation results, where the 1000 datasets were simulated from Model A for $\tau^2_j$'s similar in size to the $s^2_j$'s on average.

<table>
<thead>
<tr>
<th>Meta-analysis Model</th>
<th>No. sims</th>
<th>Mean estimate</th>
<th>No. of 95% CIs including true value (%)</th>
<th>Mean-Square Error</th>
<th>Mean s.e.</th>
<th>corr ($\hat{\beta}_1, \hat{\beta}_2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>URMA 604</td>
<td>-0.0144</td>
<td>1.981</td>
<td>-1.995</td>
<td>596 (96.7%)</td>
<td>559 (92.5%)</td>
<td>604 (100%)</td>
</tr>
<tr>
<td>Model A 604</td>
<td>-0.0163</td>
<td>1.976</td>
<td>-1.964</td>
<td>594 (98.3%)</td>
<td>559 (92.5%)</td>
<td>588 (97.4%)</td>
</tr>
<tr>
<td>Model B 604</td>
<td>-0.0119</td>
<td>1.990</td>
<td>-1.992</td>
<td>594 (98.3%)</td>
<td>580 (92.7%)</td>
<td>588 (97.4%)</td>
</tr>
</tbody>
</table>

N.B. URMA = univariate random-effects meta-analysis

R.D. Riley, Ph.D. Thesis Chapter 8
8.3.4 Missing completely at random simulation results for $\tilde{\beta}_j$ and $(\tilde{\beta}_1 - \tilde{\beta}_2)$

The complete-case simulations have shown that Model B appears to be beneficial for estimating $(\beta_1 - \beta_2)$ when the model converges and estimates $\rho$ to be between -0.95 and 0.95. I want to now assess Model B in the presence of missing data, and see whether Model B is beneficial over a URMA for estimating $\beta_j$ and $(\beta_1 - \beta_2)$ in this situation. Of course, as was the case for Model A (see Section 6.4), Model B is only suitable for evidence synthesis when the missing summary statistic is MCAR or MAR in those studies providing only one outcome. I will begin by looking at when some of the summary statistics are MCAR. Recall that in Section 6.4.1 I simulated 1000 $n = 50$ and $n = 10$ datasets from Model A where $\tilde{Y}_{i1}$ was available for all studies but $\tilde{Y}_{i2}$ was MCAR for 50% of the studies. Now, alongside Model A and a URMA, these simulated datasets were also analysed using Model B and Model A-zero, and the model results for $\tilde{\beta}_j$ and $(\beta_1 - \beta_2)$ were compared, again using the comparison criteria in Section 8.3.3, with $\tilde{\beta}_2$ of particular interest.

For simulation Setting (i) (where $\tau^2_j$ was on average much larger than the $s^2_y$s), the results from both Model B and Model A-zero were extremely similar to those results from Model A for both $n = 50$ and $n = 10$ (hence, simply see the Model A versus URMA results presented previously in Table 6.3(ii) of Section 6.4.1). This is perhaps expected given the relatively large between-study variation, and that Model A itself is not greatly affected by bias because the impact of $\rho_\beta$ being estimated as 1 or -1 is very small in this situation (see discussion in Section 5.6.4). Importantly, the simulation results therefore indicate that Model B is particularly beneficial over a URMA for estimating $\beta_j$ and $(\beta_1 - \beta_2)$ in this simulation setting, as it produces estimates with a smaller standard error, smaller MSE and an improved coverage compared to those estimates from two independent URMAs.
For \( n = 50 \) and simulation Setting (ii) (where \( r_i^2 \) was on average similar to the \( s_{ij}^2 \)'s), Model B again showed considerable benefits over a URMA for outcome \( j = 2 \).

Furthermore, Model B was also far better than Model A-zero in this situation, in particular for the standard error, MSE and coverage of \( \hat{\beta}_2 \) and \( (\hat{\beta}_1 - \hat{\beta}_2) \) (Table 8.9, Appendix C9).

Indeed, Model B was very similar to Model A in terms of \( \hat{\beta}_j \), \( (\hat{\beta}_1 - \hat{\beta}_2) \) and their coverage; the MSE and standard error of the pooled estimates were also similar, albeit slightly higher in Model B on average, but Model B was not subject to any of the problems in Model A when \( \rho_B \) was estimated as 1 or -1 (Table 8.9, Appendix C9 and C10).

Similarly, for \( n = 10 \) and simulation Setting (ii), Model B was again beneficial over a URMA, and similar to Model A (Figure 8.7). However, Model B was now slightly more conservative than for \( n = 50 \), with the standard error and coverage more noticeably higher than those from Model A, but importantly still smaller than those from Model A-zero and a URMA on average. Further analysis looking at the Model B results in comparison to those from Model A where \( \rho_B \) was estimated as 1 or -1 can be seen in Appendix C10 for both \( n = 50 \) and \( n = 10 \). Importantly, Model B was not subject to any of the problems in Model A when \( \rho_B \) was estimated as 1 or -1.

**Figure 8.7:** Model A versus Model B for the \( n = 10 \) MCAR results, described in Section 8.3.4 where convergence was achieved in both models, \( \hat{\rho}_B \neq 1 \) or -1 in Model A, and \(-0.95 < \hat{\rho} < 0.95\) in Model B (the solid line shows line of equality, dotted line shows linear regression line through the points).
Table 8.9: Summary of $\hat{\beta}_1$ and $(\hat{\beta}_1 - \hat{\beta}_2)$ from some of the MCAR simulation results described in Section 8.3.4 for all the simulations (no. sims) out of 1000 that passed the comparison criteria specified in Section 8.3.3. The true parameter values were $\beta_1 = 0$ and $\beta_2 = 2$.

(i) $n = 50$ simulation results, where the 1000 datasets were simulated from Model A for $\tau_j^2$ s similar in size to the $s_{ij}^2$ s on average.

<table>
<thead>
<tr>
<th>Meta-analysis Model</th>
<th>No. sims</th>
<th>Mean estimate</th>
<th>No. of 95% CIs including true value (%)</th>
<th>Mean-Square Error</th>
<th>Mean s.e.</th>
<th>corr($\hat{\beta}_1, \hat{\beta}_2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>URMA</td>
<td>988</td>
<td>-0.00460, 1.995, -2.000</td>
<td>993 (95.4%), 937 (94.8%), 983 (99.5%)</td>
<td>0.0103</td>
<td>0.0215</td>
<td>0.0150</td>
</tr>
<tr>
<td>Model A</td>
<td>988</td>
<td>-0.00466, 1.993, -1.998</td>
<td>932 (94.3%), 934 (94.5%), 925 (93.6%)</td>
<td>0.00959</td>
<td>0.0151</td>
<td>0.0100</td>
</tr>
<tr>
<td>Model B</td>
<td>988</td>
<td>-0.00498, 1.993, -1.998</td>
<td>932 (94.3%), 940 (95.1%), 945 (95.6%)</td>
<td>0.0101</td>
<td>0.0153</td>
<td>0.0099</td>
</tr>
<tr>
<td>Model A-zero</td>
<td>988</td>
<td>-0.00520, 1.995, -2.000</td>
<td>940 (95.1%), 945 (95.6%), 973 (98.5%)</td>
<td>0.0101</td>
<td>0.0165</td>
<td>0.0109</td>
</tr>
</tbody>
</table>

(ii) $n = 10$ simulation results, where the 1000 datasets were simulated from Model A for $\tau_j^2$ s similar in size to the $s_{ij}^2$ s on average.

<table>
<thead>
<tr>
<th>Meta-analysis Model</th>
<th>No. sims</th>
<th>Mean estimate</th>
<th>No. of 95% CIs including true value (%)</th>
<th>Mean-Square Error</th>
<th>Mean s.e.</th>
<th>corr($\hat{\beta}_1, \hat{\beta}_2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>URMA</td>
<td>680</td>
<td>-0.001, 1.994, -1.992</td>
<td>633 (93.1%), 623 (91.6%), 679 (99.9%)</td>
<td>0.0524</td>
<td>0.100</td>
<td>0.0681</td>
</tr>
<tr>
<td>Model A</td>
<td>680</td>
<td>-0.00212, 1.996, -1.998</td>
<td>635 (93.4%), 629 (92.4%), 659 (96.9%)</td>
<td>0.0524</td>
<td>0.0761</td>
<td>0.0473</td>
</tr>
<tr>
<td>Model B</td>
<td>680</td>
<td>-0.00005, 1.996, -1.996</td>
<td>642 (94.4%), 663 (97.5%), 678 (99.7%)</td>
<td>0.0548</td>
<td>0.0768</td>
<td>0.0469</td>
</tr>
<tr>
<td>Model A-zero</td>
<td>680</td>
<td>-0.00151, 1.995, -1.997</td>
<td>642 (94.4%), 663 (97.5%), 678 (99.7%)</td>
<td>0.0521</td>
<td>0.253</td>
<td>0.223</td>
</tr>
</tbody>
</table>

N.B. URMA = univariate random-effects meta-analysis

Table 8.10: Summary of $\hat{\beta}_j$ and $(\hat{\beta}_1 - \hat{\beta}_2)$ from the $n = 50$ MAR simulation results described in Section 8.3.4 that passed the criteria in Section 8.3.3; the 1000 datasets were simulated from Model A where the $\tau_j^2$ s were similar in size to the $s_{ij}^2$ s on average, and $\beta_1 = 0$ and $\beta_2 = 2$.

<table>
<thead>
<tr>
<th>Meta-analysis Model</th>
<th>No. sims</th>
<th>Mean estimate</th>
<th>No. of 95% CIs including true value (%)</th>
<th>Mean-Square Error</th>
<th>Mean s.e.</th>
<th>corr($\hat{\beta}_1, \hat{\beta}_2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>URMA</td>
<td>875</td>
<td>0.507, 1.999, -1.479</td>
<td>0 (0.0%), 824 (94.2%), 10 (1.1%)</td>
<td>0.269</td>
<td>0.0116</td>
<td>0.288</td>
</tr>
<tr>
<td>Model A</td>
<td>879</td>
<td>0.320, 1.981, -1.666</td>
<td>76 (6.6%), 825 (93.9%), 65 (7.4%)</td>
<td>0.111</td>
<td>0.0111</td>
<td>0.125</td>
</tr>
<tr>
<td>Model B</td>
<td>879</td>
<td>0.319, 1.984, -1.661</td>
<td>101 (11.5%), 831 (94.5%), 99 (13.3%)</td>
<td>0.111</td>
<td>0.0114</td>
<td>0.108</td>
</tr>
<tr>
<td>Model A-zero</td>
<td>869</td>
<td>0.422, 2.002, -1.581</td>
<td>5 (0.5%), 825 (94.9%), 47 (5.4%)</td>
<td>0.185</td>
<td>0.0117</td>
<td>0.189</td>
</tr>
</tbody>
</table>

N.B. URMA = univariate random-effects meta-analysis
8.3.5 Missing at random simulation results for $\hat{\beta}_j$ and $(\hat{\beta}_1 - \hat{\beta}_2)$

Recall that in Section 6.4.2 I simulated 1000 $n = 50$ and $n = 10$ datasets from Model A where $\bar{Y}_{i1}$ was available for all studies but $\bar{Y}_{i2}$ was only available when it was positively signed. Importantly, I knew the missing data was MAR because I had generated all the (missing and known) data from Model A (Section 6.4.2). In addition to fitting Model A and a URMA, I now also fitted Model B to these MAR datasets, and then compared the model results for $\hat{\beta}_j$ and $(\hat{\beta}_1 - \hat{\beta}_2)$ using the comparison criteria specified in Section 8.3.3.

Unfortunately, for the majority of the $n = 10$ simulations Model B, Model A and Model A-zero would not converge because the between-study variation was often small relative to the within-study variation for outcome $j = 1$, due to the missing data. The reasons why this causes non-convergence were explained at the beginning of Section 8.3. However, for $n = 50$ convergence was readily obtained and the key parameter to compare across models was $\hat{\beta}_1$, whose values from the URMA were much higher than the true value of $\beta_1$ across simulations, due to the misrepresentative sample of data available for outcome $j = 1$.

Encouragingly, Model B produced a $\hat{\beta}_1$ much closer to the true value of $\beta_1$ than the URMA on average (Table 8.10); similarly $(\hat{\beta}_1 - \hat{\beta}_2)$ was also much closer to the true value in Model B than in a URMA. The coverage of $\hat{\beta}_1$ and $(\hat{\beta}_1 - \hat{\beta}_2)$ was poor in all models but was best in Model B, i.e. closer to 95%, because Model B has a low MSE but slightly higher standard error on average for $\hat{\beta}_1$, and therefore the confidence intervals are wider and capture $\beta_1$ and $(\hat{\beta}_1 - \hat{\beta}_2)$ more often. Model B was also far superior to Model A-zero, with $\hat{\beta}_1$ and $(\hat{\beta}_1 - \hat{\beta}_2)$ much closer to the true answers with far better coverage, MSE and standard error (Table 8.10).
These results show that, when convergence is achieved and $\hat{\rho}$ is between -0.95 and 0.95, Model B is particularly beneficial (in terms of reduced MSE and more suitable coverage for $\hat{\beta}_j$ and $(\hat{\beta}_j - \hat{\beta}_2)$) over a URMA when there is data which are MAR. This was also true for Model A, however there are many occasions in the MAR results where using Model A is perhaps inadvisable because $\rho_B$ is often estimated as 1 or -1 (Table 8.11). I will now consider an exploratory analysis of the subgroups of MAR results according to the value of $\hat{\rho}_B$ in Model A (Table 8.11).

### Table 8.11: Summary of $\hat{\beta}_j$ from the $n = 50$ MAR simulation results described in Section 8.3.5 for all those simulations (no. sims) passing the comparison criteria specified in Section 8.3.3. The 1000 datasets were simulated from Model A where the $\tau^2_j$'s were similar in size to the $s^2_i$'s on average, $\beta_1 = 0$ and $\beta_2 = 2$. The results are split into subgroups defined by the value of $\hat{\rho}_B$ in Model A.

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>Meta-analysis</th>
<th>Model</th>
<th>No. of 95% CIs including true value (%)</th>
<th>Mean-Square Error</th>
<th>Mean s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL URMA</td>
<td>875</td>
<td>0.507</td>
<td>1.999</td>
<td>0 (0%)</td>
<td>0.269</td>
</tr>
<tr>
<td></td>
<td>879</td>
<td>0.320</td>
<td>1.981</td>
<td>76(6.6%)</td>
<td>0.111</td>
</tr>
<tr>
<td></td>
<td>879</td>
<td>0.319</td>
<td>1.984</td>
<td>101(11.5%)</td>
<td>0.111</td>
</tr>
<tr>
<td></td>
<td>869</td>
<td>0.422</td>
<td>2.002</td>
<td>5(0.5%)</td>
<td>0.185</td>
</tr>
<tr>
<td>ALL URMA</td>
<td>158</td>
<td>0.504</td>
<td>1.996</td>
<td>0 (0%)</td>
<td>0.265</td>
</tr>
<tr>
<td></td>
<td>158</td>
<td>0.287</td>
<td>1.973</td>
<td>22(13.9%)</td>
<td>0.0800</td>
</tr>
<tr>
<td></td>
<td>158</td>
<td>0.287</td>
<td>1.971</td>
<td>149(94.3%)</td>
<td>0.0907</td>
</tr>
<tr>
<td></td>
<td>156</td>
<td>0.402</td>
<td>1.987</td>
<td>1(0.6%)</td>
<td>0.169</td>
</tr>
<tr>
<td>ALL URMA</td>
<td>638</td>
<td>0.509</td>
<td>2.001</td>
<td>0 (0%)</td>
<td>0.272</td>
</tr>
<tr>
<td></td>
<td>642</td>
<td>0.325</td>
<td>1.991</td>
<td>54(8.4%)</td>
<td>0.114</td>
</tr>
<tr>
<td></td>
<td>642</td>
<td>0.325</td>
<td>1.986</td>
<td>75(11.7%)</td>
<td>0.113</td>
</tr>
<tr>
<td></td>
<td>638</td>
<td>0.426</td>
<td>2.010</td>
<td>4 (0.6%)</td>
<td>0.187</td>
</tr>
<tr>
<td>ALL URMA</td>
<td>79</td>
<td>0.496</td>
<td>1.984</td>
<td>0 (0%)</td>
<td>0.257</td>
</tr>
<tr>
<td></td>
<td>79</td>
<td>0.371</td>
<td>1.955</td>
<td>0 (0%)</td>
<td>0.147</td>
</tr>
<tr>
<td></td>
<td>79</td>
<td>0.345</td>
<td>1.959</td>
<td>3 (3.8%)</td>
<td>0.129</td>
</tr>
<tr>
<td></td>
<td>75</td>
<td>0.432</td>
<td>1.984</td>
<td>0 (0%)</td>
<td>0.194</td>
</tr>
</tbody>
</table>

N.B. URMA = univariate random-effects meta-analysis

R.D. Riley, Ph.D. Thesis Chapter 8 282
Firstly, consider the MAR results for which Model A did not estimate $\rho_B$ as 1 or -1, and notice that for these Model B gives the same mean value of $\hat{\beta}_i$ as that from the equivalent analyses in Model A (Table 8.11, Figure 8.8). However, for those datasets where $\rho_B$ was estimated as -1 in Model A, Model B gave a $\hat{\beta}_i$ closer to the true value than Model A on average and, conversely, where $\rho_B$ was estimated as 1 in Model A, Model B gave $\hat{\beta}_i$ further from the true value than Model A on average (Table 8.11, Figure 8.8). These results again show the influence on the Model A pooled estimates of the bias in $\hat{\tau}_j^2$ caused by $\rho_B$ being estimated as 1 or -1. The bias in Model A sometimes causes the pooled estimate to be closer to the true value than Model B, but it can also sometimes cause it to be further away than Model B (see Sections 5.6.4 and 7.5.3). Importantly in real meta-analysis situations one would not know what the true parameter values are and so it is surely more suitable to use the Model B results in this situation (of course assuming that it had converged and that $\rho$ is between -0.95 and 0.95) rather than use potentially biased Model A results where $\rho_B$ is estimated as 1 or -1, which could lead to misleading evidence-based conclusions. This means that even if the $\rho_{w_i}$s are known, Model B can still be useful over and above Model A because there will be some situations where Model A estimates $\rho_B$ as 1 or -1 (which may produce biased pooled results, see Section 5.6.4) but Model B estimates $\rho$ between -0.95 and 0.95 (which does not appear to cause bias according to the Model B simulations performed in Section 8.3).
Figure 8.8: Values of $\hat{\beta}_j$ and $\text{var}(\hat{\beta}_j)$ from Model A versus those from Model B obtained from the $n = 50$ MAR simulation results described in Section 8.3.5 for those that met the comparison criteria specified in Section 8.3.3. The 1000 simulated datasets were from Model A where the $\tau_j^2$ s were on average similar to the $s_j^2$ s. The results are split into subgroups defined by the $\hat{\rho}_B$ value in Model A (the solid line shows line of equality, dotted line shows linear regression line through the points).

(i) $\hat{\rho}_B \neq 1$ or $-1$ in Model A

(ii) $\hat{\rho}_B = 1$ in Model A

(iii) $\hat{\rho}_B = -1$ in Model A
8.3.6 Some negative correlation simulation results for Model B

From the simulations performed so far in Sections 8.3.1 to 8.3.5, Model B would seem to be a suitable BRMA model as long as \( \hat{\rho} \) is between -0.95 and 0.95 (Section 8.4 will clearly summarise the reasons why it is suitable). However, the simulations previously considered have only assessed datasets generated from Model A where \( \rho_{wi} = 0.8 \) for all studies and \( \rho_y = 0.8 \). Hence, to demonstrate that Model B is also a suitable BRMA model when there are some negative correlation values and when the \( \rho_{wi} \)s are not the same for each study (i.e. \( \rho_{wi} \neq \rho_y \)), consider now the two additional simulation scenarios where 1000 datasets were generated from Model A as follows:

(a) complete-case data for \( j = 1 \) and \( j = 2 \) for \( n = 50 \), where \( \rho_{wi} = 0.8 \) for 25 studies, \( \rho_{wi} = -0.8 \) for 25 studies, \( \rho_y = 0.8 \), \( \beta_1 = 0 \), \( \beta_2 = 2 \), and \( \tau_j^2 = 0.25 \), on average similar to the \( s_j^2 \)s (which were as described in Section 5.2.1).

(b) as for (a) except data for \( j = 2 \) was MCAR from 25 of the 50 studies (with 13 of these relating to \( \rho_{wi} = -0.8 \) and 12 relating to \( \rho_{wi} = 0.8 \)).

These two scenarios have not thus far been considered for either Model A or Model A-zero in this thesis, and assessments of Model A-zero are particularly interesting for (a) and (b) because the \( \rho_{wi} \)s are on average zero across studies. Hence, Model B, Model A, Model A-zero and two independent URMAs were applied to each of the 1000 datasets generated in each of scenarios (a) and (b), and the results are shown in Table 8.12 for those simulations that met the comparison criteria specified in Section 8.3.3.
Table 8.12: Summary of the $\hat{\beta}_1$ and $\left(\hat{\beta}_1 - \hat{\beta}_2\right)$ simulation results from the two scenarios described in Section 8.3.6 for all those simulations (no. sims) out of 1000 that passed the comparison criteria specified in Section 8.3.3. The true parameter values were $\beta_1 = 0$ and $\beta_2 = 2$.

(a) $n = 50$, complete case data where data were generated from Model A using $\rho_{wi} = 0.8$ for 25 studies, $\rho_{wi} = -0.8$ for 25 studies, $\rho_{i} = 0.8$ and $\tau^2_j$s on average similar to the $s_j^2$s

<table>
<thead>
<tr>
<th>Meta-analysis Model</th>
<th>No. sims</th>
<th>Mean Estimate</th>
<th>No. of 95% CIs including true value (%)</th>
<th>Mean-Square Error</th>
<th>Mean s.e.</th>
<th>$\text{corr}(\hat{\beta}_1, \hat{\beta}_2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>URMA</td>
<td>1000</td>
<td>-0.0031</td>
<td>2.000</td>
<td>-2.003</td>
<td>954 (55.4%)</td>
<td>949 (94.9%)</td>
</tr>
<tr>
<td>Model A</td>
<td>1000</td>
<td>-0.0025</td>
<td>2.001</td>
<td>-2.003</td>
<td>951 (55.1%)</td>
<td>957 (95.7%)</td>
</tr>
<tr>
<td>Model B</td>
<td>1000</td>
<td>-0.0024</td>
<td>2.001</td>
<td>-2.004</td>
<td>948 (94.8%)</td>
<td>942 (94.2%)</td>
</tr>
<tr>
<td>Model A-zero</td>
<td>1000</td>
<td>-0.0025</td>
<td>2.001</td>
<td>-2.004</td>
<td>953 (95.5%)</td>
<td>945 (94.5%)</td>
</tr>
</tbody>
</table>

N.B. For Model A there were 87 simulations which estimated $\rho_{i} = 1$ but none gave $-1$; for Model A-zero there were 320 simulations which estimated $\rho_{i} = 1$ but none gave $-1$.

(b) as for (a) except data for $j = 2$ was missing completely at random from 25 of the 50 studies

<table>
<thead>
<tr>
<th>Meta-analysis Model</th>
<th>No. sims</th>
<th>Mean Estimate</th>
<th>No. of 95% CIs including true value (%)</th>
<th>Mean-Square Error</th>
<th>Mean s.e.</th>
<th>$\text{corr}(\hat{\beta}_1, \hat{\beta}_2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>URMA</td>
<td>997</td>
<td>-0.0030</td>
<td>1.994</td>
<td>-1.997</td>
<td>951 (95.4%)</td>
<td>936 (93.9%)</td>
</tr>
<tr>
<td>Model A</td>
<td>997</td>
<td>-0.0036</td>
<td>2.000</td>
<td>-2.003</td>
<td>947 (95.0%)</td>
<td>925 (92.8%)</td>
</tr>
<tr>
<td>Model B</td>
<td>997</td>
<td>-0.0029</td>
<td>1.996</td>
<td>-1.999</td>
<td>949 (95.2%)</td>
<td>945 (94.8%)</td>
</tr>
<tr>
<td>Model A-zero</td>
<td>997</td>
<td>-0.0028</td>
<td>1.998</td>
<td>-2.000</td>
<td>953 (95.6%)</td>
<td>946 (94.9%)</td>
</tr>
</tbody>
</table>

N.B. For Model A there were 198 simulations which estimated $\rho_{i} = 1$ but none gave $-1$; for Model A-zero there were 401 simulations which estimated $\rho_{i} = 1$, and 3 estimated $\rho_{i} = -1$.

URMA = univariate random-effects meta-analysis

R.D. Riley, Ph.D. Thesis Chapter 8 286
Scenario (a) results

For scenario (a) there is complete-case data, and the results again show that on average there is negligible benefit for estimating $\beta_j$ using a BRMA model rather than two independent URMAs (Table 8.12); for example, the precision, MSE, and coverage of $\hat{\beta}_j$ across simulations are very similar in Model A, Model A-zero, and Model B compared to the URMA. The clear benefit of the BRMA models over URMA is again in the estimate $(\hat{\beta}_1 - \hat{\beta}_2)$. The coverage of $(\hat{\beta}_1 - \hat{\beta}_2)$ from the URMA is too high (98.5%) because $corr(\hat{\beta}, \hat{\beta}_2)$ is not incorporated, unlike in the three BRMA models. Interestingly, Model A-zero gives the most appropriate coverage for $(\hat{\beta}_1 - \hat{\beta}_2)$ (94.7% compared to 93.2% for Model A and 95.8% for Model B) and it also has a much larger precision and smaller MSE of $(\hat{\beta}_1 - \hat{\beta}_2)$ than the URMA. The suitability of Model A-zero here is perhaps because the true $\rho_{wi}$s are on average zero across studies, and thus the assumption that $\rho_{wi} = 0$ is appropriate on average. However, it is surprising that Model A-zero appears to outperform Model A (where the $\rho_{wi}$s are known) in terms of the coverage of $(\hat{\beta}_1 - \hat{\beta}_2)$, as this has not previously occurred in other simulation settings (e.g. see Table 8.8). This poor coverage for Model A is likely to be caused by those 87 occasions where Model A estimates $\rho_b$ as 1 because when removing these the coverage of the remaining simulations was 94.1%, substantially better than before. This indicates that given large differences in the $\rho_{wi}$s, one needs to be particularly cautious about using Model A when $\hat{\rho}_b$ is equal to either 1 or −1.

Despite the good statistical properties of Model A-zero, there were unfortunately 320 occasions where $\rho_b$ was estimated as 1 by this model, and previous findings indicate a need to be cautious about using such results (e.g. see Section 8.3.5). Furthermore, in reality one would not know what the true $\rho_{wi}$s were and would therefore not know how suitable
the assumption that $\rho_{w1} = 0$ was for Model A-zero; in other settings, for example where the true $\rho_{w1}$s are all highly positive (e.g. see Table 8.9), Model A-zero does not perform as well. Hence, given that in Model B one does not need to specify the $\rho_{w1}$s, the results for scenario (a) show that, where $\tilde{\rho}$ is between $-0.95$ and $0.95$, Model B is perhaps the most suitable of all the meta-analysis models to estimate $(\beta_1 - \beta_2)$. The coverage of $(\tilde{\beta}_1 - \tilde{\beta}_2)$ is $95.8\%$ in Model B, closer to the $95\%$ than both the URMA coverage of $98.5\%$ and the Model A coverage of $93.2\%$. Furthermore, the $(\tilde{\beta}_1 - \tilde{\beta}_2)$ results from Model B have a precision and MSE that are on average only slightly inferior to Model A-zero (Table 8.12), but importantly it does not have the problem estimating the between-study correlation as $1$ that is common in the Model A-zero results. The simulation results from scenario (a) therefore provide further evidence that Model B is a suitable BRMA model where $\rho$ is between $-0.95$ and $0.95$ (e.g. as compared to a URMA it obtains a more reliable estimate of $(\beta_1 - \beta_2)$, with a coverage much closer to $95\%$), and that it may be more appropriate than Model A even when the $\rho_{w1}$s are known (e.g. the Model B results gave a coverage for $(\tilde{\beta}_1 - \tilde{\beta}_2)$ closer to $95\%$ than Model A, whilst Model A also often estimated $\rho_B$ as $1$).

Scenario (b) results

Similar conclusions regarding the estimation of $(\beta_1 - \beta_2)$ can be drawn for scenario (b) as for scenario (a). In particular, the coverage is closest to $95\%$ in Model A-zero (95.2\%) followed by Model B (96.2\%), whereas the coverage is again quite poor for both Model A (91.5\%) and the URMA (98.0\%). However, despite the good statistical properties of Model A-zero, for 404 of the simulations Model A-zero estimated $\rho_B$ as either $1$ or $-1$ and one would need to be cautious about using such results in practice (see Section 5.6.4). Perhaps of most interest for scenario (b) is the pooled estimate for outcome $j = 2$ because there are $25$ of the $50$ studies which do not report this outcome. All the BRMA models allow the
‘borrowing of strength’ from the $j = 1$ outcome available from all 50 studies, and thus they all reduce the MSE and increase the precision of $\tilde{\beta}_2$ compared to the URMA (Table 8.12).

Model B and Model A-zero also give appropriate coverage for $\tilde{\beta}_2$ (94.8% and 94.9% respectively). However, the coverage of $\tilde{\beta}_2$ from Model A is again quite poor (92.8%). This poor coverage is likely to be caused by the 198 occasions where Model A estimates $\rho_b$ as 1 because when removing these the coverage of the remaining simulations was 94.5%, substantially better than before. This therefore indicates that one needs to be cautious about using Model A given large differences in the $\rho_w$ s across studies and a $\tilde{\rho}_B$ equal to either 1 or $-1$. Model A-zero also has 404 simulations for which $\tilde{\rho}_B$ is equal to either 1 or $-1$, however despite this problem the coverage, MSE and standard error of $\tilde{\beta}_2$ from Model A-zero appear appropriate for this particular scenario (Table 8.12). This is again possibly because the true $\rho_w$ s are on average zero across studies, and thus the assumption that $\rho_w = 0$ is appropriate on average. Of course, in reality one may not know that this assumption is plausible, and in other settings, for example where the true $\rho_w$ s are all highly positive, Model A-zero does not perform as well in terms of coverage and MSE (e.g. see Table 8.9), and can be greatly affected by a $\tilde{\rho}_B$ equal to 1 or $-1$ (e.g. see Table 8.2). Conversely, where $\tilde{\rho}$ is between $-0.95$ and 0.95, Model B has appropriate coverage, and smaller MSE and standard error for $\tilde{\beta}_2$ than a URMA, but it does not require any assumptions to be made about the $\rho_w$ s and it is not subject to the problem of $\tilde{\rho}_B$ being estimated as 1 or $-1$. Hence, the results for scenario (b) emphasise again the suitability of Model B for BRMA (e.g. as it produces a coverage of $\tilde{\beta}_2$ close to 95%, unlike Model A), and, given data MCAR and a $\tilde{\rho}$ between $-0.95$ and 0.95, Model B will on average obtain $\tilde{\beta}_j$ s with smaller MSE and standard error than a URMA.
8.4 Discussion of the major benefits and limitations of Model B

8.4.1 The major benefits of Model B identified by the simulations

During Sections 8.2 and 8.3 I have introduced and assessed Model B, an alternative model for BRMA, and have identified through a variety of different simulations that, when it converges and \( \hat{\rho} \) is between -0.95 and 0.95:

(i) Model B produces very similar values of \( \hat{\beta}_j \) and \( \hat{\beta}_1 - \hat{\beta}_2 \) to those from Model A where \( \hat{\rho}_B \) is not equal to either 1 or -1 (e.g. see Figure 8.2(i)).

(ii) The estimate of \( \text{var}(\hat{\beta}_j) \) from Model B is similar but generally slightly higher (i.e. more conservative) than \( \text{var}(\hat{\beta}_j) \) from Model A (e.g. see Table 8.9).

(iii) Model B is not subject to the potential bias affecting the parameter estimates in Model A when \( \hat{\rho}_B \) is equal to either 1 or -1 (e.g. see Table 8.11).

Model B is therefore beneficial over two independent URMAs for the same situations I identified Model A to be beneficial over URMA when it does not estimate \( \rho_B \) as 1 or -1 (see Sections 6.4, 6.5 and 6.6). In these situations, Model B is potentially more suitable than URMA for forming conclusions about the true values of \( \beta_j \) and \( \beta_1 - \beta_2 \). In particular, where \( \hat{\rho} \) is between -0.95 and 0.95 then:

(a) where some data is MCAR, Model B will on average obtain \( \hat{\beta}_j \) s with greater precision and smaller MSE than the \( \hat{\beta}_j \) s from two independent URMAs (see Section 8.3.6).

(b) where some data is MAR, Model B will on average obtain \( \hat{\beta}_j \) s with smaller MSE and more appropriate coverage (i.e. closer to 95%) then two independent URMAs (see Section 8.3.5).
(c) in terms of \( \bar{\beta}_1 - \bar{\beta}_2 \), Model B is clearly always preferred (i.e. in complete-case, MCAR or MAR situations) to URMA because, unlike in a URMA, it allows the 
\[ \text{corr}(\bar{\beta}_1, \bar{\beta}_2) \]
to be estimated and therefore a more appropriate estimate of 
\[ \text{var}(\bar{\beta}_1 - \bar{\beta}_2) \]
is obtained (see Tables 8.8 to 8.10).

It is, however, also worth noting that, as for Model A, there is only negligible benefit of Model B over a URMA for estimating \( \beta_j \) when there is complete-case data, as in this situation the reduction in MSE and gain in precision of \( \bar{\beta}_j \) are both very small on average (see Table 8.8).

Most important of all, however, is that Model B will be a readily applicable BRMA model in real meta-analysis situations because it does not require the within-study correlations (i.e. the \( \rho_{w_i} \)'s in Model A) to be known; indeed, the only information required to fit Model B is that which would have been required for two independent URMAs. It would also appear far superior in general (e.g. in terms of MSE and coverage of the pooled estimates) than Model A-zero (e.g. see Table 8.9), another method currently used to apply BRMA when the \( \rho_{w_i} \)'s are unknown in Model A. Furthermore, as long as \( \hat{\rho} \) is between -0.95 and 0.95, Model B is not subject to the potential bias Model A or Model A-zero has when \( \hat{\rho}_{B} \) is equal to either 1 or -1 (see Table 8.11). This means that even if the \( \rho_{w_i} \)'s are known, Model B can still produce more suitable (i.e. least biased) evidence-based results than Model A because there will be some situations where Model A estimates \( \rho_{B} \) as 1 or -1 (which may lead to biased pooled results, see Section 5.6.4) but Model B estimates \( \rho \) between -0.95 and 0.95 (which does not lead to biased results, see Section 8.3.5).
8.4.2 The major limitations of Model B identified by the simulations

Model B is not without its limitations, and these also need to be clearly discussed here. Firstly, if one is interested in a measure of the between-study variation, Model B will not be suitable because it does not allow the $\tau_j^2$ s of Model A to be estimated (see Section 8.2.1). Furthermore, Model B fails to achieve convergence when the within-study variation is much larger than the between-study variation (see Section 8.3). Given this, Model B is not a suitable model for a bivariate fixed-effects meta-analysis, because for this approach it really is essential to know the within-study correlations.

Another limitation of Model B is that, when it does converge, it may occasionally give a $\hat{\rho}$ very close (but not equal) to 1 or −1. This can cause biased and potentially misleading values of $\hat{\beta}_j$ and $\text{var}(\hat{\beta}_j)$ (see Section 8.3.2), similar to the problem of $\hat{\rho}_b$ being equal to either 1 or −1 in Model A (see Section 5.6.4), and happens most often when the within-study variation is much larger than the between-study variation. However, it is important to note that the problem of a $\hat{\rho}$ being very close to 1 or −1 happens less frequently than the problem of $\hat{\rho}_b$ being equal to either 1 or −1 in Model A (see Section 8.3.2).

Another important point worth emphasising is that, as for Model A, Model B is only applicable when the missing summary statistics are either MCAR or MAR in those studies providing only one outcome. Model B is therefore not a suitable model for evidence synthesis when the missing summary statistics are NMAR in these studies, and furthermore Model B does not take into account those studies for which neither outcome were available. These issues will be considered further in Sections 8.7.3 and 8.8, and also Chapter 9.
8.4.3 Discussion on the choice of -0.95 to 0.95 as a ‘safe’ range for $\hat{\rho}$

For my analyses, I only evaluated Model B results where $\hat{\rho}$ was between -0.95 and 0.95 because there did not appear to be any misleading Model B results within this range. This range was a subjective choice based on the simulation results observed and the finding that a $\hat{\rho}$ very close to 1 or -1 causes the analytic solutions for the parameter estimates to have denominators very close to zero (Section 8.3.2). However, the use of this restricted range poses the question of what to do when $\hat{\rho}$ is at the extremes of the range or just outside it?

From my simulations there did not appear to be any misleading answers arising from those occasions where $\hat{\rho}$ was very close to -0.95 or 0.95, and so it is possible that my range chosen was conservative, with some $\hat{\rho}$ values outside the range potentially also producing valid results.

The simulations that gave most problems were those where $\hat{\rho}$ was larger than 0.99, again a consequence of the denominators in the analytic solutions being virtually zero (Section 8.3.2). In reality, there are perhaps very few situations where the true underlying correlation will actually be as high as 0.99 or -0.99. However, if one fits Model B given a small number of studies then the problem of flat profile likelihoods may cause $\hat{\rho}$ to be close to 1 or -1 (see Section 5.6.2), because the model may struggle to estimate $\hat{\rho}$ given the small number of $\bar{Y}_{ii}$s and $\bar{Y}_{i2}$s available. The problem of flat profile likelihoods is likely to be most common where the between-study variation is much smaller than the within-study variation, and indeed the model will often not converge in this situation (see Section 8.3). However, when the number of studies is large there is less chance of a flat likelihood and in this situation it will be less common to observe a $\hat{\rho}$ very close to 1 or -1.

Where one uses Model B and obtains a $\hat{\rho}$ at the extremes or outside of the range -0.95 to 0.95, a sensible next step would be to assess the robustness of the Model B results given
slight changes in the value of $\hat{\rho}$. For example, if Model B gave $\hat{\rho}$ equal to 0.98, one could compare the results to those from Model B when $\hat{\rho}$ was fixed at 0.90, 0.95, 0.96 and 0.97. If Model B was truly tending to misleading values at $\hat{\rho} = 0.98$, due to the analytic solutions dividing by a number very close to zero, then small reductions in the value of $\hat{\rho}$ (e.g. to 0.97, 0.96, 0.96 and 0.90) should demonstrate large discrepancies to those parameter estimates and standard errors obtained at $\hat{\rho} = 0.98$. Figure 8.6 in Section 8.3.2 demonstrates this clearly, with large discrepancies in parameter estimates for very small increases in $\hat{\rho}$ as it nears 1; for example, when $\hat{\rho}$ equals 0.99 $\hat{\beta}_1$ is approximately -4.5 but when $\hat{\rho}$ equals 0.9 $\hat{\beta}_1$ is approximately -1.8, with the large difference due to the 'dividing by zero' problem when $\hat{\rho}$ equals 0.99.

In such situations where the Model B results are highly sensitive to small changes in $\hat{\rho}$, one needs to be extremely cautious about using the Model B results because they could be biased and misleading, something from which evidence-based conclusions should clearly not be made. On the other hand, if a sensitivity assessment showed consistent parameter estimates and consistent standard errors for small changes in $\hat{\rho}$, then this would clearly strengthen the appropriateness of using the Model B results. Examples of these situations are given in Section 8.5.

Given my simulation results, I recommend a sensitivity assessment of Model B whenever $\hat{\rho}$ is at the extremes or outside of the range -0.95 to 0.95. Some readers may wish to shorten this range further (e.g. from -0.9 to 0.9) to be even more cautious if they consider this appropriate. However, the range should certainly not be widened unless proper analytical reasoning is provided, something that would make interesting further research (see Chapter 10).
8.5 Application of Model B to datasets from the neuroblastoma prognostic marker review and other clinically relevant situations

The suitability of Model B as a BRMA model is particularly appealing to those meta-analysts of prognostic marker studies, which are clearly hampered by the problems of missing outcomes, missing summary statistics and unavailable within-study correlations. Indeed application of BRMA to the OS and DFS HRs from the neuroblastoma review has not been possible so far, however Model B offers the chance to rectify this. In this section I will appropriately apply Model B to the prognostic marker datasets for neuroblastoma, and I will also consider three other applications of Model B to real clinical datasets outside the prognostic marker field. These examples will show some of the potential applications of Model B, whilst emphasising the key benefits and limitations of the approach that were established in Sections 8.2 to 8.4.

8.5.1 Joint modelling of OS and DFS hazard ratios

The results when applying Model B to the OS and DFS HR estimates from the neuroblastoma review of Chapter 2 are shown in Table 8.13 for MYCN, chromosome 1p (Ch1p) and multi-drug resistance (MDR). I only considered these markers because they were the only ones for which at least four studies gave a HR for both OS and DFS. I deemed that 4 studies was the least acceptable number of studies for which \( \rho \) could be estimated sensibly, although this was a subjective choice. Given that the assumption of MCAR or MAR in those studies only providing one outcome was difficult to verify for the prognostic marker datasets (see Section 6.3.1), the Model B results in Table 8.13 should only be treated as a sensitivity analysis to assess how the URMA results would change if the MCAR or MAR assumption was indeed correct. In Chapter 9, sensitivity analyses will be introduced to assess the impact of missing data when it is assumed NMAR, and indeed when dissemination bias is assumed to be influencing which summary statistics are available.
Table 8.13: Pooled OS and DFS loge(HR) results from Model B and a univariate (URMA) random-effects meta-analysis for markers MYCN, Ch1p and MDR in neuroblastoma.

<table>
<thead>
<tr>
<th>Marker</th>
<th>n</th>
<th>Model</th>
<th>DFS</th>
<th>OS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\hat{\beta}_1$ (s.e.)</td>
<td>$\hat{\tau}_1^2$</td>
</tr>
<tr>
<td>MYCN</td>
<td>81</td>
<td>URMA</td>
<td>1.478 (0.127)</td>
<td>0.386</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Model B</td>
<td>1.474 (0.113)</td>
<td>-</td>
</tr>
<tr>
<td>Ch1p</td>
<td>15</td>
<td>URMA</td>
<td>1.338 (0.344)</td>
<td>0.303</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Model B</td>
<td>1.481 (0.258)</td>
<td>-</td>
</tr>
<tr>
<td>MDR</td>
<td>7</td>
<td>URMA</td>
<td>1.851 (0.276)</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Model B</td>
<td>1.713 (0.227)</td>
<td>-</td>
</tr>
</tbody>
</table>

$Ch1p = \text{chromosome 1p}; MDR = \text{multi-drug resistance.}$

For MYCN, the Model B results strengthen the conclusions that can be drawn from the URMA as the pooled HR estimates barely changed but their standard errors were much smaller (Table 8.13). For Ch1p, the pooled HR estimate for DFS moved even higher than the URMA value, which was due to one study where DFS was missing having a very high OS HR relative to the other OS values across studies. Hence, when Model B utilised the high $\hat{\rho}$ to ‘borrow strength’ for $\hat{\beta}_2$ from the OS values, this DFS pooled estimate moved slightly higher than the original URMA estimate. The standard errors of the Ch1p pooled estimates were also greatly reduced for both OS and DFS (Table 8.13).

The Model B results for MDR were questionable because $\hat{\rho} = 0.994$, very close to 1 (see Section 8.4.3). The high correlation is perhaps not surprising given the small number of MDR studies providing both OS and DFS summary statistics. To assess whether the MDR results were potentially biased or misleading when $\hat{\rho} = 0.994$, I fitted Model B again to the MDR dataset but fixed $\hat{\rho}$ at 0.95 and then also at 0.90, and observed that on both occasions the pooled estimates were very similar to those when $\hat{\rho}$ was 0.994. In particular, the DFS pooled estimate was still substantially smaller than the value from the URMA.

This was due to one of the studies with DFS missing having an OS result very different to
all the other OS results, and thus Model B borrows strength from this extreme OS value to reduce the DFS pooled estimate. Importantly all the MDR DFS and OS results from Model B (for all of the $\hat{\rho}$ equal to 0.9, 0.95 or 0.994 results) concur with the URMA conclusion that MDR is a potentially important prognostic marker in neuroblastoma.

In conclusion, if the MCAR or MAR assumption is indeed plausible for the MDR, Ch1p and MYCN datasets, the Model B results strengthen the URMA results that these markers are very strong indicators of OS and DFS in patients with neuroblastoma. The Model B analysis has, under the assumption of MCAR or MAR, allowed the incorporation of all those studies only providing one of OS or DFS in the estimation of the pooled HR for both outcomes. Furthermore, it has also enabled the correlation between outcomes to be utilised in the estimation procedure and this allows the 'borrowing of strength' to limit the missing data problem, something that is not possible in a URMA. The Model B results strengthen the URMA conclusion that the role of MDR, Ch1p and MYCN in clinical practice should be researched further, in particular for the development of treatment strategies for specific subgroups of patients and to aid patient counselling (see Sections 2.4.3 and 2.7). In order to help achieve this I would ideally like to extend Model B to a bivariate meta-regression in order to help explain the between-study variation.

To demonstrate that a meta-regression using Model B is possible, I fitted Model B again for the MYCN dataset but included a covariate for 'year of publication – 1995', which may be considered a proxy for treatment received. This type of analysis was originally done for Model A using the Berkey data (see equation (6.3) in Section 6.7.4), and Model B can be extended similarly. The bivariate meta-regression results from Model B indicate that year of publication does not explain a statistically significant amount of the between-study heterogeneity for the MYCN results (Table 8.14). However, there are two major concerns for bivariate meta-regression using Model B in general, and specifically for the prognostic
marker datasets. Firstly, as was shown when Model A was extended to a meta-regression (Section 6.7.4), adding extra covariates to explain the between-study variation may actually prevent Model B from converging because the within-study variation will become increasingly large in relation to the unexplained between-study variation, which unfortunately is the situation where Model B does not converge very easily (Section 8.3). Hence, if most of the between-study variation can be explained in a meta-regression, a univariate rather than bivariate meta-regression will most likely be the most suitable way forward. Other things being equal, only when the between-study variation remains similar or larger than the within-study variation will a bivariate meta-regression be more productive than a univariate approach.

Table 8.14: Pooled OS and DFS loge(HR) MYCN results from Model B, and a bivariate meta-regression using Model B but including an extra covariate for 'year of publication - 1995'.

<table>
<thead>
<tr>
<th>No. of studies</th>
<th>Model</th>
<th>DFS</th>
<th>OS</th>
</tr>
</thead>
<tbody>
<tr>
<td>81</td>
<td>Model B</td>
<td>$\hat{\beta}_1$ (s.e.)</td>
<td>$\hat{\tau}^2_1$</td>
</tr>
<tr>
<td>17 both, 39 just OS, 25 just DFS</td>
<td>Meta-regression using Model B</td>
<td>1.474 (0.113)</td>
<td>-</td>
</tr>
</tbody>
</table>

The second concern for bivariate meta-regression using Model B is that Lambert et al. have shown that meta-regression using summary statistics and aggregated patient-level covariates has low statistical power to detect any relationships between patient characteristics and effect estimates across studies [54]. This was emphasised in the bivariate meta-regression using Model A (Section 6.7.4) and will also be true if Model B is used. For example, for the prognostic marker meta-analyses it would be ideal if study specific patient characteristics such as treatment, age of patient and stage of disease could also be assessed in relation to the prognostic markers. However, this is practically impossible if the IPD is not available because there are often overlapping subgroups of patients across studies; for example, some studies may assess those aged under 1 year of age, whereas others group all ages together. For this exact situation, where patient-level
characteristics are aggregated in a study, meta-regression will have low statistical power to
detect whether age of patients is truly is related to the benefit of the prognostic marker
[54]. For this reason I do not consider bivariate meta-regression any further in this thesis.

8.5.2 Joint modelling of two correlated prognostic markers

Although I have commonly focused on the bivariate meta-analysis approach to ‘borrow
strength’ across two correlated outcomes, other novel applications possible. For example,
consider the potentially prognostic tumour markers assessed in the neuroblastoma review
again. Many of these will be closely related because abnormalities in one marker will be
linked with abnormalities in another; hence, Model B could be used to ‘borrow strength’
across markers. Consider just the outcome of DFS for both MYCN and Ch1p. For Model
B, the DFS HR for MYCN can now be considered as \( j = 1 \) and the DFS HR for Ch1p can
be considered \( j = 2 \), and so Model B can be applied to ‘borrow strength’ from MYCN for
Ch1p and vice-versa. The results are shown in Table 8.15 and the main advantage from
Model B over a URMA is in the DFS pooled estimate for Ch1p, which is able to ‘borrow
strength’ from the large number of studies that only report MYCN results. I repeated this
approach for OS, and found that the MYCN pooled estimate increased and the uncertainty
in the Ch1p estimate decreased in Model B (Table 8.15). All the results strengthen the
conclusion that MYCN and Ch1p are potentially important prognostic markers in
neuroblastoma, worthy of future research.

Table 8.15: Pooled MYCN and chromosome 1p (Ch1p) loge(HR) results from Model B and a
univariate (URMA) random-effects meta-analysis for outcomes DFS and OS

<table>
<thead>
<tr>
<th>Outcome</th>
<th>No. of studies</th>
<th>Model</th>
<th>MYCN</th>
<th>Ch1p</th>
</tr>
</thead>
<tbody>
<tr>
<td>DFS</td>
<td>43 (8 both, 34 just MYCN)</td>
<td>URMA</td>
<td>1.476 (0.127)</td>
<td>0.386</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Model B</td>
<td>1.473 (0.128)</td>
<td>-</td>
</tr>
<tr>
<td>OS</td>
<td>58 (9 both, 47 just MYCN)</td>
<td>URMA</td>
<td>1.627 (0.118)</td>
<td>0.374</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Model B</td>
<td>1.676 (0.127)</td>
<td>-</td>
</tr>
</tbody>
</table>

R.D. Riley, Ph.D. Thesis  Chapter 8  299
The opportunity to 'borrow strength' across markers as well as across outcomes suggests that trivariate and other multivariate extensions to Model B will be useful [150]. A trivariate version of Model A has recently been used by Arends et al. [161], and this enabled multiple correlations to be utilised in the same model, and so equivalent extensions to Model B would make particularly interesting further research.

In Sections 8.5.1 and 8.5.2 I have analysed the prognostic marker datasets from the neuroblastoma review using the BRMA approach of Model B. This has importantly helped to assess the problems of missing outcomes and missing summary statistics, however there are, of course, numerous other methodological problems affecting clinically relevant meta-analysis of this data. These were outlined in Figure 3.2 and include such problems as heterogeneous cut-off levels, indirect HR estimates, and heterogeneous adjustment factors. Meta-analysis methods also need to be developed to limit these problems, and Model B and BRMA methods in general will be placed in context of the remaining methodological issues not addressed by this thesis in Chapter 10.

8.5.3 Joint modelling of sensitivity and specificity

Glas et al. apply a BRMA model within a systematic review of tumour markers used for the diagnosis of primary bladder cancer, where the sensitivity and specificity were of primary interest [167]. For each marker identified, the authors model the logit-transformed sensitivity \( \hat{Y}_{1i} \) and logit-transformed specificity \( \hat{Y}_{12} \) values from each study as a bivariate normally distributed response, but it is unclear whether Model A is used and, if so, whether the \( \rho_{\omega} \) were known. However, a BRMA approach is desirable because it would allow the structural correlation (from the \( 2 \times 2 \) tables) between sensitivity and specificity to be incorporated in the estimation of their pooled values, and thus it would allow the 'borrowing of strength' between sensitivity and specificity estimates, something that is not possible in a URMA.
I was able to apply Model B to the Glas et al. dataset because sensitivity and specificity values from each study were provided in the Glas et al. publication. I chose one marker to assess, bladder tumour antigen, and 6 studies were available, all of which gave both sensitivity and specificity for this marker (thus no assumption about MCAR or MAR is necessary as there is complete-case data here). The Model B results were practically identical to the URMA results, which was perhaps not surprising given there was complete-case data and a relatively low value of $\hat{\rho}$ was obtained (Table 8.16). The results obtained were also similar to those reported in the Glas et al. paper, and relate to a sensitivity of 47.9% and specificity of 80.0%, which Glas et al. suggest are not sufficient for bladder tumour antigen to be of clinical use. This example has shown a novel application of Model B because it is outside the prognostic marker focus of this thesis and also does not involve two log-odds as the responses, as most previous BRMA applications have [148]. However, the example has also emphasised again that given complete-case data there is only negligible benefit of using Model B rather than a URMA to estimate $\beta_j$ (see Table 8.8 for further evidence of this).

Table 8.16: Pooled logit-sensitivity and logit-specificity values from Model B and a univariate (URMA) random-effects meta-analysis for bladder tumour antigen in bladder cancer

<table>
<thead>
<tr>
<th>No. of studies</th>
<th>Model</th>
<th>Logit-sensitivity</th>
<th>Logit-specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$\hat{\beta}_1$ (s.e.)</td>
<td>$\tau^2_1$</td>
</tr>
<tr>
<td>6 (all studies reported both sensitivity and specificity)</td>
<td>URMA</td>
<td>-0.0575 (0.450)</td>
<td>0.859</td>
</tr>
<tr>
<td></td>
<td>Model B</td>
<td>-0.083 (0.451)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Glas*</td>
<td>0 (0.432)</td>
<td>-</td>
</tr>
</tbody>
</table>

* The model used by Glas et al. was a BRMA model but it was difficult to ascertain if it was Model A or another approach; equally it was difficult to ascertain how the confidence intervals were calculated and so the s.e. estimates for Glas et al., above are approximate, based on a normal distribution. The Pearson correlation coefficient for the data was quoted as 0.12 by Glas et al. [167].

8.5.4 Joint modelling of the log-odds for a treatment and control group

Van Houwelingen et al. were the first to suggest Model A as a general framework for BRMA, and the main example used in their paper used the log-odds for an intervention
group and the log-odds for a control group as the bivariate response [148]. The intervention group received a vaccination for tuberculosis whilst those in the control group were not vaccinated. Given each of the 13 studies (which all provided complete-case data) in the meta-analysis used a randomised parallel study design, Van Houwelingen et al. assume all the $\rho_{w}$s are zero and thus use Model A-zero. This assumption is not necessary when using Model B because the $\rho_{w}$s are not needed and the overall correlation ($\rho$) is estimated directly (Section 8.2.1). The results when applying a URMA, Model A-zero and Model B are shown in Table 8.17.

Table 8.17: Pooled log-odds results from Model B, Model A-zero and a univariate (URMA) random-effects meta-analysis for the main dataset used by Van Houwelingen et al. [148]. Model A-zero was the BRMA model used for the original analysis in their paper.

<table>
<thead>
<tr>
<th>No. of studies</th>
<th>Model</th>
<th>Intervention Group</th>
<th>Control Group</th>
<th>Log-odds ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\beta}_1$ (s.e.)</td>
<td>$\hat{\tau}_1^2$</td>
<td>$\hat{\beta}_2$ (s.e.)</td>
<td>$\hat{\tau}_2^2$</td>
</tr>
<tr>
<td>13 (all gave the log-odds in both groups)</td>
<td>URMA</td>
<td>-4.889 (0.344)</td>
<td>1.448</td>
<td>-4.090 (0.459)</td>
</tr>
<tr>
<td></td>
<td>Model A-zero</td>
<td>-4.836 (0.353)</td>
<td>1.550</td>
<td>-4.096 (0.453)</td>
</tr>
<tr>
<td></td>
<td>Model B</td>
<td>-4.855 (0.346)</td>
<td>-</td>
<td>-4.062 (0.463)</td>
</tr>
</tbody>
</table>

Interestingly the between-study variation was very high, much larger than the within-study variation on average, and this is why Model A-zero and Model B give very similar results (see Sections 8.1.1 and 8.4.1), although Model B results give a slightly more conservative precision of the pooled estimates. All the results strongly indicate that the risk of obtaining tuberculosis is greatly reduced when vaccination is used, with the odds ratio from Model B suggesting a 55% reduction in risk compared to the non-vaccinated group (95% CI: 0.31 to 0.67). Importantly, this example clearly shows the benefit of Model B for estimating $(\beta_1 - \beta_2)$, as the standard error of this estimate is reduced by 66% in Model B compared to the URMA, and one does not need to make any strong assumptions regarding the within-study correlations as in Model A-zero.

R.D. Riley, Ph.D. Thesis Chapter 8
8.5.5 Joint modelling of two log-odds ratios

Nam et al. apply Model A-Bayes where the two summary statistics of interest from each study were the log-odds ratio for developing asthma and the log-odds ratio of developing lower respiratory disease, comparing children exposed and unexposed to passive smoking [147]. For their analysis the authors placed a prior distribution on the unknown $\rho_m$'s to limit the fact that they were unknown, but this is again not necessary in Model B because the $\rho_m$'s are not required. In the Nam et al. dataset, only 8 studies gave both $\hat{Y}_{i1}$ and $\hat{Y}_{i2}$, yet 51 studies gave only one of these statistics.

I applied (classically) Model B to this dataset and the results are presented in Table 8.18. Unfortunately, the results obtained were potentially biased because $\bar{\rho}$ was 0.997, very close to 1. Indeed a sensitivity assessment revealed large discrepancies in the $\hat{\beta}_i$'s for small changes in $\bar{\rho}$. This is highlighted by $\hat{\beta}_2$ and especially $\text{var}(\hat{\beta}_1)$ and $\text{var}(\hat{\beta}_2)$ being very different in Model B to those from the URMA, further suggesting that the ‘dividing by zero’ problem could be causing misleading answers (see Section 8.3.2) and that the Model B results should therefore not be trusted here. This example illustrates that Model B does not always overcome the problems in Model A, and in particular when there are only a few studies providing both $\hat{Y}_{i1}$ and $\hat{Y}_{i2}$ relative to the other studies providing only one of these, a $\bar{\rho}$ close to 1 or −1 may occur, potentially causing misleading pooled estimates.

Table 8.18: Pooled log-odds ratio results from Model B and a univariate (URMA) random-effects meta-analysis for the dataset used by Nam et al. [147]; LRD = lower respiratory disease

<table>
<thead>
<tr>
<th>No. of studies</th>
<th>Model</th>
<th>Risk of asthma</th>
<th>Risk of LRD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$\hat{\beta}_1$ (s.e.)</td>
<td>$\tau_1^2$</td>
</tr>
<tr>
<td>59 (8 presented both, 24 just asthma, 27 just LRD)</td>
<td>URMA</td>
<td>1.26 (0.046)</td>
<td>0.036</td>
</tr>
<tr>
<td></td>
<td>Model B</td>
<td>1.184 (0.145)</td>
<td>-</td>
</tr>
</tbody>
</table>
8.5.6 Issues arising from applying Model B in the five examples

The five applications in Sections 8.5.1 to 8.5.5 have shown some common and also some novel ways in which bivariate meta-analysis models, and Model B in particular, are applicable. They have also highlighted how Model B can potentially produce clinically relevant evidence-based results. There will be many other situations where Model B could be applied, and other measures may be of interest such as $\bar{\beta}_1/\bar{\beta}_2$ or baseline risk [148;169]. Meta-analysts are encouraged to consider whether multiple correlated summary statistics are of interest from their own projects, and if Model B is applicable in particular. Of course, the existence of multiple summary statistics does not suggest Model B should always be applied, and meta-analysts needs to clearly justify the approach by considering model assumptions in particular. For instance, those applying Model B when there is missing data must consider whether the assumption of MCAR or MAR is plausible, and where this cannot be ascertained Model B is potentially only suitable as a sensitivity analysis (see Section 8.7.3) [43]. The threat of publication and other related dissemination biases should also be considered (see Chapter 9).

It was particularly encouraging that Model B converged and gave interpretable results (i.e. $\hat{\rho}$ was not very close to 1) for four of the five examples above. A concern of mine was that the situations where Model B does not converge or produces a $\hat{\rho}$ close to 1 or $-1$ were those situations most commonly met in reality. However, this does not appear to be the case and for the datasets I considered the between-study variation was often at least similar in size to the within-study variation, and thus Model B was highly applicable. For Model B to appropriately converge, the examples do suggest, though, that one may need a reasonable number of studies for which both summary statistics of interest were available. In general meta-analyses involve fewer than 10 studies and so it may often be difficult to identify a sufficient number of studies that provide both summary statistics. This problem
may well be the most prominent hurdle to the use of Model B in practice. Indeed, when the
number of studies is very small no BRMA model (be it Model A, Model B or Model A-
zero) can overcome the fact that there are only a few \( \hat{Y}_n \) s and \( \hat{Y}_i \) s to estimate the
parameters. In this situation what is truly needed is additional data or external (prior)
information about the values of the parameters to be estimated. This highlights that a
Bayesian approach to Model B may also be very useful.

8.6 A Bayesian approach to Model B

Model B is expressed in a Bayesian framework (denoted ‘Model B-Bayes’) in equation
(8.10), using ‘vague’ prior distributions similar to those for Model A-Bayes (see equation
(7.1) in Section 7.2) but these can be changed for more informative prior distributions
where prior information is available. Model B-Bayes is also relatively straightforward to
implement in WinBUGS (Appendix C11).

\[
 Model \text{ B-Bayes} \\

\text{Likelihood:} \quad Y_{ij} \sim N(\beta_j, \psi_j) \\
\text{where } \beta_j = \begin{pmatrix} \beta_{i1} \\ \beta_{i2} \end{pmatrix} \\
\psi_j = \begin{pmatrix} s_{i1}^2 + \psi_{i1}^2 & \rho \sqrt{(s_{i1}^2 + \psi_{i1}^2)(s_{i2}^2 + \psi_{i2}^2)} \\ \rho \sqrt{(s_{i1}^2 + \psi_{i1}^2)(s_{i2}^2 + \psi_{i2}^2)} & s_{i2}^2 + \psi_{i2}^2 \end{pmatrix} \\
\text{for } i = 1 \text{ to } n \text{ studies and } j = \begin{cases} 1 & \text{for outcome 1} \\ 2 & \text{for outcome 2} \end{cases} \\
\text{Vague prior distributions:} \quad \psi_j^{-2} \sim \text{Gamma}(0.001, 0.001) \\
\rho \sim \text{Uniform}(-1,1) \\
\beta_j \sim N(0,1000000)
\]

As I discussed for Model A, when there is no prior information available about the
unknown parameters in Model B, in particular for \( \hat{\rho} \), there will often be little benefit of
Model B-Bayes over Model B because one has the additional unwanted problem of the
‘vague’ prior distributions being potentially influential on the posterior results (see Section 7.3). This will be particularly pertinent when there are a small number of studies in the meta-analysis. On the other hand, when prior information is available Model A-Bayes is highly desirable because it can formally incorporate such information through the specification of appropriate prior distributions, which are rightly informative in this situation. Indeed, the incorporation of prior information about \( \rho \) alongside the data is particularly advantageous because it could help prevent \( \hat{\rho} \) being very close to 1 or -1, as sometimes occurs in Model B when there are a relatively small number of studies available (see Sections 8.3.2 and 8.5.5).

To demonstrate these points, consider one of the \( n = 5 \) complete-case simulated datasets (id no. = 809) from Section 5.3. When Model B was applied to this dataset \( \hat{\rho} \) was equal to 0.998, and thus the Model B parameter estimates were not trustworthy. In particular, the pooled estimates and their standard errors appeared somewhat different than those from both a frequentist and Bayesian URMA (Table 8.19). Model B-Bayes was also applied to this dataset using the ‘vague’ prior distributions specified in equation (8.10), except a more informative uniform(0.6, 0.9) prior distribution for \( \rho \) was used. The incorporation of this strong prior information led to more sensible pooled estimates and standard errors than observed from the frequentist Model B, most likely due to the posterior mean of \( \hat{\rho} \) being 0.799, well away from 1. In particular, the standard errors from Model B-Bayes were much larger than the biased and unreliable values from Model B, so if the frequentist Model B results had been used they may have led to strong but potentially wrong conclusions.
Table 8.19: Pooled estimates from a univariate random-effects meta-analysis (URMA), Model B and Model B-Bayes for one of the $n = 5$ complete-case simulated datasets (id = 809) in Section 5.3, where the $\tau^2_j$'s were similar in size to the $s_j^2$'s on average. For the Bayesian analyses a 50000 burn-in was used and 50000 samples were then drawn, with ‘vague’ prior distributions used as specified in equation (8.10), except the prior distribution for $\rho$ was as indicated in the table.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\hat{\beta}_1$ (s.e.)</th>
<th>$\hat{\beta}_2$ (s.e.)</th>
<th>$\tau^2_1$ (s.e.)</th>
<th>$\tau^2_2$ (s.e.)</th>
<th>$\hat{\rho}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>URMA (frequentist)</td>
<td>-0.0250 (0.202)</td>
<td>2.096 (0.268)</td>
<td>0.258</td>
<td>0.537</td>
<td>-</td>
</tr>
<tr>
<td>URMA (Bayesian)</td>
<td>-0.0191 (0.336)</td>
<td>2.085 (0.439)</td>
<td>0.161</td>
<td>0.423</td>
<td>-</td>
</tr>
<tr>
<td>Model B</td>
<td>-0.269 (0.0434)</td>
<td>1.862 (0.0426)</td>
<td>-</td>
<td>-</td>
<td>0.998</td>
</tr>
<tr>
<td>Model B-Bayes</td>
<td>-0.0505 (0.235)</td>
<td>2.055 (0.268)</td>
<td>-</td>
<td>-</td>
<td>0.799</td>
</tr>
</tbody>
</table>

This analysis had demonstrated the advantage Model B-Bayes has over Model B when there is prior information to be included in the meta-analysis. There is still, however, the need to critically assess whether convergence of the parameter estimates has been achieved in Model B-Bayes; indeed, the history of posterior samples does indicate a slight concern for convergence (Figure 8.9), as occasionally some values are sampled quite far away from the mean estimates. This issue may just be indicative of the small sample size of five, and reassuringly I obtained very similar estimates from other Model B-Bayes analyses using different prior distributions for $\psi_j^2$, different starting values and increased burn-in lengths.

Figure 8.9: Samples from the posterior distribution of $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\rho}$ (‘rho’) from Model B-Bayes for one of the datasets (id = 809) from the $n = 5$ complete-case simulations described in Section 5.3. A 50000 burn-in was used and 50000 samples were then drawn. ‘Vague’ prior distributions were used as specified in equation (8.10), except the prior distribution for $\hat{\rho}$ was uniform(0.6,0.9).
8.7 Further discussion of Model B in relation to Model A, with appraisal of my research and suggestions for additional work

8.7.1 Benefits of my research

By obtaining analytic solutions, performing simulations and using real and hypothetical datasets, I have evaluated the benefits and limitations of two BRMA models (Model A and Model B), both against each other and also with a URMA, using both frequentist and Bayesian frameworks. This has enabled me to consider the questions posed by the previous literature and to address a number of the major issues they highlighted (see Section 3.7). This work should therefore help meta-analysts understand when and why BRMA should be considered instead of a URMA, and to appreciate why BRMA results sometimes do or do not differ to those from a URMA. Indeed, the foremost benefit of Model A and Model B is their ability to utilise the correlation between the two summary statistics of interest, and thereby allow the ‘borrowing of strength’ across outcomes when estimating the pooled values of interest. This is not possible in a URMA, and thus the BRMA approach has the potential to produce more appropriate pooled meta-analysis results (i.e. the least biased pooled estimates with smaller MSE and possible larger precision) than two independent URMAs. The BRMA approach is therefore potentially very important for forming evidence-based results that can be used to help inform clinical decisions, public health policies and ultimately, one would hope, the improvement of patient care. This makes BRMA, and Model B in particular, especially relevant for the synthesis of prognostic marker studies, as it allows a joint synthesis between OS and DFS HRs and can thereby incorporate the ‘borrowing of strength’ between these correlated outcomes.

As meta-analysts become more aware of the benefits, I expect to see an increase in the use of BRMA over the coming years. In particular, one of the most fruitful aspects of Model B is its applicability to a wide variety of meta-analysis situations where any two correlated
summary statistics are of interest (see Section 8.5), not just in the prognostic marker field. Indeed, as the benefits I demonstrated for BRMA are likely to generalise to where three or more correlated summary statistics are desired from each study (see Section 6.7.3), the appropriate application of trivariate or higher-order multivariate models may also become increasingly evident in the meta-analysis literature [150;161]. The application and extension of Model B to incorporate three or more correlated summary statistics would therefore make interesting further research. For example, as illustrated in Section 8.5.2, extensions to Model B could also facilitate the 'borrowing of strength' across correlated tumour markers as well as across outcomes, in particular where there are missing results for both markers and outcomes across studies.

8.7.2 Using Model A and Model B in practice, with suggestions for further research

The main benefits and limitations of Model A and Model B are summarised in Figure 8.10. It is safe to say that neither Model A nor Model B is ideal, with both limited by estimation problems associated with the correlation parameters. It is therefore difficult to firmly recommend one model over another in general. Clearly, if the within-study correlations are available from each study then Model A should be primarily applied because this is a true hierarchical model, unlike Model B, and thus it has a more appropriate model structure which can lead to better statistical properties such as increased precision and reduced MSE (see Section 8.2.1) [152]. For the same reasons, if an estimate of the between-study variance is desired then only Model A is suitable as Model B does not provide a direct estimate of this (see Section 8.2.1). Furthermore, when wanting to explain the between-study variation, a bivariate meta-regression using Model A or Model B may be more problematic than using a univariate meta-regression, although neither approach is recommended for assessing patient level characteristics unless IPD is available (see Sections 6.7.4 and 8.5.1) [54].
Figure 8.10: Summary of the key benefits and limitations of Model A and Model B compared with a univariate random-effects meta-analysis (URMA).

The following hold for Model A when $\hat{P}_b$ is not equal to either 1 or -1, and they hold for Model B when $\hat{P}$ is < 0.95 or > -0.95 (see the right hand column for further discussion on this):

**Main findings about Model A and Model B for $\hat{P}_1$ and $\hat{P}_2$:**

- On average the variance and mean-square error (MSE) of $\hat{P}_j$ from Model A or Model B will be less than or equal to the variance and MSE of $\hat{P}_j$ from a URMA.

- **Complete-case data** - there is negligible benefit in using Model A or Model B over two independent URMAs, as precision and MSE are only slightly reduced on average.

- **Missing Data** - Model A and Model B are particularly useful where the missing summary statistic is missing completely at random (MCAR) or missing at random (MAR) in those studies only providing one outcome, as they can improve the precision, MSE and coverage of $\hat{P}_j$ compared to two independent URMAs.

- The benefits (e.g. gain in precision, reduction in MSE) of Model A over a URMA are greater the larger the within- and between-study correlations, and similarly the benefits of Model B are greater the larger the overall correlation.

**Main findings about Model A and Model B for ($\hat{P}_1 - \hat{P}_2$):**

- Model A and Model B are particularly useful for obtaining a more appropriate precision and MSE for ($\hat{P}_1 - \hat{P}_2$) than two independent URMAs, in complete-case, MCAR or MAR situations.

- The var($\hat{P}_1 - \hat{P}_2$) is much more suitable from Model A and Model B as it incorporates corr($\hat{P}_1$, $\hat{P}_2$), which is unavailable from a URMA. The coverage of ($\hat{P}_1 - \hat{P}_2$) is therefore also more suitable (i.e. closer to 95%).

- The benefits of Model A or Model B over a URMA are again greater the larger the correlation estimates.

Difficulties in applying Model A and Model B:

- Model A requires the within-study correlations (i.e. the $\rho_{wi}$ s) to be known from each study, a situation that is highly unlikely in reality. However, in some situations it may be possible to assume the within-study correlations are all zero (see Section 3.6.2).

- Model B does not require the $\rho_{wi}$ s, and only needs the information required to fit a separate URMA for each outcome. However an estimate of $\tau^2_j$ is not possible from Model B.

- Neither Model A nor Model B is suitable if the missing summary statistic is not missing at random (NMAR) in those studies only providing one outcome.

**Estimation problems and caution points for Model A and Model B:**

- It is common for the between-study correlation estimate ($\hat{P}_b$) to be equal to either 1 or -1 from Model A. There are two potential reasons for this (see Section 5.6), and it occurs more frequently when there are a small number of studies, and when the $\tau^2_j$ s are small relative to the $S^2_j$ s.

- Whenever this problem occurs one needs to be extremely cautious about using the Model A results because the $\tau^2_j$ s will on average be upwardly biased, which could potentially lead to misleading values for $\hat{P}_j$, ($\hat{P}_1 - \hat{P}_1$) and their associated variance.

- Convergence of Model B also becomes increasingly difficult for relatively small $\tau^2_j$ s and small number of studies. Values of the overall correlation ($\hat{P}$) very close to 1 or -1 from Model B can also cause misleading pooled estimates. My simulation results for Model B limited to those occasions where $\hat{P}$ was < 0.95 or > -0.95 did not appear to have any problems of bias (see Section 8.3).

- When Model B or Model A are extended to a bivariate meta-regression, it will be very common for the correlation parameters to be close or equal to 1 or -1, or even for non-convergence to occur because the between-study variation will become relatively smaller (see Sections 6.7.4 and 8.5.1).

- A Bayesian approach to Model A or B incorporating pertinent external information about $\hat{P}_b$ or $\hat{P}$ is particularly useful to overcome the difficulty in estimating these parameters (see Chapter 7 and Section 8.6). However, the influence of any ‘vague’ prior distributions on the posterior pooled estimates should always be thoroughly and critically assessed (see Section 6.3).
In many situations, meta-analysts will primarily be interested in the pooled estimates. In this situation, even if the within-study correlations are known in every study, Model B may be necessary in addition to Model A because Model A may still estimate $\rho_B$ as 1 or -1 in this situation, and could thus produce misleading pooled estimates. Indeed, from a purely practical point of view, Model B will likely be the most readily applicable BRMA model because it does not require the within-study correlations to be known, and actually only requires the data one would have required to fit a separate URMA to each outcome (Section 8.2.1). Furthermore, when the simulations in Section 8.4 are placed in context with the literature review in Section 8.1, this thesis suggests that Model B is in general the best of all the currently proposed options for BRMA when the within-study correlations are unknown. For these reasons, Model B is highly relevant for the synthesis of prognostic marker studies, and could thereby help facilitate the most reliable (i.e. least biased) evidence-based results from future systematic reviews in this field, where missing summary statistics and missing outcomes are a major problem (see Section 3.1).

Given that Model B is applicable even when the within-study correlations are unknown, it is unrealistic to therefore expect authors of primary prognostic marker (or other) studies to report the within-study correlation between summary statistics in their published articles, especially as there is already a wealth of information they should report (see Figure 2.11 in Section 2.9). Furthermore, the calculation of this value is non-standard and may be difficult for those (often non-statisticians) who perform the statistical analyses of primary studies, whilst it may also be an irrelevant statistic to the main purposes and messages of the paper itself; this is true for other measures of correlation elsewhere, for example the intra-class correlation coefficient in cluster RCTs [192;193].
I have tried to extend my work where possible, for example to bivariate meta-regression, but I am aware that there are important areas that I have not investigated in this thesis for Model A and Model B. For example, my simulations from Model A assumed that $s_{i1}^2$ and $s_{i2}^2$ are independent (see Section 5.2) but this may not be the case. Indeed, in Chapter 9 I will show that $s_{i1}^2$ and $s_{i2}^2$ are highly correlated for the OS and DFS HRs from the neuroblastoma review, and it would make interesting further research to consider utilising such correlation in a multivariate framework, perhaps by extending Model A or Model B. My simulations of Model A and Model B also assumed that the $s_{y}^2$s were independent to the $\tilde{Y}_{ij}$s, however in many real meta-analysis situations, such as the synthesis of log-odds ratios or indeed log-HRs, the size of $s_{y}^2$ is likely to be related to the size of $\tilde{Y}_{ij}$ (see Section 5.2). The fact that Model A does not utilise the correlation between $s_{i1}^2$ and $s_{i2}^2$, nor that between the $\tilde{Y}_{ij}$ and the $s_{y}^2$, means that it is highly plausible that the conclusions formed about Model A and Model B from the simulations in this thesis (which involved uncorrelated $s_{i1}^2$ and $s_{i2}^2$ and uncorrelated $\tilde{Y}_{ij}$ and $s_{y}^2$) are generalisable (see Section 5.2); it would make interesting further research to explicitly show this, however.

I have also not considered the implications of the assumption that the $s_{i1}^2$s and the $s_{i2}^2$s are known, when of course these within-study variances are only estimates themselves. Hardy and Thompson consider the incorporation of the uncertainty in the $s_{i1}^2$s and $s_{i2}^2$s to be unnecessarily sophisticated for a URMA [45], however whether this is the case for BRMA remains to be seen. The uncertainty of the $s_{y}^2$s could easily be incorporated in the Model B-Bayes framework if required. However, if the uncertainty was acknowledged I would be concerned about the potential additional impact on the estimation of the correlation parameters, for example in Model B the increased uncertainty may cause $\hat{\rho}$ to move
toward 1 or -1 more often, and in Model B-Bayes it could make the posterior pooled estimates even more sensitive to any 'vague' prior distributions specified.

I have applied Model B to the datasets from the neuroblastoma review, and also to some other clinically relevant situations (Sections 8.5.1 to 8.5.5). Another situation of interesting application would be when there is external clinical knowledge to be included alongside the data, a situation that would make Model B-Bayes particularly pertinent (Section 8.6). Other potentially appealing applications of Model B are to situations where baseline risk is of interest (see Section 3.6.3 and also [159; 170-173]), and also when indirect comparisons are needed; for example, Model B could potentially allow pooled estimates of unreported treatment comparisons to be made from the pooled estimates of the individual treatment-effects available [194]. The latter particularly warrants consideration because there is currently very little application of the multivariate meta-analysis approach to indirect comparisons but it could be beneficial [169]. For example, consider Lim et al. who estimate the difference ($\beta_1 - \beta_2$) between the log-relative risk of low dose aspirin ($\beta_1$) and the log-relative risk of medium dose aspirin ($\beta_2$) on graft occlusion after coronary artery surgery [195]. The authors only perform separate univariate meta-analyses because there were no studies reporting both a log-relative risk for low dose aspirin and a log-relative risk for medium dose aspirin, meaning a bivariate meta-analysis approach was not possible as no correlation parameters could be estimated. However, in other situations where some studies report both treatment-effects a bivariate approach would be possible, although one would need to be especially considerate of the model assumptions. For example, is the underlying difference between treatments the same in those studies just examining one of the treatments as in those studies that examine both? This may not be true if a study examining just one of the treatments has slightly different clinical or methodological characteristics that modify the treatment-effect considered. Essentially this is asking
whether the missing data can be assumed MCAR or MAR. This question was asked of the missing OS and DFS HRs in Section 6.3.1 and is particularly important to consider in extreme MAR situations, where the shrinkage in pooled estimates is explicitly linked to the assumption that the observed relationship in those studies providing both outcomes also applies in those studies only providing one of the two outcomes (Section 6.2). The consequences of wrongly using BRMA in this case may be to obtain pooled estimates more misleading than one would have obtained from a URMA (see Table 6.2 in Section 6.2).

In my simulation studies I knew explicitly that the data was truly from a BRMA model because I had generated the datasets directly from Model A. In real-life situations one will not know this and therefore the plausibility of the BRMA model needs to be checked, in particular by considering the model assumptions. Model checking methods for mixed models have currently not been developed in-depth. Brown and Prescott provide some good starting points including some visual checks [152], however it may be more additionally complicated for Model B, given that this is not a true hierarchical model. For the URMA and BRMA models presented I have also made the distributional assumption that the \( Y_{ij} \) s (or equivalently their associated random-error) are normally distributed. This normality assumption may be checked by using normal probability plots of the standardised residuals in each model, with a straight line of unit gradient through the origin indicative that the normal assumption is plausible [45]. However, assuming one would have applied two independent URMA models anyway, the main additional assumption for using Model A or Model B rather than a URMA is that there is some overall correlation between the two summary statistics, with Model A going a step further than Model B by assuming this can be decomposed into within- and between-study correlation. The gains from using Model A and Model B essentially come from including these correlation
parameters, so it is important to consider whether their inclusion is justified. For my OS and DFS HRs there were strong biological and statistical reasons why these summary statistics should be (positively) correlated both overall and also within- and between-studies (see Section 3.2).

One further concern to highlight is that there may be an issue of leverage in BRMA models, where some studies far away from the majority of other studies may have an increased influence on the value of the parameter estimates, especially $\hat{\rho}_B$ in Model A or $\hat{\rho}$ in Model B. For this reason it may be interesting to look at the hat matrix $H$, where $\hat{y} = Hy$ and $H = X(X^T V^{-1} X)^{-1} X^T V^{-1}$, and assess which of the first off-diagonal elements are large and therefore have more influence [196]. However, things are made more complicated in a meta-analysis framework because each study is already forced to have a leverage or weighting based on their within-study variance, so one would need to assess any leverage in addition to this. The issue of leverage in both mixed models and meta-analysis is the subject of on-going research, and is beyond the scope of this thesis; more detailed consideration is given elsewhere [196].

8.7.3 The use of Model A and Model B as a sensitivity analysis

Unless one can justify the assumptions made (e.g. plausibility of MAR), it may be sensible to only use Model A or Model B as a sensitivity analysis to assess the robustness of the URMA conclusions when Model A or Model B are indeed valid. One could argue that the need to perform sensitivity analyses defeats the key benefits of BRMA; for example, one could not use the Model B pooled results directly themselves and so the gain in precision of pooled estimates over URMA would be redundant. However, a sensitivity analysis approach using Model A or Model B can still be very important, perhaps by uncovering
some potential concerns that may otherwise have been ignored and help prevent misleading conclusions from the URMA.

For example, the summary statistics in those studies providing only one outcome may be very different than the summary statistics in those studies providing both outcomes. In this situation, \( \hat{\beta}_j \) from Model B may be very different to the \( \hat{\beta}_j \) from a URMA and this would highlight the need to be cautious about the URMA results, as they may not be valid if Model B was indeed a plausible model. Conversely, if the summary statistics were similar across all types of studies, then the sensitivity approach using Model B could potentially strengthen the URMA conclusions as in this situation \( \hat{\beta}_j \) from Model B may be very similar to the \( \hat{\beta}_j \) from a URMA; this would then make the URMA conclusions valid even assuming Model A was plausible. I am therefore fully supportive of using Model A or Model B (estimation issues aside) as a sensitivity analysis of the URMA results because the sensitivity approach is driven by the desire to make the most reliable (i.e. least biased) evidence-based conclusions [65;197].

8.8 Summary and rationale for subsequent chapters

In this chapter I have reviewed and critically assessed a number of options for BRMA when the within-study correlations are unknown. In particular, I have shown that, where \( \hat{\rho} \) is between -0.95 and 0.95, Model B is a particularly appealing option because it does not require the within-study correlations to be available and it can utilise the overall correlation (\( \hat{\rho} \)) between outcomes to ‘borrow strength’ in order to obtain pooled estimates which have a greater precision, smaller MSE and more suitable coverage than those from two independent URMA.s. Indeed, even when the within-study correlations are known, Model B can still potentially produce less biased evidence-based results than Model A because
there will be some situations where Model A estimates $\rho_B$ as 1 or -1 (which can lead to biased pooled estimates) but Model B estimates $\rho$ between -0.95 and 0.95 (which does not lead to biased estimates).

Model A and Model B are both useful meta-analysis tools, and meta-analysts should consider whether they are applicable for their own research, for example where OS and DFS HRs are to be synthesised from a prognostic marker review. However in practice, rather than deciding whether to apply BRMA models instead of a URMA, meta-analysts of prognostic marker and other studies may be more preoccupied with the threat of publication bias and other related dissemination biases on their analyses. Where the BRMA approach of Model A or Model B is considered plausible, or where they are used as a sensitivity analysis (see Section 8.7.3), the assumption is that the missing summary statistic is either MCAR or MAR in those studies for which only one outcome was available. However, how does one assess the robustness of such BRMA results to the potential impact of dissemination bias from those studies not included in the analysis, i.e. those studies for which neither outcome was available? Furthermore, how does one perform assessments for dissemination bias when all the missing summary statistics are assumed NMAR, even in those studies with only one outcome missing, and can a bivariate meta-analysis framework be useful in this situation?

Such issues regarding how to assess dissemination bias within a bivariate meta-analysis framework now motivate the following, penultimate chapter of this thesis. It is very important to consider such issues because dissemination bias is a very real concern for evidence synthesis of prognostic marker studies (see Sections 2.3.3 and 3.1), and thus for BRMA models to be useful in this field, there must also be supplementary methods available to help assess the potential impact of dissemination bias on the BRMA results.
Chapter 9

USING THE BIVARIATE META-ANALYSIS FRAMEWORK TO HELP ASSESS DISSEMINATION BIAS IN META-ANALYSIS OF PROGNOSTIC MARKER STUDIES

Chapter overview

In this chapter I will consider how to apply current methods for assessing dissemination bias in a meta-analysis when two correlated summary statistics are desired from each study, such as DFS and OS HRs from prognostic marker studies. In particular, I will consider how one may assess the potential impact of dissemination bias on the results of a bivariate random-effects meta-analysis (BRMA) such as Model A or Model B. I will also discuss how and why the bivariate framework may be useful within a selection model to help assess dissemination bias even if Model A and Model B are not applicable because data is not missing at random (NMAR) in those studies where only one of the two desired summary statistics are available.

9.1 Missing summary statistics and the threat of dissemination bias

Missing summary statistics across studies are a common problem for meta-analysis and are caused predominately by poor reporting of primary studies (see Chapter 2) and dissemination bias, which refers to the various ways the reporting and publishing of an individual study can be influenced by the nature and direction of its results [42;86]. Evidence synthesis of prognostic markers is an area particularly vulnerable to these problems (Section 3.1). Publication bias is the most recognised form of dissemination bias and exists because studies that do not generate statistically significant or clinically valuable findings are less likely to be published [41]; other types of dissemination bias include outcome reporting bias, subgroup reporting bias, time lag bias, language bias, citation bias,
and duplicate (multiple) publication bias [119]. There have been numerous methods developed to help assess the potential impact of dissemination bias on the results from a univariate meta-analysis, with the majority of methods motivated by the so-called ‘funnel plot’ of the individual study summary statistics against their standard error or another measure of uncertainty [198]. For example, Duval and Tweedie present a ‘Trim and Fill’ method of testing and adjusting for dissemination bias based on funnel plot asymmetry [199], whilst Egger et al. suggest a hypothesis test to help detect whether dissemination bias is likely to exist [200].

Like dissemination bias, poor statistical reporting is a related problem that restricts evidence-based research by limiting the availability of desired summary statistics. Dissemination bias is most likely to lead to summary statistics that are NMAR, but poor reporting may cause some summary statistics to be MCAR or MAR and also some summary statistics to be NMAR; for the latter situation, the poor reporting is therefore clearly contributing to the overall dissemination bias problem.

When there are missing summary statistics the key concern is that the set of available summary statistics extracted from the literature do not actually reflect the overall evidence-base, i.e. the truth [86]. In particular, when many unavailable summary statistics are NMAR there is a considerable potential for misleading meta-analysis results to be obtained and for wrong evidence-based conclusions to be made. A full assessment of the potential impact of missing summary statistics is therefore clearly important when performing and presenting any meta-analysis so that the least biased and therefore most reliable conclusions can be made.

In this chapter I want to consider how to apply current methods for assessing dissemination bias when two correlated summary statistics are sought from each study. Indeed, for this
situation I have previously used the BRMA approach of Model A and Model B to limit the problem of missing data in those studies where only one of two desired summary statistics were available, an approach which assumed the missing data was MCAR or MAR in these studies. This poses two questions:

(i) If the assumption of MCAR or MAR is true for Models A and B, then is it possible to assess the results from these models for the possible problem of dissemination bias from the remaining studies not incorporated in the meta-analysis, i.e. those studies for which no summary statistics were available?

(ii) If the assumption of MCAR or MAR is not true for Models A and B, then would it still be possible to use a bivariate meta-analysis framework to help assess dissemination bias in this situation, where one would now be concerned that all missing summary statistics from all studies were NMAR?

These are the two questions I want to consider in this chapter, and they are particularly pertinent for meta-analysis of observational studies, such as those for prognostic markers, where the problems of missing data and dissemination bias are likely to be much greater than in RCTs [102]. Hence, to illustrate the methods and issues introduced, I will again use the dataset for MYCN, the most commonly studied prognostic marker from the neuroblastoma review in Chapter 2. This dataset was introduced in Section 3.2 and it contains 42 HRs for DFS and 56 HRs for OS (Table 3.1). One consequence of the limited reporting in the neuroblastoma literature was that OS and DFS HRs were not always both available from an individual primary study; indeed, there were three 3 different primary studies to be considered from my review of MYCN:

(1) Studies for which both OS and DFS HRs were available (17 studies).

(2) Studies for which only OS HRs were available (39 studies), and studies for which only DFS HRs were available (25 studies).
Studies with no summary statistics available for either outcome, i.e. known studies which reported prognostic results or IPD, but not sufficiently to allow either OS or DFS HRs to be obtained (70 studies), and also unidentified studies which have not been published (unknown number of studies).

The possible reasons why studies in (2) only provide one of OS or DFS were discussed in Section 6.3.1, but of most concern is that dissemination bias is the cause, with the missing summary statistics NMAR, and possibly unavailable because they were not clinically or statistically significant (e.g. at the 5% level). Dissemination bias may also have influenced those studies in (3), as the decision to not publish the study, or report results in sufficient detail, may again have been influenced by the summary statistics not being statistically or clinically significant. This threat of dissemination bias means that some or all of the missing summary statistics may have been NMAR, and therefore those OS and DFS HRs available might not form an accurate representation of the overall evidence-base, and thus any subsequent meta-analysis results and conclusions about MYCN might be misleading.

Despite these problems I have still performed univariate and bivariate meta-analyses using the MYCN dataset throughout this thesis. As discussed previously (see Section 2.4.2), this decision was made because of the great need to identify those markers of potential importance so to help prioritise the future research agenda for prognostic markers in neuroblastoma [1]. In order to continue to meet this aim and clarify whether MYCN is potentially important, it is therefore necessary to assess the robustness of the univariate and bivariate MYCN meta-analysis results to the possible impact of dissemination bias. For instance, the Model B analysis for MYCN in Section 8.5.1 assumed the missing summary statistic was either MCAR or MAR in those studies only providing one outcome, but this might not be the case and therefore the impact when such missing summary statistics are assumed NMAR needs to also be considered. I will begin my assessments by applying
current methods proposed for assessing the potential impact of dissemination bias on the results of a univariate meta-analysis, and will then consider how one may apply or extend these assessments alongside or within a bivariate meta-analysis framework.

9.2 Assessing the potential impact of dissemination bias on the results of a univariate random-effects meta-analysis

The univariate random-effects meta-analysis (URMA) results indicate that high levels of the marker MYCN are associated with a statistically significant increased risk of death (56 OS studies: pooled loge(HR) = 1.63, 95% CI 1.39 to 1.87, p<0.0001), and also risk of either death or recurrence of disease (42 DFS studies: pooled loge(HR) = 1.48, 95% CI 1.23 to 1.74, p< 0.0001). These pooled results clearly suggest that MYCN is a potentially important prognostic marker for neuroblastoma, but they are open to criticism given the abundance of missing summary statistics across studies and the underlying threat of dissemination bias. To assess dissemination bias and its potential impact on these URMA results, I used the 'Trim and Fill' method [199], alongside a visual inspection of the funnel plot and the hypothesis test proposed by Egger et al. [200]. I also considered whether study characteristics, and not dissemination bias, could explain the funnel plot asymmetry by looking at subgroups of patients [41].

9.2.1 Visual inspection of the funnel plots

Funnel plots of the DFS and OS estimates against their standard error are shown in Figure 9.1 (a) and (b). The assumption is that these plots should form a funnel shape if there is no dissemination bias present, as estimates from smaller studies will be more widely spread about the mean effect due to larger standard errors. However, the plots for DFS and OS were not indicative of a very funnel-like shape, and asymmetry was apparent, with a gap in the bottom right of each plot. Hence, it appeared that some studies with less positive results than those obtained were potentially missing from my analyses (Figure 9.1).
**Figure 9.1**: Original and 'filled' funnel plots for overall (OS) and disease-free (DFS) survival; the 'filled' studies are those suggested by the Trim and Fill analysis.

(a) DFS  
42 known studies

(b) OS  
56 known studies

(c) DFS  
42 known, 8 'filled' studies

(d) OS  
56 known, 12 'filled' studies

**9.2.2 Egger's test for the presence of dissemination bias**

This method uses a linear regression approach in which the standard normal deviates (study summary statistic (\( \tilde{Y}_j \))/standard error of summary statistic (\( s_{\tilde{Y}} \))) are regressed against precision (\( s_{\tilde{Y}}^{-1} \)) for each outcome \( j \) separately [200]. This approach corresponds to a weighted regression (\( \tilde{Y}_j = \alpha + \beta s_{\tilde{Y}} \)) of summary statistic on standard error (\( s_{\tilde{Y}} \)) where the weights are inversely proportional to the variance (\( s_{\tilde{Y}}^2 \)) of the summary statistic. The degree of asymmetry in the funnel plot, and therefore the potential for dissemination bias, is indicated by the magnitude and statistical significance of the intercept coefficient \( \alpha \).

Applying this test to the summary statistics for each outcome separately produced statistically significant evidence of asymmetry for DFS \( (p = 0.003) \) but not for OS \( (p = \)
0.45. Furthermore, even when the extreme observation in the top right of the funnel plots in Figure 9.1 was removed, Egger’s Test for DFS and OS produced similar p-values to before.

9.2.3 The Trim and Fill method

This method imputes summary statistics from studies considered missing based on funnel plot asymmetry [199]. At first, an iterative algorithm is used to estimate the number of studies considered ‘missing’ from the funnel plot; these studies can broadly be considered to be those with no counterpart on the opposite side of the funnel. This number of studies is then ‘Trimmed’ from the asymmetric outlying part of the funnel, leaving a symmetric remainder, which is then used to produce an ‘adjusted’ estimate of the ‘true centre’ of the funnel using standard meta-analysis techniques. The ‘Trimmed’ studies are then replaced, and their ‘missing’ counterparts imputed or ‘Filled’ as mirror images of the ‘Trimmed’ studies, with the mirror axis placed along the ‘adjusted’ pooled estimate. This last stage is necessary to calculate an ‘adjusted’ standard error and an ‘adjusted’ confidence interval for the ‘adjusted’ pooled estimate. I want to emphasise here that these ‘adjusted’ results are only obtained to assess if and how the univariate meta-analysis conclusions change in light of the potential dissemination bias, and they should not be used on their own but only for a comparison with the univariate meta-analysis results.

I applied the Trim and Fill method using a URMA model to the MYCN estimates in each outcome. There were 8 DFS studies and 12 OS studies, with values of loge(HR) close to and less than zero, estimated as missing (Figure 9.1 (c) and (d)). When these ‘missing’ studies were imputed and incorporated into the URMAs, the pooled estimates were considerably less than those originally calculated for both OS (original pooled loge(HR) = 1.63, ‘adjusted’ pooled loge(HR) = 1.37, 95% CI 1.12 to 1.62, p < 0.0001) and DFS
(original pooled loge(HR) = 1.48, ‘adjusted’ pooled loge(HR) = 1.27, 95% CI 1.01 to 1.53, p<0.0001), but still indicative of MYCN being a potentially important prognostic marker.

The Trim and Fill method emphasises the findings from the visual inspection of the funnel plots, and also from Egger’s Test for DFS, that the OS and DFS funnel plots are asymmetric. Assuming this is caused by dissemination bias, it is likely that a number of studies with a small or negative loge(HR) are potentially missing from my analyses, and that the pooled loge(HR) estimates from the original URMA are both likely to be an overestimate of the true effect for OS and DFS. However, MYCN would still appear to be a potentially important prognostic marker because the pooled results were especially large with confidence intervals still far above a loge(HR) value of 0, both from the original URMA and also from the Trim and Fill sensitivity analysis.

9.2.4 Other possible causes of funnel plot asymmetry beside dissemination bias

There may be other explanations for funnel plot asymmetry other than dissemination bias [201]. Sterne et al. recommend using meta-regression to investigate possible factors causing the between-study heterogeneity of study estimates, as it could be these factors causing the observed asymmetry [41]. For example, the larger summary statistics with larger standard errors could be related to a particular subgroup of patients and by removing this subgroup the funnel plot may become more symmetric. However, Lambert et al. have shown that meta-regression using aggregated patient-level covariates has low statistical power to detect the true underlying relationships between patient characteristics and effect estimates [54]. Unfortunately, as I did not have IPD available for the majority of studies, I could only consider aggregated patient level covariates for a meta-regression of the MYCN data, and subsequently my analyses were fraught with the problems highlighted by Lambert et al. [54]. For example, most studies involved patients above and below 1 year old at diagnosis, although five did only involve patients older than 1 year. Hence, when
considering age as a covariate, the only two groups I could consider for age were (i) those studies just involving patients older than one year, versus (ii) all other studies. However, the majority of studies in (ii) also involved patients older than one year old so the meta-regression results could be very misleading and misinterpreted. For these reasons, I do not present any meta-regression analyses here.

Instead, I did consider funnel plots for subgroups of the *MYCN* studies to assess whether asymmetry was different for each subgroup separately. I looked at subgroups formed by the number of patients in each study (e.g. those with > 50), cut-off level used (e.g. just 10 copies of *MYCN* used), and year of study publication (e.g. those since 1990). These were the only study characteristics for which sufficient information was available from the literature and whose subgroups were distinct (i.e. not overlapping), unlike age and stage of disease for example. There was still considerable evidence of funnel plot asymmetry in all the subgroups considered, for both OS and DFS, and practically the same answers were found from Egger's Test and the Trim and Fill method to before (for an example see Figure 9.2). Hence, although I could only consider a few study characteristics, it appeared that dissemination bias was still the most likely cause of funnel plot asymmetry for the *MYCN* dataset. In an ideal world, and to best guide clinical policy, one would want to consider separate meta-analyses and separate funnel plots for the mutually exclusive subgroups of factors in this situation (e.g. age > 1, stage 4 disease, n > 50). However, this was practically very difficult because of the large number of possible subgroups, the potential for very small numbers of studies in each, and the amount of missing study information. I therefore felt it was sensible to keep all the study estimates together and continue to assess dissemination bias collectively. As the majority of studies included all types of patients (e.g. all ages, all stages), the pooled OS and DFS estimates should therefore be considered to relate to an 'average' individual with neuroblastoma, as previously discussed (Section 2.4.2).
Figure 9.2: Evidence of funnel plot asymmetry after excluding studies with less than 50 patients.

- **DFS (22 studies)**
  - LogHR vs. SE(logHR)
  - Poled logHR = 1.54

- **OS (18 studies)**
  - LogHR vs. SE(logHR)
  - Poled logHR = 1.69

Egger's Test gave p<0.005 for both DFS and OS.

N.B. Of the 42 original DFS studies included, 19 studies involved less than 50 patients and one study was also omitted as the number of patients was unclear. Of the 56 original OS studies included, 38 studies involved less than 50 patients.

### 9.3 A method for assessing the potential impact of dissemination bias on the results of a bivariate random effects meta-analysis

Section 9.2 used a variety of methods to assess dissemination bias in relation to a URMA, and in this section I will now consider how one could apply or extend those methods to help assess the robustness of the results from a BRMA to the threat of dissemination bias.

The Model B analysis for MYCN in Section 8.5.1 assumed that in those studies providing only one of the OS and DFS summary statistics, the other missing summary statistic was either MCAR or MAR. The Model B results again strongly indicated that MYCN was potentially an important prognostic marker in neuroblastoma for both OS (pooled loge(HR) = 1.65, 95% CI to 1.42 to 1.88, p<0.0001), and also DFS (pooled loge(HR) = 1.47, 95% CI 1.25 to 1.70, p< 0.0001). However, for this Model B analysis, MCAR or MAR was indeed only an assumption (see Section 6.3), and furthermore the Model B framework does not consider at all those studies for which neither OS nor DFS HRs were available. Hence, the two important questions (i) and (ii) of Section 9.1 are now posed for this situation. Firstly, if the MCAR or MAR assumption in Model B is true, how does one assess the MYCN
results from Model B for possible dissemination bias from the remaining studies which do not provide either OS or DFS summary statistics? Secondly, if the assumption of MCAR or MAR is not true for Model B, then is it possible to use a bivariate meta-analysis framework to help assess the URMA MYCN results for the possible impact of dissemination bias from all missing data studies, where one would now be concerned that all the missing summary statistics were NMAR?

9.3.1 The 'three-step approach'

I will firstly consider the situation where the assumption of MCAR or MAR is indeed correct for the missing summary statistic in those studies only providing one of the two outcomes. In this situation Model B is suitable and its multivariate structure allows it to inherently 'borrow strength' between OS and DFS summary statistics, thereby allowing potentially more reliable (i.e. least biased) meta-analysis results than a URMA (see Section 8.4.1). However, how can one assess the potential impact of dissemination bias on Model B results? Unfortunately it is not straightforward to immediately apply the methods previously introduced (e.g. Trim and Fill, Egger’s Test) because in Model B the 'borrowing of strength' framework does not directly translate back onto the funnel plots necessary for these methods. For instance, as the 'borrowing of strength' is inherent in the analytic solutions (see Section 8.2.2), Model B does not formally produce predicted estimates of the missing summary statistics and also does not produce predicted estimates of the missing standard errors of the missing summary statistics (this is discussed further in Section 9.3.2). It is therefore not straightforward to incorporate on the funnel plot how Model B is 'borrowing strength' for the missing summary statistics from those studies only providing one of OS or DFS.
In order to apply a BRMA followed by current methods for assessing dissemination bias, one ideally requires the ‘borrowing of strength’ in the bivariate framework to be made explicit on a funnel plot alongside the summary statistics already known, thus then facilitating a dissemination bias assessment such as Trim and Fill. As a suggestion of one possible approach to meet this need, and to illustrate some of the difficulties involved, consider now the following ‘three-step approach’ that does not use Model B directly but uses a similar bivariate framework to obtain predicted estimates and predicted standard errors of the missing summary statistics, which are then back transformed onto a funnel plot alongside the known summary statistics to facilitate dissemination bias assessments such as Trim and Fill and also Egger’s Test.

**Step 1 - Using the observed relationship between OS and DFS estimates**

Assume, as for Model B, that the relationship between OS and DFS summary statistics from the 17 MYCN studies providing both outcomes is the same for studies where only one outcome is available (i.e. MCAR or MAR is assumed). In these 17 studies there was an observed linear relationship between the loge(HR) estimates between outcomes, and similarly there was an observed linear relationship between the standard error of the loge(HR) between outcomes (Figure 9.3). Indeed, assuming normally distributed errors, using simple linear regression to model loge(HR) for OS against loge(HR) for DFS (and vice-versa) suggested a linear relationship was plausible (adjusted $R^2 = 0.79$); similarly for the standard error of the loge(HR) for OS against that for DFS, and vice-versa (adjusted $R^2 = 0.81$) (Table 9.1). I decided to use the standard error rather than loge(standard error) here as it made little difference to the regression models and I wanted to keep the approach as clear as possible; in practice the residuals from a regression of loge(standard error) are more likely to meet the required normality assumption.
Table 9.1: Linear regression equations for the relationship between overall (OS) and disease-free (DFS) survival estimates, with $R^2$ model-fit statistics.

<table>
<thead>
<tr>
<th>Model</th>
<th>Predicted estimate</th>
<th>$= a + \beta \times$ covariate</th>
<th>s.e. ($a$)</th>
<th>s.e. ($\beta$)</th>
<th>$R^2$</th>
<th>Adj-$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i)</td>
<td>log$_e$HR(OS)</td>
<td>0.32 + 0.93 x log$_e$HR(DFS)</td>
<td>0.24</td>
<td>0.12</td>
<td>0.81</td>
<td>0.79</td>
</tr>
<tr>
<td>(ii)</td>
<td>log$_e$HR(DFS)</td>
<td>0.038 + 0.87 x log$_e$HR(OS)</td>
<td>0.25</td>
<td>0.11</td>
<td>0.81</td>
<td>0.79</td>
</tr>
<tr>
<td>(iii)</td>
<td>se(log$_e$HR(OS))</td>
<td>0.11 + 0.94 x se(log$_e$HR(DFS))</td>
<td>0.082</td>
<td>0.11</td>
<td>0.84</td>
<td>0.81</td>
</tr>
<tr>
<td>(iv)</td>
<td>se(log$_e$HR(DFS))</td>
<td>0.021 + 0.88 x se(log$_e$HR(OS))</td>
<td>0.084</td>
<td>0.10</td>
<td>0.84</td>
<td>0.81</td>
</tr>
</tbody>
</table>

s.e. = standard error, HR = hazard ratio, adj = adjusted

Figure 9.3: Relationship between overall (OS) and disease-free (DFS) survival estimates from the 17 studies providing both outcomes

(a) log$_e$(HR): OS versus DFS

(b) se(log$_e$(HR)): OS versus DFS

se = standard error; HR = hazard ratio

For those 64 studies where only one outcome was available (25 only DFS, 39 only OS), I used these observed relationships to predict the missing data for the outcome currently unavailable, an approach that is suggested by Pigott [202]. Using the linear regression equations (i)-(iv) in Table 9.1, I predicted the log$_e$(HR) for OS when only DFS was available (i), and vice-versa (ii), and similarly the standard error of the log$_e$(HR) for OS when only DFS was available (iii), and vice-versa (iv) (see Table 9.2 for predicted estimates). Hence, 39 DFS estimates and 25 OS estimates, previously unavailable, were predicted. For example, for study 18 only DFS was available, but I predicted for OS a log$_e$(HR) of 0.55 (using (i)) with a standard error of 0.38 (using (iii)).
Table 9.2: The 42 disease-free survival (DFS) and 56 overall survival (OS) estimates of the loge(hazard ratio) (loge(HR)) and its standard error (s.e.) for each study, with predicted estimates from the ‘three-step approach’ where applicable (see step 1, Section 9.3.1).

<table>
<thead>
<tr>
<th>STUDY ID</th>
<th>DFS providing both outcomes</th>
<th>OS providing both outcomes</th>
<th>STUDY ID</th>
<th>DFS providing only DFS</th>
<th>OS predicted* DFS predicted*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.11 (0.67)</td>
<td>-0.14 (0.81)</td>
<td>18</td>
<td>0.25 (0.29)</td>
<td>NA</td>
</tr>
<tr>
<td>2</td>
<td>0.30 (0.26)</td>
<td>0.43 (0.81)</td>
<td>19</td>
<td>0.29 (0.59)</td>
<td>NA</td>
</tr>
<tr>
<td>3</td>
<td>0.41 (0.82)</td>
<td>0.67 (0.29)</td>
<td>20</td>
<td>0.52 (0.41)</td>
<td>NA</td>
</tr>
<tr>
<td>4</td>
<td>0.47 (0.53)</td>
<td>0.70 (0.56)</td>
<td>21</td>
<td>0.55 (0.38)</td>
<td>NA</td>
</tr>
<tr>
<td>5</td>
<td>0.76 (0.49)</td>
<td>0.71 (0.63)</td>
<td>22</td>
<td>0.84 (0.26)</td>
<td>NA</td>
</tr>
<tr>
<td>6</td>
<td>1.06 (0.54)</td>
<td>1.32 (0.51)</td>
<td>23</td>
<td>0.93 (0.32)</td>
<td>NA</td>
</tr>
<tr>
<td>7</td>
<td>1.46 (0.41)</td>
<td>1.36 (0.37)</td>
<td>24</td>
<td>1.18 (0.57)</td>
<td>NA</td>
</tr>
<tr>
<td>8</td>
<td>1.64 (0.64)</td>
<td>1.51 (0.48)</td>
<td>25</td>
<td>1.34 (0.51)</td>
<td>NA</td>
</tr>
<tr>
<td>9</td>
<td>1.64 (0.64)</td>
<td>1.54 (0.52)</td>
<td>26</td>
<td>1.43 (0.37)</td>
<td>NA</td>
</tr>
<tr>
<td>10</td>
<td>1.64 (0.51)</td>
<td>1.82 (0.71)</td>
<td>27</td>
<td>1.44 (1.17)</td>
<td>NA</td>
</tr>
<tr>
<td>11</td>
<td>1.70 (0.39)</td>
<td>1.83 (0.47)</td>
<td>28</td>
<td>1.45 (0.57)</td>
<td>NA</td>
</tr>
<tr>
<td>12</td>
<td>1.85 (0.66)</td>
<td>2.08 (0.67)</td>
<td>29</td>
<td>1.52 (0.35)</td>
<td>NA</td>
</tr>
<tr>
<td>13</td>
<td>1.90 (0.46)</td>
<td>2.59 (1.04)</td>
<td>30</td>
<td>1.60 (0.49)</td>
<td>NA</td>
</tr>
<tr>
<td>14</td>
<td>1.90 (0.88)</td>
<td>2.75 (1.10)</td>
<td>31</td>
<td>1.62 (0.42)</td>
<td>NA</td>
</tr>
<tr>
<td>15</td>
<td>2.19 (0.42)</td>
<td>2.90 (1.10)</td>
<td>32</td>
<td>1.77 (0.46)</td>
<td>NA</td>
</tr>
<tr>
<td>16</td>
<td>2.25 (1.06)</td>
<td>2.99 (0.51)</td>
<td>33</td>
<td>1.90 (0.58)</td>
<td>NA</td>
</tr>
<tr>
<td>17</td>
<td>5.70 (1.73)</td>
<td>5.70 (1.73)</td>
<td>34</td>
<td>1.92 (0.34)</td>
<td>NA</td>
</tr>
<tr>
<td>35</td>
<td>2.04 (0.62)</td>
<td>2.21 (0.69)</td>
<td>36</td>
<td>2.19 (0.35)</td>
<td>NA</td>
</tr>
<tr>
<td>37</td>
<td>2.37 (1.00)</td>
<td>2.51 (1.05)</td>
<td>38</td>
<td>2.39 (0.73)</td>
<td>NA</td>
</tr>
<tr>
<td>39</td>
<td>2.50 (0.76)</td>
<td>2.63 (0.82)</td>
<td>40</td>
<td>2.56 (0.55)</td>
<td>NA</td>
</tr>
<tr>
<td>41</td>
<td>2.98 (0.58)</td>
<td>3.07 (0.65)</td>
<td>42</td>
<td>3.29 (0.50)</td>
<td>NA</td>
</tr>
<tr>
<td>43</td>
<td>NA</td>
<td>-0.84 (0.85)</td>
<td>44</td>
<td>NA</td>
<td>0.05 (0.40)</td>
</tr>
<tr>
<td>45</td>
<td>NA</td>
<td>0.73 (0.71)</td>
<td>46</td>
<td>NA</td>
<td>0.76 (0.20)</td>
</tr>
<tr>
<td>47</td>
<td>NA</td>
<td>0.91 (0.66)</td>
<td>48</td>
<td>NA</td>
<td>0.93 (0.27)</td>
</tr>
<tr>
<td>49</td>
<td>NA</td>
<td>0.96 (0.47)</td>
<td>50</td>
<td>NA</td>
<td>1.05 (0.86)</td>
</tr>
<tr>
<td>51</td>
<td>NA</td>
<td>1.16 (1.18)</td>
<td>52</td>
<td>NA</td>
<td>1.22 (0.22)</td>
</tr>
<tr>
<td>53</td>
<td>NA</td>
<td>1.26 (0.49)</td>
<td>54</td>
<td>NA</td>
<td>1.26 (0.38)</td>
</tr>
<tr>
<td>55</td>
<td>NA</td>
<td>1.27 (1.28)</td>
<td>56</td>
<td>NA</td>
<td>1.31 (0.82)</td>
</tr>
<tr>
<td>57</td>
<td>NA</td>
<td>1.52 (0.46)</td>
<td>58</td>
<td>NA</td>
<td>1.54 (0.55)</td>
</tr>
<tr>
<td>59</td>
<td>NA</td>
<td>1.55 (0.70)</td>
<td>60</td>
<td>NA</td>
<td>1.63 (0.83)</td>
</tr>
<tr>
<td>61</td>
<td>NA</td>
<td>1.67 (1.13)</td>
<td>62</td>
<td>NA</td>
<td>1.72 (0.67)</td>
</tr>
<tr>
<td>63</td>
<td>NA</td>
<td>1.74 (0.45)</td>
<td>64</td>
<td>NA</td>
<td>1.75 (0.72)</td>
</tr>
<tr>
<td>65</td>
<td>NA</td>
<td>1.75 (0.64)</td>
<td>66</td>
<td>NA</td>
<td>1.87 (0.57)</td>
</tr>
<tr>
<td>67</td>
<td>NA</td>
<td>2.07 (0.69)</td>
<td>68</td>
<td>NA</td>
<td>2.13 (0.23)</td>
</tr>
<tr>
<td>69</td>
<td>NA</td>
<td>2.19 (0.12)</td>
<td>70</td>
<td>NA</td>
<td>2.25 (0.87)</td>
</tr>
<tr>
<td>71</td>
<td>NA</td>
<td>2.31 (0.50)</td>
<td>72</td>
<td>NA</td>
<td>2.33 (0.88)</td>
</tr>
<tr>
<td>73</td>
<td>NA</td>
<td>2.36 (0.57)</td>
<td>74</td>
<td>NA</td>
<td>2.37 (0.72)</td>
</tr>
<tr>
<td>75</td>
<td>NA</td>
<td>2.63 (0.75)</td>
<td>76</td>
<td>NA</td>
<td>2.66 (0.68)</td>
</tr>
<tr>
<td>77</td>
<td>NA</td>
<td>2.77 (1.10)</td>
<td>78</td>
<td>NA</td>
<td>2.80 (0.52)</td>
</tr>
<tr>
<td>79</td>
<td>NA</td>
<td>3.33 (0.71)</td>
<td>80</td>
<td>NA</td>
<td>3.54 (0.91)</td>
</tr>
<tr>
<td>81</td>
<td>NA</td>
<td>5.04 (1.10)</td>
<td>82</td>
<td>NA</td>
<td>5.43 (0.98)</td>
</tr>
</tbody>
</table>

In addition, (i) 70 studies allowed neither a OS HR nor a DFS HR to be extracted, and (ii) unknown unpublished studies could also exist.

* Predicted using the ‘three-step approach’; NA = not applicable.

R.D. Riley, Ph.D. Thesis
Chapter 9 331
**Step 2 - Sensitivity analysis using the predicted values**

On the assumption of MCAR or MAR and assuming there truly is an underlying linear relationship (with normally distributed errors) between estimates and also between standard errors, the predicted values from step 1 form the best estimate of the \( \log_e(\text{HR}) \) and its standard error for the missing outcome in those studies for which the other outcome was available. *For the purposes of sensitivity analysis*, I treated these predicted values as known and included these values in the *MYCN* dataset (Table 9.2) and updated the meta-analyses. I only used the predicted standard errors and *deliberately* did not take the actual uncertainty of the predicted estimates into account, as for the sensitivity assessment I wanted to assess how the original URMA results would change if the predicted values were indeed correct; incorporating all uncertainty would defeat this purpose, as the predicted estimates would have very little weight in the meta-analysis (this is considered further in Section 9.3.2).

I now updated the URMA for OS including the estimates from 81 studies (56 known plus 25 predicted), and similarly for DFS (42 known plus 39 predicted). The updated pooled estimates obtained for OS (pooled \( \log_e(\text{HR}) = 1.67 \), 95% CI 1.48 to 1.86, \( p < 0.0001 \)) and DFS (pooled \( \log_e(\text{HR}) = 1.49 \), 95% CI 1.31 to 1.66, \( p < 0.0001 \)) were very similar to those from the original URMA and also to those from Model B (Table 9.3). Similar pooled estimates were also seen when just considering the subset of predicted estimates, and just the subset of known values from (a) those studies reporting only one outcome and (b) those studies reporting both outcomes (Table 9.4). Interestingly there was a smaller estimate of between-study variation (\( \hat{\tau}_j^2 \)) in the subset of just the predicted estimates and this leads to the updated meta-analyses also having a smaller estimate of between-study variation than originally. This may not be true for other datasets as the between-study variation in the subgroup of predicted estimates clearly depends on the slope of the regression lines from which the predicted summary statistics and their predicted standard errors were calculated.
Table 9.3: Overall (OS) and disease-free (DFS) survival meta-analysis and sensitivity analysis results from URMA, Model B, Trim and Fill, and the ‘three-step approach’.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Description of the approach</th>
<th>Number of known estimates</th>
<th>Number of predicted estimates*</th>
<th>Number of 'Fill' studies</th>
<th>Meta-analysis results for log(HR)</th>
<th>Between-study variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>DFS</td>
<td>URMA of just the known</td>
<td>42</td>
<td>-</td>
<td>-</td>
<td>1.48 [1.23, 1.74]</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>estimates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DFS</td>
<td>Trim and Fill analysis</td>
<td>42</td>
<td>-</td>
<td>8</td>
<td>1.27 [1.01, 1.53]</td>
<td>0.57</td>
</tr>
<tr>
<td></td>
<td>using the known estimates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DFS</td>
<td>Model B using just the</td>
<td>42</td>
<td>-</td>
<td>-</td>
<td>1.47 [1.25, 1.70]</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>known estimates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DFS</td>
<td>URMA of the known plus</td>
<td>42</td>
<td>39</td>
<td>-</td>
<td>1.49 [1.31, 1.66]</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td>predicted estimates from</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>the ‘three-step approach’</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DFS</td>
<td>Trim and Fill analysis</td>
<td>42</td>
<td>39</td>
<td>19</td>
<td>1.22 [1.04, 1.40]</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td>of the known and predicted</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>estimates from the ‘three-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>step approach</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OS</td>
<td>URMA of just the known</td>
<td>56</td>
<td>-</td>
<td>-</td>
<td>1.63 [1.40, 1.87]</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>estimates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OS</td>
<td>Trim and Fill analysis</td>
<td>56</td>
<td>-</td>
<td>12</td>
<td>1.37 [1.12, 1.62]</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>using the known estimates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OS</td>
<td>Model B using just the</td>
<td>56</td>
<td>-</td>
<td>-</td>
<td>1.65 [1.42, 1.88]</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>known estimates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OS</td>
<td>URMA of the known plus</td>
<td>56</td>
<td>25</td>
<td>-</td>
<td>1.67 [1.48, 1.86]</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td>predicted estimates from</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>the ‘three-step approach’</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OS</td>
<td>Trim and Fill analysis</td>
<td>56</td>
<td>25</td>
<td>18</td>
<td>1.41 [1.21, 1.61]</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>of the known and predicted</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>estimates from the ‘three-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>step approach</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* predicted from step 1 of the ‘three-step approach’

HR = hazard ratio; N.B. The results from Model B, Trim and Fill, and those from the ‘three-step approach’ are sensitivity analyses to assess the robustness of the original URMA analysis.

Table 9.4: Meta-analysis results for 3 subsets of the known and predicted estimates

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Subset of the estimates</th>
<th>Number of known estimates</th>
<th>Number of predicted estimates*</th>
<th>URMA results for log(HR)</th>
<th>URMA results for $\overline{\tau^2}$, the between-study variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>DFS</td>
<td>Just the predicted</td>
<td>-</td>
<td>39</td>
<td>1.49 [1.24, 1.74]</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td>estimates</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DFS</td>
<td>Just estimates from</td>
<td>17</td>
<td>-</td>
<td>1.36 [0.94, 1.79]</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td>studies providing both</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>outcomes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DFS</td>
<td>Just estimates from</td>
<td>25</td>
<td>-</td>
<td>1.56 [1.23, 1.88]</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>studies only providing</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>DFS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OS</td>
<td>Just the predicted</td>
<td>-</td>
<td>25</td>
<td>1.74 [1.44, 2.04]</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td>estimates</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OS</td>
<td>Just estimates from</td>
<td>17</td>
<td>-</td>
<td>1.53 [1.11, 1.97]</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>studies providing both</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>outcomes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OS</td>
<td>Just estimates from</td>
<td>39</td>
<td>-</td>
<td>1.68 [1.39, 1.97]</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td>studies only providing</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>OS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* predicted using step 1 of the ‘three-step approach’;

OS = overall survival, DFS = disease-free survival, HR = hazard ratio
Figure 9.4: Funnel plots for the known, predicted and 'filled' estimates

(a) DFS
42 known estimates

(b) OS
56 known estimates

(c) DFS
42 known, 39 predicted* estimates

(d) OS
56 known, 25 predicted* estimates

(e) DFS
42 known, 39 predicted*, 19 'filled' estimates

(f) OS
56 known, 25 predicted*, 18 'filled' estimates

* predicted in step 1 of the 'three-step approach'
The 'filled' estimates were from the Trim and Fill analysis which included both the known and 'predicted' estimates.
Step 3 – Assessment of dissemination bias for the results from Step 2

The updated meta-analysis results from step 2 are similar to those from Model B, which also made the MCAR or MAR assumption in those studies only providing one outcome (Table 9.3). However, unlike for Model B, steps 1 and 2 above immediately provide a way of transforming the ‘borrowing of strength’ from a BRMA framework back on to the original funnel plots. The predicted estimates and their predicted standard errors can be plotted alongside those values already known on the funnel plot (Figure 9.4), and an updated assessment of the potential impact of dissemination bias (from those studies for which neither OS nor DFS was available) can now be made.

The visual inspection of these extended funnel plots indicated that asymmetry was still a concern (Figure 9.4 (c) and (d)). Indeed, applying Egger’s Test to the extended datasets suggested that asymmetry was still a problem for DFS ($p = 0.046$) and possibly also for OS, with Egger’s p-value for OS ($p = 0.16$) now considerably smaller than when the predicted values were not included ($p = 0.45$). Furthermore, applying the Trim and Fill method to the extended datasets estimated 19 DFS and 18 OS studies to still be ‘missing’ (Figure 9.4 (e) and (f)). This was a greater number of studies than estimated when the predicted values were not included, although the proportion ‘Filled’ to the overall number studies was very similar; for example, originally 8 DFS studies were ‘Filled’ in addition to the 42 already known (19%) whereas for the extended dataset 19 studies were ‘Filled’ in addition to the 81 known (23%). The asymmetry in the funnel plots could again not be explained by any study characteristics and so dissemination bias was deemed the most likely cause. It is likely that the remaining 70 known studies for which neither OS nor DFS was available, combined with some unknown and unpublished studies, are those contributing to this dissemination bias problem observed for the extended datasets.
To conclude my assessments of dissemination bias on these extended datasets, I also incorporated the ‘missing’ study estimates from the updated Trim and Fill analysis into the meta-analyses, alongside the known and predicted values. The URMA pooled estimates of the \( \log_e(\text{HR}) \) for both DFS (100 studies: pooled \( \log_e(\text{HR}) = 1.22, \ 95\% \ CI \ 1.03 \ to \ 1.40, \ p<0.0001 \)) and OS (99 studies: pooled \( \log_e(\text{HR}) = 1.41, \ 95\% \ CI \ 1.21 \ to \ 1.61, \ p<0.0001 \)) were similar to the Trim and Fill sensitivity results from when the predicted values were not included (Table 9.3), but still indicative that \textit{MYCN} may be of possible prognostic value.

**9.3.2 Critical discussion of the ‘three-step approach’**

In conclusion, after assuming MCAR or MAR is true in those studies only providing one outcome and after assessing the threat of dissemination bias from other studies providing no summary statistics, the ‘three-step approach’ suggests that the true OS and DFS pooled estimates are likely to be somewhat lower than the original URMA and Model B results. However, they are still strongly indicative of \textit{MYCN} being of \textit{potential} prognostic value and that \textit{MYCN} should be a high priority for further prognostic research in neuroblastoma. This conclusion is the same as when methods for dissemination bias were originally applied to just the known HRs (see Section 9.2), however the ‘three-step approach’ conclusion was derived differently, as it specifically takes into account those known studies for which only one outcome was available. For example, the original Trim and Fill analysis suggested that there were 8 DFS and 12 OS studies ‘missing’ based on funnel plot asymmetry; however, I knew specifically that there were at least 95 studies missing for OS and 109 studies missing for DFS because 70 studies provided neither outcome, 39 just provided OS, 25 just provided DFS and other unidentified, unpublished studies may exist. Hence, I have tried to use the ‘three-step approach’ to specifically consider the problem of missing information from \textit{known} studies before applying the Trim and Fill method. Indeed, by predicting missing studies in the ‘three-step approach’ first I might have accounted for...
some of those MYCN studies that Trim and Fill originally suggested were ‘missing’ from the funnel plots; however this turned out to not be the case, as funnel plot asymmetry remained even after the ‘three-step approach’.

The fact that all the sensitivity analyses still concurred that MYCN was potentially valuable is a very important finding, one that strengthens the original URMA and Model B conclusions because they remain robust even when considering the possible impact of the missing data. This concurrence is perhaps not surprising given the magnitude of the original known HRs for MYCN; in other less clear-cut situations, the sensitivity approaches may be even more valuable and may not produce consistent findings [203;204]. The ‘three-step approach’ also illustrates the need and potential benefit of sensitivity assessments of bivariate meta-analysis results, and it also provides a helpful insight into the difficulties of translating current dissemination bias methods within a univariate meta-analysis framework onto a bivariate meta-analysis framework. The ‘three-step method’ is not without its concerns, however. Firstly, the prediction model used in step 1 is very sensitive to extreme observations, due to the small number of studies providing both OS and DFS; for example, when study 17 is removed, equation (ii) changes considerably (\( \alpha \) from 0.038 to 0.34, \( \beta \) from 0.87 to 0.65, adj-\( R^2 \) from 0.79 to 0.54). However, when predicting estimates from prediction models formed without study 17, the updated URMA results changed very little to before (e.g. DFS pooled loge(HR) changes from 1.49 to 1.45).

Secondly, in general meta-analyses of prognostic or other studies will usually involve a considerably smaller number of overall studies than I had available for MYCN, thus limiting the opportunity to firstly observe and then to model relationships between outcomes as I have. This problem also limits Model A and Model B because the correlation parameters cannot be estimated very well when the number of studies is small (see Sections 5.3 and 8.5.6). Another issue for the ‘three-step approach’ is that if the known estimates used for predicting others are mostly extreme observations themselves,
there is then the additional concern of making predictions beyond the range of the observed

data, and one would have to assume here that the observed relationship between outcomes

continued above and below the observed range. Of the predicted estimates I calculated for

MYCN, only one (study 43) was outside the range of the HRs from the 17 studies that

provided both outcomes and which were used to estimate the regression model.

There are also two issues to do with statistical uncertainty that I will now discuss. Firstly,

for each predicted summary statistic, there is clearly a difference between its predicted

standard error from step 1 and the actual standard error of the predicted value itself. For

instance, for study 18 the predicted standard error for OS was 0.38 (Table 9.2), but this is a

predicted value from step 1 using the regression line in Figure 9.3(b), and the actual

standard error of the predicted summary statistic would be far greater than 0.38 because it

would take into account the uncertainty in the estimates used to make the prediction, i.e. $\alpha$

and $\beta$ in equation (i) and the $\log_e(HR)$ for DFS. Of course, for my updated meta-analyses

in step 2 I used the predicted standard error of the predicted summary statistics and not the

actual standard error of the predicted summary statistics. This was because if the actual

standard errors had been used the predicted summary statistics would have very little

weight in the updated meta-analysis, as the actual standard error of the $\log_e(HR)$ would be

relatively large, and the pooled results would therefore be very similar to before, defeating

the need to assess the robustness of the original URMA results given the problem of

missing data. It would also mean that the Trim and Fill method could not be subsequently

used, as this method would then impute unrealistic ‘missing’ studies based on the increased

asymmetry in the funnel plots caused by the large actual standard errors being used. Hence,

by using the predicted standard errors I avoided this problem and was able to see the most

likely impact of the missing summary statistics, assuming of course the prediction models

were valid.
The second discussion point regarding uncertainty is that, for both the missing summary statistics and the missing standard errors, I have only considered their mean predicted values and not other possible values. For example, I could have taken predicted values other than the mean of the predicted distribution and used these in the updated URMA and funnel plot assessments. Hence, perhaps it would be more suitable to take an approach similar to multiple imputation, and perform multiple updated URMAs and multiple updated funnel plot assessments across a range of multiple predicted values used. One could then assess the average results from across all the multiple updated analyses to make final assessments regarding the robustness of the original URMA results. In this situation it would be very interesting to see whether the average conclusions from such multiple assessments are any different to the conclusion from a single updated assessment which used the mean predicted values as I did in step 3. This is likely to depend on the assumption that the predicted distributions are normally distributed and are therefore symmetric. Multiple assessments would of course be far more computationally intense than a single assessment, and perhaps a Bayesian approach would most naturally facilitate samples being drawn from the predicted distributions and then used in an updated meta-analysis. However, even the Bayesian approach may struggle to easily include and report an additional Trim and Fill analysis for each of the multiple assessments.

The prediction models used in step 1 may not be considered particularly sophisticated but they provide a possible method of obtaining the missing information required, something not so easily done in other models. For example, Model B or Model B-Bayes was not used for prediction because there is a stumbling block for predicting missing summary statistics from such random-effects models. Consider the Bayesian random-effects approach of Model B-Bayes, which would seem a particularly sensible way of predicting missing summary statistics (i.e. missing \( \tilde{\gamma} \) values) because one could obtain a mean predicted
value of the missing $\hat{Y}_j$ from its predictive distribution. However, unfortunately this predictive distribution will be *shrunk* away from the known summary statistics toward the posterior pooled estimate, $\hat{\beta}_j$. Put another way, the mean predictive estimate for the missing summary statistics will not be $\hat{Y}_j$ but rather $\hat{\theta}_j$, the mean posterior shrunken or empirical Bayes study estimate (see Section 4.6). Unfortunately the $\hat{\theta}_j$s are *not* what are required for the funnel plot and dissemination bias assessments of Trim and Fill or Egger’s Test. Of course the problem of predicted values being shrunk is not just a problem of random-effects models; for example, where a bivariate fixed-effects model (e.g. Model A where $r_j^2 = 0$) is used, its mean predicted value of the missing $\hat{Y}_j$s will again be shrunk to $\bar{\beta}_j$, the overall underlying pooled estimate. An additional problem for predicting values from both random- and fixed-effects models is that they cannot predict the missing standard error of the missing $\hat{Y}_j$s. Indeed, further exploration regarding how one can appropriately predict missing $\hat{Y}_j$s and their standard error to aid funnel plot assessments would make interesting further research. However, as it avoids the problems discussed, the prediction model in step 1 of the ‘three-step approach’ would seem a simple yet very suitable way of estimating the most likely values for the missing summary statistics and their missing standard error.

The ‘three-step approach’ therefore forms a potential starting point for further research of how to assess dissemination bias following a bivariate meta-analysis approach, and it provides a novel way of specifically considering *known* studies for which summary statistics were MCAR or MAR alongside additional studies for which the missing information may be NMAR due to dissemination bias.
9.4 The use of selection models to assess dissemination bias

9.4.1 Selection models

I want to now discuss possible solutions to question (ii) posed in Section 9.1, which asked how one may use a bivariate meta-analysis framework to assess dissemination bias when none of the missing summary statistics can be assumed MCAR or MAR. In this situation neither Model B nor the ‘three-step approach’ would be valid because they both assume MCAR or MAR for the unavailable summary statistic in those studies providing only one of the two outcomes. One option when data is NMAR in known studies is to predict a range of possible values for the missing summary statistics in these studies using a selection model, which assumes a specific selection procedure has been used when reporting results and that this has ultimately led to a biased set of available summary statistics (e.g. just those which were statistically significant) [197]. In the univariate meta-analysis setting, selection models have been considered by relatively few authors and it is an under-researched part of the meta-analysis field [65;197;204-206], which is despite their being increasing evidence that within-study selective reporting does happen [64]. The Trim and Fill method is not a selection model because it simply assumes that a biased availability of summary statistics across studies will be evident by an asymmetry in the funnel plot. Formal selection models in meta-analysis consider specific selective within-study reporting for those studies that were known but for which summary statistics were not available. Indeed Terrin et al. have recently indicated that selection models are likely to perform better than the Trim and Fill method in practice [201]. I will now discuss two specific examples of how selection models have been used in the meta-analysis literature, and I will indicate why the approaches used help to assess the potential impact of summary statistics that are NMAR on the results of a URMA.
9.4.2 Two examples of the use of selection models in meta-analysis

Hahn et al. consider a meta-analysis of five studies for which an odds-ratio was desired from each, but was ultimately only available from three [204]. Specifically considering the two studies for which the odds-ratio was missing, the authors consider a worst-case scenario and simulate a range of possible values for the missing statistic and its standard error assuming it was unavailable because it was not statistically significant at the 5% level. They do this by utilising other data available in each published article (e.g. total number of patients in each group) to fit constraints about the missing information (e.g. number of deaths in each group), which if known could have been used to estimate the log-odds ratio and it standard error. Using values across the constrained range identified, a number of simulated values for the missing log-odds ratio and standard error were then generated for each study. For each permutation of the simulated values in the two studies, a separate updated meta-analysis was performed which included them alongside the other three studies for which the log-odds ratio was known. The meta-analysis results across all permutations of the missing values were then assessed and the authors conclude that, if the selection procedure was truly based on p-values being greater than 0.05, the pooled estimate from the original analysis, which included only three studies, is most likely an overestimate of the true effect, and that misleading evidence-based conclusions may therefore have been made originally.

A slightly different approach is taken by Hutton and Williamson [65], who consider a selection model for the situation when m outcomes are considered by a publication but only the one with the smallest significance level is reported. The authors assume that the summary statistics from each outcome are actually distributed about the same underlying mean value, and propose how to adjust the one available summary statistic and its standard error assuming (m-1) summary statistics with larger p-values are not available. In two real meta-analysis examples presented, the authors assess the robustness of the original meta-
analysis results when including each set of ‘bias corrected’ summary statistics. The authors found that the original conclusions could easily be reversed if within-trial selective reporting has taken place, and in one of the examples the original conclusion was particularly dependent on the missing result from one large trial which was not originally included in the meta-analysis.

9.4.3 The use of selection models when two correlated summary statistics are needed for a meta-analysis

The selection models of Hahn et al. and Hutton and Williamson are both essentially using information that is known in a study to help generate possible values of the missing summary statistic based on the assumption that a biased reporting procedure was used. These are the main characteristics of selection models in meta-analysis [197], and it is therefore clearly possible to use them when meta-analysis of two correlated summary statistics is desired. Indeed, it is conceivable that this situation facilitates the use of selection models because there are those studies for which only one summary statistic is missing, and therefore for these studies one could use the summary statistic available to help produce possible values of the missing summary statistic, again assuming a specific biased reporting procedure had taken place. For example, taking a similar approach to Hutton and Williamson, one could assume that the known summary statistic has a smaller p-value than the missing summary statistic and use this to help generate possible values of the missing statistic for a range of larger p-values. The application of a URMA or even a BRMA model (such as Model A or Model B) could then be applied using both the known and the generated summary statistics, with a separate meta-analysis model fitted for each permutation of the generated values across studies. The overall meta-analysis results across all simulated datasets could then be used to assess the robustness of the original (URMA or BRMA) results presented. Of course, when there are a large number of missing summary statistics they will also be a large number of generated values, and then an even larger
number of permutations of the generated values across studies that need to be separately fitted in a meta-analysis model. The number of simulations required here indicates that a Bayesian framework may be most suitable for this situation, as a single model could be used to firstly generate all possible missing values and to then incorporate all the different permutations of these within the meta-analysis model specified. Indeed, Hahn et al. use a Bayesian framework for their selection model analyses for similar reasons [204]. Of course, when using such a Bayesian approach the influence of any 'vague' prior distributions also needs to be assessed (see Chapter 7). Furthermore, it should be noted again here that if Model A or Model B was fitted to each of the generated datasets then one would still need to be cautious about problems of bias when the correlation parameters were poorly estimated (see Figure 8.10 in Section 8.7.2).

9.4.4 Two common obstacles for using selection models in practice

There are perhaps two main problems that limit the use of selection models for meta-analysis. Firstly, not only do values for the missing summary statistics need to be generated, but also values for their missing standard errors are needed. Hahn et al. do not have this problem because they generate possible values of the missing information in $2 \times 2$ tables [204], which can then be used to calculate values for both the missing log-odds ratio and its associated standard error. In other situations it may not be so straightforward. For example, Hutton and Williamson have to use the observed relationship across studies between known standard errors and other study information to then predict in another study the most likely value of the missing standard error given the related information [65]. This predicted value of the standard error is then used alongside a predicted value of the missing summary statistic within an updated meta-analysis. Of course, this approach is similar to how I predicted the standard error in step 1 of the 'three-step approach', and it again emphasises that selection models may be more readily applicable where two correlated summary statistics are required from each study, because more related information may
commonly be available within a study to help generate predicted values. Of course, it
should be noted that if Model A was desired to be fitted to the generated summary statistics
then not only would a predicted standard error be necessary but also a predicted within-
study correlation. This suggests that Model B may be the most useful BRMA approach
within a selection model framework, as the within-study correlations are not required for
this model (see Section 8.2.1).

The second major problem that those using selection models will face is how to assess any
additional dissemination bias from those studies for which no information is available, or
from those unknown or unpublished studies? This was the same question asked of the
results from Model B after MCAR or MAR was assumed in those studies providing one of
the two summary statistics. From the selection model NMAR is now assumed in these
studies, but what about those studies for which the selection model has not generated
possible values for the missing summary statistics? The investigation of how to answer this
question would make interesting further research, and it would be one of the first
extensions of this thesis I would consider, with particular application to the MYCN dataset
again important. Neither Hutton and Williamson nor Hahn et al. consider any additional
dissemination bias assessments after fitting their selection models [65;204]. However, I am
aware that Jackson et al. have recently submitted a paper that does propose a selection
model followed by a further dissemination bias assessment using updated funnel plots for a
meta-analysis where two (uncorrelated) summary statistics are desired from each study but
are often not reported [207].

9.5 Summary and final thoughts on sensitivity analyses in meta-analysis

Systematic reviews of prognostic marker studies should seek to incorporate all the
published and unpublished evidence [34]; however, the example of MYCN in
neuroblastoma has illustrated the difficulty of achieving this given missing summary
statistics caused by dissemination bias and poor statistical reporting in primary studies. In this type of situation, when presented with an incomplete evidence-base, those performing meta-analysis should consider performing sensitivity analyses that consider the possible impact the missing summary statistics and dissemination bias might have on the meta-analysis results and conclusions. In this chapter I have considered how to perform such sensitivity analyses to help assess the robustness of both URMA and also BRMA results regarding prognostic marker MYCN (Appendix D1) [208]. In particular, where meta-analysis of two correlated summary statistics is affected by missing data, I have introduced a 'three-step approach' which assumes part of the missing data is either MCAR or MAR and that the remainder is NMAR, and then I have discussed the potential use of selection models when one assumes all the missing data is NMAR. The need for methods to help assess missing data in relation to a bivariate meta-analysis framework is a very real one if BRMA models, such as Model A or Model B, are to be more commonly used in practice and especially if they are to be used to inform the evidence-based use of prognostic markers. However, conversely my work has shown that the bivariate framework itself can actually facilitate suitable assessments of dissemination bias. For example, meta-analyses involving two correlated summary statistics may be particularly receptive to the application of selection models because more information is likely to be available in these studies to help generate possible values of any missing summary statistics (see Section 9.4.3).

This chapter has also emphasised that sensitivity analyses for the problem of missing data are not always straightforward and they must only be considered as a guide to the problems as they are themselves vulnerable to the validity of the underlying assumptions and may have low statistical power of detecting problems [209]. For example, funnel plot asymmetry may not be caused by dissemination bias but rather by the true heterogeneity present, poor design of small studies or even chance [41]. Indeed, this has caused the Trim
and Fill method to be criticised, as it can detect missing studies even in the absence of bias [201]. This emphasises that selection models may be the most suitable way to make sensitivity assessments in meta-analysis, where values of missing summary statistics are specifically generated for those known studies in which the required summary statistics are unavailable [201].

Considering the MYCN analysis and prognostic markers more specifically, there are some other bias related problems that are still a concern. For example, what if the choice of which prognostic markers to report was also affected by dissemination bias alongside the choice about which outcomes to report? Furthermore, what if there was evidence of biased choice of cut-off levels across studies, and how could sensitivity analyses account for this problem, if at all? Such questions are difficult to answer, but they clearly emphasise that reporting standards in primary studies need to improve (see Section 2.9). Furthermore, IPD should be sought from primary prognostic marker studies in order to form the most reliable and clinically relevant evidence-based results possible from a synthesis of these studies (see Section 2.9). However, there are of course additional concerns for IPD reviews (see Section 2.9.5), especially cost and time [143;144], and even within the IPD itself there is likely to be missing outcome and missing covariate information for some individuals [129], causing some dissemination bias issues to remain a concern. Sensitivity analyses will therefore play a continually important role in systematic reviews and meta-analysis of prognostic marker studies in the future, and inevitably also in other research fields, especially those involving observational studies [102]. Further research into sensitivity methods is therefore required to help produce the least biased evidence-based conclusions from meta-analysis. In particular, for those conducting meta-analysis of multiple outcomes and correlated summary statistics, sensitivity methods that utilise the known summary statistics to help predict the missing summary statistics may be an especially valuable and informative approach to consider (Appendix D1) [208].

R.D. Riley, Ph.D. Thesis Chapter 9
Chapter 10

DISCUSSION AND SUGGESTIONS FOR FURTHER RESEARCH

Chapter overview

In this thesis I have formally demonstrated the methodological difficulties of performing an evidence synthesis of prognostic markers, and have then developed appropriate guidelines and statistical methods to help facilitate clinically relevant evidence-based results in the future. In this concluding chapter I will summarise and discuss the most important findings and issues that have arisen during the thesis, and will particularly emphasise why the research undertaken benefits future evidence syntheses of prognostic markers. I will also consider the numerous methodological issues that remain unaddressed by the research in this thesis, and will identify the main further research priorities that have specifically developed out of the work I have presented.

10.1 Benefits of the neuroblastoma review and the reporting guidelines developed

'The quality of reviews of observational studies must be improved to address questions about aetiology, diagnostic accuracy, risk prediction and prognosis'

Chalmers, 2001 [210]

This quote from Chalmers emphasises the rationale behind why I have studied the problems and issues associated with an evidence synthesis of prognostic marker studies [210]. Prior to this thesis the literature had clearly indicated that systematic reviews and meta-analysis would be difficult for this field [76]. In order to assess and address this problem, I performed a large-scale systematic review of prognostic markers in neuroblastoma (see Chapter 2), and have used this to identify and demonstrate the specific
factors which limit clinically relevant evidence-based results for prognostic markers, most of which are related to the primary studies themselves. This research is important because it provides clear evidence that primary prognostic studies need to be substantially improved if evidence-based prognostic marker results are to be obtained in the future.

There are perhaps three main areas where improvements in primary prognostic marker studies are needed: study design, clinical relevance, and study reporting. The neuroblastoma review particularly highlighted the reporting problems, such as poorly reported summary statistics (e.g. no confidence intervals presented, or only inexact p-values provided) and large heterogeneity in terms of statistical (e.g. adjustment factors), clinical (e.g. age, stage of disease) and methodological (e.g. cut-off levels) characteristics across studies (see Section 2.3.1). However, the reporting problems are clearly related to the problems associated with study design and clinical relevance (see Section 2.6), and the key aspects of how each of these areas needs to improve are summarised in Figure 10.1 [76;86;116]. Importantly, forward steps in each of these three areas would not only improve the clinical relevance and validity of the primary studies themselves, but they would also greatly facilitate evidence-based reviews in this field. Firstly, improved standards of study design would allow the synthesis of higher methodological quality and more appropriately targeted primary studies, giving more credibility to the meta-analysis results produced. Secondly, if each study were of direct relevance to clinical practice then this would also facilitate the production of clinically relevant meta-analysis results and thus appropriate clinical decisions; for example, if new markers in neuroblastoma were consistently assessed in relation to MYCN, then this would potentially enable meta-analysis results about the benefit of using those markers either instead of or in combination with MYCN. Thirdly, improved reporting would allow a larger number of studies to be included in a meta-analysis, as more summary statistics and outcomes would be available across studies to be incorporated in the synthesis.
Figure 10.1: Key areas where improved research and better statistical practice are needed within primary prognostic marker studies to facilitate clinically useful evidence-based results

- **Study design** – see Altman and Lyman [76]. Improvements need to arise through greater consideration of, amongst other things:
  - study type (e.g. Phase I, II or III)
  - study purpose (e.g. what is the primary objective?)
  - sample size (e.g. what is the desired power to detect meaningful differences in outcome)
  - inclusion/exclusion criteria
  - prior hypotheses and the prior specification of which markers and outcomes are to be considered (e.g. in a study protocol)

- **Clinical relevance** – see Windeler [116]. Primary studies need to produce more clinically useful results by, amongst other things:
  - being high-quality, large-scale Phase III studies and relating prognostic marker assessments to specific treatment strategies
  - collaborating across research groups to achieve larger sample sizes and consistency in method of measuring the marker, cut-off levels, adjustment factors and outcomes assessed
  - building on the results of previous studies, and assessing new markers in relation to those markers and treatments currently used in practice

- **Study reporting** – see Riley et al. [86]. Improved reporting of prognostic marker studies is greatly needed with regard to, amongst other things:
  - reporting the results of all markers and outcomes considered, not just those statistically significant (to reduce the threat of publication bias and within-study selective reporting)
  - reporting an effect estimate (e.g. hazard ratio) with confidence interval, rather than solely a p-value
  - reporting marker results adjusted for other prognostic factors and treatments of recognised and accepted clinical importance
  - making individual patient data available for those performing evidence synthesis

The development and publication of guidelines can help achieve the necessary standards in each of these areas, assuming of course that authors adhere to them (see Section 2.9.6).

Previous guidelines for primary prognostic marker studies have focused on their design, purpose and clinically relevance [76;116], and to complement these, in this thesis I have developed guidelines about how to report prognostic marker results (see Figure 2.11 in
Section 2.9). These reporting guidelines are particularly important because they should improve the interpretability of the findings from individual studies, whilst enabling more summary statistics and other pertinent information to be available for meta-analysis. In particular, I have encouraged summary statistics such as the HR and the actuarial survival at \( n \) years to be presented within each published article, together with a measure of uncertainty (e.g. standard error, confidence interval) for each statistic. These are two basic but very important clinical measures that are currently not clearly reported, as emphasised by the neuroblastoma review (see Sections 2.2 and 2.3). They are both greatly needed to allow each primary study to be of practical use, either on its own or as part of an evidence synthesis, and I deemed them to be the minimum reporting standard required.

For those meta-analysts of prognostic marker studies, perhaps the most important recommendation in my reporting guidelines is that IPD should be made more readily available from primary studies, and even stored in a central data repository [136]. For the systematic review I performed in neuroblastoma (see Chapter 2), IPD was sometimes available within the published literature itself, because many primary studies had a small sample size, and this enabled me to calculate a number of summary statistics that would otherwise have been difficult to obtain (see Section 2.9.1). However, for evidence synthesis of prognostic markers in non-paediatric oncology and for non-rare diseases settings, it is perhaps less likely for IPD to be published because journals will not usually have the necessary publication space to accommodate the larger IPD available.

Stewart and Parmar recommend that, whenever possible, meta-analysis using IPD is preferred because it produces the least biased answers and therefore the most suitable way of addressing questions that have not been or could not be resolved by individual studies [128]. The main advantages for an IPD evidence synthesis of prognostic marker studies are
shown in Figure 10.2, and these would help reduce those problems which exist when performing a meta-analysis using summary statistics (see Figure 3.2 of Section 3.1). In particular, perhaps the greatest advantage of IPD is that it would help both reduce and explain the heterogeneity across studies.

**Figure 10.2:** Summary of the main benefits of having individual patient data (IPD) for a meta-analysis of prognostic marker studies.

Availability of IPD from the primary studies has the potential to allow one to:
- Obtain estimates for those missing or poorly reported outcomes and summary statistics across studies; it may thus reduce the problem of selective within-study reporting
- Obtain more direct estimates (e.g. of the hazard ratio) where previously only indirect estimates were available
- Produce more adjusted estimates where previously only unadjusted estimates were available
- Standardise the definition of outcomes across studies
- Use a consistent and appropriate method for statistical analysis across studies
- Use a consistent set of adjustment factors across studies
- Use a consistent cut-off level across studies, or produce continuous marker results where originally a cut-off level was used (or vice versa).
- Assess specific subgroups of patients across studies (e.g. for age < 1, stage 4 disease), and assess whether patient level characteristics (such as age and treatment) are effect modifiers across studies [54]
- Identify those studies which contain the same or overlapping sets of patients
- Assess model assumptions in each study, such as proportional hazards
- Assess markers over time and in relation to time-dependent covariates such as treatment received and stage of disease

The systematic review in neuroblastoma identified three main causes of heterogeneity across prognostic marker studies: (a) statistical, (b) clinical, or (c) structural or methodological. **Statistical heterogeneity** arises out of the fact that individual primary studies report quantitatively different results purely because of random variation, and this common problem is why random effects meta-analysis methods are frequently used [43].

**Clinical heterogeneity** refers to the differences between the populations and their characteristics across primary studies, e.g. age, treatment received, and stage of disease.

Finally, **structural or methodological heterogeneity** refers to differences in the actual
study design, methodology or reporting of results/analysis used by the different primary studies (e.g. prospective/retrospective studies, adjusted/unadjusted results, and different cut-off levels). When any combination of the three types of heterogeneity outlined is present it can severely limit the statistical and clinical appropriateness of the meta-analysis conclusions. However, the availability of IPD from the primary studies can help alleviate the heterogeneity in each of these areas. For example, IPD could potentially reduce methodological heterogeneity by allowing one to be consistent in the cut-off levels and adjustment factors used across studies (see Sections 2.9.2 and 2.9.3). IPD would also allow the analysis of sub-populations across studies (e.g. age < 1) to limit the clinical heterogeneity (see Section 2.9.4). Perhaps most importantly, IPD would also allow one to explain the statistical heterogeneity by facilitating a meta-regression that assesses how clinical and patient characteristics (e.g. age, stage of disease, treatment received) are associated with the benefit of a prognostic marker, something that has low statistical power if only aggregated patient level data are available [54].

The extensive heterogeneity in the neuroblastoma review, combined with the poor reporting, mean that the neuroblastoma meta-analysis results presented in this thesis should be treated with caution and can only be used to identify those markers which are potentially important, with further more clinically relevant research required (see Section 2.4). This may seem like a very disappointing conclusion from such a large-scale systematic review of 260 published articles, and the lack of clinically relevant results is disconcerting. However, the review has helped frame the future research priorities for prognostic markers in neuroblastoma and initiated reporting guidelines to help generally overcome the problems, so there are still a number of benefits derived from the research. Furthermore, for this particular review, it is debatable whether even having IPD would have enabled clinically relevant conclusions, simply because the quality and design of the neuroblastoma primary studies were also likely to be inadequate (e.g. only Phase I or Phase...
II studies performed, see Section 2.6.2). IPD has many advantages (Figure 10.2), but it cannot address the problems of poor study design or that a study does consider relevant clinical questions.

There are, of course, other limitations to IPD evidence-based reviews, and my reporting guidelines will not solve all the problems (see Section 2.9.5); for example, there will inevitably be missing patient data in the IPD itself and similarly not every study will have recorded information for all the outcomes, markers and clinical characteristics of interest. IPD may also not be available for every study, and so the meta-analyst may still have to collect summary statistics for some studies even when taking the IPD approach (see discussion on joint synthesis of IPD and summary statistics in Section 10.4.4). This last point emphasises again the importance of improving the reporting of summary statistics, something my reporting guidelines specifically target (see Figure 2.11 in Section 2.9). Indeed, as IPD reviews may be very costly and time-consuming, future evidence-based reviews of prognostic markers may still have to prioritise the collection and synthesis of published summary statistics. Hence, alongside improvements in the quality of primary studies, there is a need to develop novel meta-analysis methods that enable the least biased results to be obtained from a meta-analysis of summary statistics, and which limit some or all of the reporting problems that the neuroblastoma review faced. It is in this context that one should place the meta-analysis research that followed the neuroblastoma review in this thesis.

10.2 Benefits of the research into bivariate random-effects meta-analysis

There are a number of problems facing meta-analysis of summary statistics from prognostic marker studies (see Figure 3.2 in Section 3.1), and in this thesis I have considered methods to help address the specific issues of missing outcomes, missing summary statistics and the threat of dissemination bias. Although this leaves many other
problems unresolved (see ‘further research’ in Section 10.4.1), the methods developed in this thesis complement the reporting guidelines and provide the initial steps towards less biased and more clinically relevant evidence-based prognostic marker results.

The large part of the thesis (Chapters 3 to 9) studied the benefits and limitations of bivariate random-effects meta-analysis (BRMA), and the motivation for this was to utilise the strong correlation between OS and DFS HR estimates that was commonly observed in the neuroblastoma review. In particular, I wanted to facilitate the ‘borrowing strength’ across outcomes when one of either OS or DFS was missing in a study, and thereby potentially enable less biased meta-analysis results than would have otherwise been possible from a univariate random-effects meta-analysis (URMA). Usefully, as BRMA generally allows the joint modelling of two correlated summary statistics, the findings for BRMA in this thesis are also applicable away from the prognostic marker field and to other summary statistics beside the HR (see Sections 3.6.3 and 8.5).

In Section 3.6.1 I introduced Model A, a general frequentist framework for BRMA that has been the most commonly used bivariate meta-analysis approach in the current literature [148]. Model A is a fully hierarchical model, incorporating both within- and between-study correlation and variance parameters. Alongside the information required to apply a URMA for each outcome independently (i.e. a summary statistic and its associated standard error for each outcome from each primary study), one also requires the within-study correlation between the two summary statistics to be known from each primary study in order to apply Model A. This is highly unlikely in most situations, and this problem is the main methodological hurdle for the application of Model A to OS and DFS prognostic marker results. However, alongside this issue, a review of the literature highlighted that there were many other unresolved questions about using Model A in practice (see Section 3.6.3); in particular, there was little evidence about the benefits and limitations of the approach even
for when the within-study correlations were known. Hence, in Chapters 4 to 7 of this thesis I have specifically tried to answer a number of the unresolved questions about Model A posed by the previous literature (see list of unanswered questions in Section 3.7) for when the within-study correlations were known, and to this end work toward being in a better position to help address the problem of unknown within-study correlations for application of BRMA in the evidence synthesis of prognostic markers.

The main benefits and limitations of Model A that were identified by this thesis are shown in Figure 8.10 of Section 8.7.2. The main benefit is in the estimation of \((\beta_1 - \beta_2)\) because the model allows the incorporation of \(\text{corr}(\hat{\beta}_1, \hat{\beta}_2)\), which is not available when fitting two independent URMAs (see Section 4.8.6). However, \((\hat{\beta}_1 - \hat{\beta}_2)\) may not be of interest in many situations, and especially in the synthesis of prognostic marker studies, where the individual DFS and OS pooled HRs (i.e. \(\hat{\beta}_1\) and \(\hat{\beta}_2\)) would be of primary interest.

However, my research in Chapter 6 showed for the first time that Model A can produce more appropriate results for \(\hat{\beta}_1\) and \(\hat{\beta}_2\) than a URMA when there is missing data. In particular, there can be considerable gain in precision, reduction in mean-square error (MSE) and more suitable coverage for \(\hat{\beta}_1\) and \(\hat{\beta}_2\) (see Sections 6.4 and 6.5). However, there are two limitations to this result. Firstly, Model A assumes that the missing summary statistic is either missing completely at random (MCAR) or missing at random (MAR) in those studies only providing one outcome (see Section 6.2). Secondly, one needs to be extremely cautious about using Model A results where the between-study correlation \((\rho_B)\) is estimated to be 1 or -1, because in this situation the simulation studies clearly show that there is an upward bias in the between-study variance estimates (see Sections 5.5 and 5.6). Indeed, when the within-study correlations are known, these two problems would appear to be the main issues preventing Model A from being useful in practice.
The demonstration that the between-study correlation is often estimated as 1 or −1 in Model A is one of the most important findings from this thesis. Model A has been applied in a wide-variety of evidence synthesis contexts (see Section 3.6.3), but the problem of bias when the between-study correlation is estimated as 1 or −1 has not been demonstrated in detail. Thus, by highlighting this problem and providing potential reasoning for it (see Section 5.6), the research in this thesis should help facilitate the proper use and application of Model A results in real meta-analysis situations in practice. Importantly, there are no problems of bias in Model A when the between-study correlation is not estimated as 1 or −1 (see Section 5.5). Furthermore, a Bayesian approach to Model A (Model A-Bayes) that incorporates external information can help to prevent the between-study correlation being estimated as 1 or −1 (see Section 7.5). However, strong external information may be difficult to illicit in practice (See Section 7.5.4), and one must be cautious about using Model A-Bayes when only ‘vague’ prior information exists as otherwise misleading conclusions may be made (see Section 7.3).

In terms of applying BRMA to prognostic marker studies, Chapter 8 is by far the most constructive chapter of the thesis because it specifically considers the various options for BRMA when the within-study correlations are unknown. A review of the literature showed that none of the previous methods for dealing with unknown within-study correlations are ideal. Even the commonly used Model A-zero, which assumes the within-study correlation is zero in every study, is problematic because it increases the problem of the between-study correlation being estimated as 1 or −1 (see Section 8.1.1). In response to this I have introduced Model B, an alternative approach to BRMA, which instead of having within-and between-study correlation parameters only has a single correlation parameter (\( \rho \)) that models directly the overall correlation between the two outcome summary statistics (i.e. the \( \bar{Y}_{ni} \)'s and the \( \bar{Y}_{i2} \)’s; see Section 8.2.1). This means Model B does not require the within-
study correlations to be available from each study, and indeed to fit the model one only needs the same information required to fit two independent URMAs (one URMA for each outcome). This makes Model B capable of the joint synthesis of OS and DFS summary statistics from prognostic marker studies (see Sections 8.5.1 and 8.5.2) and so, alongside the reporting guidelines, Model B is potentially the most important part of this thesis for facilitating evidence-based prognostic marker results.

Sections 8.2 and 8.3 critically assessed Model B primarily through simulation studies, and this research identified that Model B was a suitable approach to BRMA (see summary of Model B benefits and limitations in Figure 8.10 of Section 8.7.2). In particular, where \( \rho \) was estimated between \(-0.95\) to \(0.95\), Model B has the following main benefits:

- **Model B produces very similar results to Model A when the latter is not subject to bias (e.g. see Section 8.3.5).** Hence, the benefits of Model B over a URMA are exactly the same as those benefits for Model A. Firstly when data is either MCAR or MAR and estimates of the pooled \( \tilde{\beta} \), values are of interest, Model B is preferred because it will on average obtain \( \tilde{\beta} \) s with greater precision, smaller MSE and more suitable coverage than the \( \tilde{\beta} \) s from two independent URMAs (see Section 8.3.5). Furthermore, in terms of estimating \( (\beta_1 - \beta_2) \), Model B is clearly always preferred (i.e. in complete-case, MCAR and MAR situations) to a URMA because, unlike in a URMA, it allows the \( \text{corr}(\tilde{\beta}_1, \tilde{\beta}_2) \) to be estimated and therefore a more appropriate estimate of \( \text{var}(\tilde{\beta}_1 - \tilde{\beta}_2) \) is obtained, which leads to a more suitable coverage (i.e. closer to 95%).
- **Even when the within-study correlations are known, Model B may be useful over and above Model A** because there will be some situations where Model A estimates \( \rho_B \) as 1 or \(-1\) (which may lead to biased pooled estimates, see Section 5.6.4) but
Model B estimates $\rho$ between -0.95 and 0.95 (which does not lead to biased pooled estimates, see Section 8.3.5).

- **Model B produces more appropriate pooled estimates than Model A-zero (i.e. the estimates are generally less biased with smaller MSE and more suitable coverage),** and would appear the best current option available for BRMA when the within-study correlations are unknown (see Sections 8.3.3 to 8.3.6).

These are very important findings that will help the increased and proper application of BRMA in real meta-analysis situations in practice, both to prognostic marker evidence syntheses and in general. For instance, the findings enabled me to sensibly apply Model B to the prognostic marker datasets from the neuroblastoma review and also to other clinically relevant datasets (see Section 8.5). It is a particularly important finding that Model B is more suitable than Model A-zero, which is the most common BRMA model used by researchers in practice [148], and also that Model B can even be more suitable than Model A when the within-study correlations are known. It is, however, worth emphasising that, as for Model A, there is only negligible benefit of Model B over a URMA for estimating $\beta_j$ when there is complete-case data, as in this situation the reduction in MSE and gain in precision of $\hat{\beta}_j$ are both very small on average (see Table 8.8). There are also some limitations of Model B that need to be re-emphasised here:

- **An estimate of the between-study variance cannot be obtained from Model B,** as the method is only suitable to obtain the pooled estimates (see Section 8.2.1).

- **When $\hat{\rho}$ is very close to 1 or -1, Model B can produce misleading and biased pooled estimates (e.g. see Section 8.3.2).** The use of Model B-Bayes, a Bayesian approach to Model B, and the incorporation of external information can help prevent this problem (see Section 8.6). My simulations suggest that $\hat{\rho}$ between -0.95 and
0.95 in Model B or Model B-Bayes is a ‘safe’ range to use, although further research on this issue is recommended (see Section 8.4.3).

- Model B often does not converge when the within-study variation is very large relative to the between-study variation (see Section 8.3); this is also the situation where Model A commonly estimates $\rho_B$ as 1 or −1. Hence, Model B is not suitable where a bivariate fixed-effects meta-analysis is needed, because in this situation the between-study variation is zero.

- One needs to justify the MCAR or MAR assumption in those studies only providing one outcome (see Section 8.8), as is the case for Model A (see Sections 6.2 and 6.3).

- Model B does not take into account those studies for which neither outcome is available (see Section 9.3), as is the case for Model A (see Section 6.2).

Where appropriate convergence of Model B has been obtained, the largest drawback to using the model results to form evidence-based conclusions will often be the need to assume MCAR or MAR in those studies only providing one outcome. The assumption of MCAR or MAR is a very strong conjecture, especially for prognostic marker studies (see Section 6.3.1), unless one can truly justify this from the published primary studies or from their authors. Where MCAR or MAR cannot be justified, Model B should only be used as a sensitivity analysis to assess the robustness of the URMA conclusions when the MCAR or MAR assumption is indeed valid (see Section 8.7.3). Some authors may object to this as it involves producing new pooled estimates and confidence intervals to which the URMA results can be compared, even though they may not be valid [211]. However, as long as they are treated appropriately, I believe that producing ‘adjusted’ pooled estimates alongside standard URMA estimates enables the least biased conclusions because they help to assess the robustness of URMA to the problems of missing data [203;212]. Ultimately, real-life decisions for clinical practice have to be made from evidence-based
reviews of prognostic markers, and sensitivity analyses can help make the most reliable recommendations where standard evidence-based results are limited (see Section 8.7.3 and Chapter 9).

10.3 Benefits of the research into how the bivariate framework can be used to help assess dissemination bias

Alongside the need for sensitivity analyses when missing data is assumed MCAR or MAR, there is also the need for sensitivity analyses assuming data is not missing at random (NMAR) (see Section 8.8). Data NMAR is most likely caused by dissemination bias, and this is one of the major problems affecting meta-analysis of prognostic marker studies, and indeed most other areas where evidence synthesis is required [41]. Hence, for BRMA to be useful in practice, both for meta-analysis of prognostic studies and in general, there is a real need for supplementary methods to be available to help assess the potential impact of dissemination bias on the BRMA results. The research in Chapter 9 lays the foundation toward meeting this need, something that has not been considered in detail before. In particular, I have considered two settings: (i) MCAR or MAR is assumed for the missing summary statistic in those studies providing one outcome, but NMAR is assumed in those studies providing neither outcome; and (ii) NMAR is assumed for the missing summary statistics in all studies where there are missing outcomes.

For setting (i) I have introduced how the ‘three-step approach’ can allow standard dissemination bias assessments, such as Trim and Fill and Egger’s test, following a BRMA approach similar to Model B (see Section 9.3). The main advantage of this approach is that it specifically takes into account those studies where one of the two summary statistics was known to be missing, something that is not possible following a standard URMA (see Section 9.3.2). For setting (ii), I have discussed why selection models may be the most suitable way to undertake sensitivity assessments [201], and also why the bivariate
framework can facilitate this approach (see Section 9.4). Selection models predict values of missing summary statistics for those known studies in which the required summary statistics are unavailable. Meta-analyses involving two correlated summary statistics may be particular receptive to the application of selection models because more information is likely to be available in these studies to help generate possible values of any missing summary statistics (see Section 9.4.3). However, methods need to be developed that allow a selection model to be followed by an assessment of the potential impact of dissemination bias from both: (i) known studies for which predicted summary statistics were not possible in the selection model, and (ii) unknown studies missed by the systematic review, perhaps due to publication bias (Section 9.4.4). Similar approaches to the ‘three-step approach’ may therefore be useful even after a selection model is used.

Sensitivity analyses in meta-analysis are an under-utilised and under-researched area, especially in relation to selection models and bivariate meta-analysis, and the research in Chapter 9 should help make researchers aware of this issue. Even if reporting standards improve, sensitivity analyses will play a continually important role in systematic reviews and meta-analysis of prognostic marker studies in the future (see Section 9.5), and inevitably also in other research fields, especially those involving observational studies [102]. Hence, the methods developed and discussed in Chapter 9 should hopefully provide a platform for further, novel sensitivity methods to build on, something that can only help facilitate the most reliable evidence-based clinical results in the future.

10.4 Where next for evidence synthesis of prognostic markers? Identification of the future research priorities toward clinically relevant evidence-based results

In Sections 10.1 to 10.3 I have summarised the main findings from the research undertaken in this thesis and discussed how it will help produce more statistically appropriate and clinically relevant evidence-based prognostic marker results in the future. However, there
is still much further work required, and in Figure 10.3 I summarise what I believe are the main priorities to facilitate evidence synthesis of prognostic marker studies based on my findings in this thesis. I have already mentioned the need to improve the design, clinical relevance and reporting of primary studies in Section 10.1. However, I will now briefly discuss some of the other priorities for the research that needs to follow on from this thesis.

**Figure 10.3: The main research priorities to facilitate a clinically relevant evidence synthesis of prognostic marker studies in the future**

- Continual improvements in the design, clinical relevance and reporting of primary prognostic marker studies.

- Development of sensitivity analyses to help assess the robustness of meta-analysis results to the potential problems commonly affecting meta-analysis of summary statistics from prognostic marker studies (e.g. dissemination bias, missing outcomes, heterogeneous cut-off levels and adjustment factors etc.).

- Demonstration of the clinical and statistical benefits of IPD for evidence synthesis of prognostic marker studies, to thereby help the drive toward IPD being commonly made available from primary studies.

- Development of meta-analysis methods for combining IPD with summary statistics.

- Development of decision models and cost-benefit models to help funders and leaders of evidence synthesis projects in the prognostic marker field to: (i) ascertain how much IPD to collect in addition to or instead of summary statistics, or (ii) decide whether to initiate an evidence synthesis or a new large-scale primary study, or even (iii) do neither of the above and use resources for a different project in perhaps a different field.

- Formation of collaborative groups within each research area need to work together toward prospectively planned pooled analyses for prognostic markers.

**10.4.1 Further development of sensitivity methods in meta-analysis**

In Figure 3.2 of Section 3.11 I summarised the main methodological problems affecting meta-analysis of summary statistics from prognostic marker studies. In this thesis I have only considered methods to address the specific problems of missing outcomes, missing summary statistics and the threat of dissemination bias. As well as extending these methods (see further research suggestions for Model B, the ‘three-step approach’ and
selection models in Sections 8.7.2, 9.3.2 and 9.4 respectively), there is also a need to consider the other methodological problems this thesis has not addressed (Figure 3.2). For example, how does one limit the problem of heterogeneous and often biased choice of adjustment factors and cut-off levels across studies? There are no easy answers to this question, and a sensitivity approach is again likely to be the most fruitful approach. For instance, if one assessed those studies which present results for a range of cut-off levels, one may identify a relationship between cut-off level and effect size (e.g. HR estimate). Thus, taking a similar stance as the ‘three-step approach’ (see Section 9.3.1), it may be possible use this relationship to predict estimates for a consistent cut-off level across studies. However, such an approach would undoubtedly require assumptions to be made, and many of these may not be plausible if a biased cut-off level procedure was used (see Sections 2.8.3 and 2.9.2). This again points to the potential benefit of selection models (see Section 9.4), where estimates for unavailable cut-off levels could be predicted based on the assumption that the cut-off levels reported were subject to a biased within-study selection procedure. The investigation of how to limit the cut-off problem would make particularly interesting and relevant further research, although the drive toward IPD being made available may help alleviate this problem in the long-term (see Section 2.9.2).

10.4.2 Formal demonstration of the benefits of IPD over summary statistics

I have placed great emphasis throughout this thesis about how beneficial it would be to have IPD rather than summary statistics available from primary studies (see Section 2.9) [213]. However, given the limited IPD available for this thesis, it has not been possible to actually demonstrate this. Hence, the formal demonstration of the benefits of IPD over summary statistics is, at least in the short-term, a high priority for the prognostic marker field. In particular, it is needed to reinforce the reasons why researchers of primary studies should make their IPD available.
In other research fields, the comparison between an IPD and a summary statistics evidence synthesis have tended to focus on statistical aspects, such as bias and precision of pooled estimates [100;128;214]. Alongside this, there is also the need to consider benefit from a clinical point of view. For example, consider hypothetically that clinicians in neuroblastoma were not interested in clinical heterogeneity across primary studies because their treatment strategies are based on the average person with the disease, with prognostic markers used consistently for all types of patients (e.g. for all age groups, for all stages of disease, and for all types of treatment). Other things being equal, in this situation where just an average pooled estimate is required without concern for clinical heterogeneity, evidence synthesis using summary statistics has the potential to be equally beneficial to one using IPD from both a clinical and statistical viewpoint. The example proposed here is perhaps unlikely in clinical practice, but would be more common if prognostic marker results were needed for public health policies, where decisions are made at the population level rather than the individual level. The most likely situation for clinical practice is that clinical heterogeneity is very important, perhaps because different treatment strategies are available for different risk groups. In this situation, an IPD evidence synthesis would likely be far superior both clinically and statistically given the need for IPD when exploring heterogeneity [54]. Both statistical and clinical benefits therefore need consideration when evaluating the use of IPD over summary statistics for evidence synthesis of prognostic markers.

10.4.3 Development of methods to weigh benefit against cost for an evidence synthesis using IPD rather than summary statistics

Given the potential benefits of IPD over summary statistics, it seems sensible to recommend that future evidence synthesis of prognostic marker studies should seek IPD wherever possible. Even where clinical heterogeneity is not important, the likely problem of structural/methodological heterogeneity is a strong argument toward an IPD review.
However, there are a number of issues that may restrict IPD reviews [144], and these were discussed at length in Section 2.9.5. In particular, one needs to weigh the clinical and statistical benefits of IPD against the cost to actually obtain, clean and analyse the IPD itself. For example, if the additional benefits of IPD do not outweigh the cost of obtaining the IPD then it may be hard to justify an evidence synthesis using IPD rather than summary statistics.

The need to assess benefit against cost naturally points to decision models and cost-benefit analyses to help funders and leaders of evidence synthesis projects decide the most appropriate strategies for their work [53;74]. Such models need to incorporate information and parameters about both cost and benefit, but neither the information nor the model definition will necessarily be straightforward. Parameters regarding benefit of IPD over summary statistics need to incorporate both statistical and clinical aspects (Section 10.4.2). Statistically, the improvement in precision or reduction in bias from having IPD over summary statistics should be included. In terms of clinical benefit, one needs to include parameters which acknowledge the additional ability of IPD to answer more appropriate clinical questions of interest; such information will undoubtedly require collaboration with and elicitation from clinicians in the field, and may not be straightforward [187]. Bayesian models, which can naturally incorporate elicited prior information, could be very useful for this situation, and indeed they have already been used for cost-benefit models in evidence synthesis [74]. An expected value of information (EVI) approach may also be taken in this framework [215], where the expected value of having a complete set of IPD from each primary study could be weighed against the cost needed to obtain it.

In practice, however, it may be very difficult to find and ascertain information about the cost of the proposed IPD review over and above an equivalent summary statistics review.
One could base a cost figure on those from previous IPD reviews, but unless they have been performed in exactly the same research area the figures and budgets used may vary considerably. Perhaps the major expense is the cost of multiple meetings with the authors and trialists of primary studies to firstly explain the reasoning for wanting their IPD and to then later disseminate the final evidence synthesis results [128;143;144]. Ascertaining all the specific cost issues for an IPD evidence synthesis of prognostic markers studies is non-trivial, but ultimately, as for most decisions in healthcare, an assessment of cost has to be made in order to make the most appropriate decisions for research strategy. Investigations into cost and IPD, together with the development of decision and cost-benefit models, therefore form a high priority for future evidence-based prognostic marker research, and indeed other areas of healthcare evaluation.

10.4.4 Development of meta-analysis methods to combine IPD and summary statistics

The decision about whether to choose IPD rather than summary statistics may not always be clear-cut, and instead of choosing solely IPD or solely summary statistics, it may be most beneficial from a cost and time perspective to seek IPD for some studies and summary statistics for others. Indeed, even after deciding to obtain IPD from every study it is highly unlikely that all the IPD required would be obtained, and therefore summary statistics may still be needed from some studies even when taking the full IPD approach.

Meta-analysis methods for combining IPD with summary statistics are therefore greatly needed, particularly as relatively few publications currently tackle this [142;153], although similar methods developed for other ES reasons may be applicable (e.g. the synthesis of epidemiological and toxicological evidence [216], or the synthesis of qualitative and quantitative evidence [217]). Regression models are one viable way to synthesise IPD and summary statistics, and they have been used in an 'adaptation method' which improves the
precision of pooled results by utilising the correlation between univariate and multivariate regression coefficients [142]. More generally, multi-level and multivariate models may provide a suitable way to jointly model IPD and summary statistics as they can model correlation within and between studies [153]. Adopting a Bayesian framework may again be useful here as it would allow the incorporation of expert opinion or external knowledge alongside the available data (see Chapter 7). It may be that evidence from summary statistics fits more appropriately in this framework, perhaps to form sensible prior beliefs for inclusion alongside the IPD [216;217]. Also, observed relationships from the IPD collected (e.g. between outcomes, markers) may be modelled and could be used to predict estimates of desired but missing information from the studies only providing summary statistics. Such a sensitivity analysis approach may also facilitate selection models and dissemination bias assessments, like those I suggested in Chapter 9 [208]. The development of novel methods for the synthesis of IPD and summary statistics from prognostic marker studies would therefore make particularly pertinent further research.

10.4.5 Development of methods to weigh benefit against cost for an evidence synthesis rather than one new, high quality primary study

There is an alternative option to performing an evidence synthesis of prognostic marker studies. For example, what if an evidence-based review was likely to be severely limited using summary statistics, and yet an equivalent IPD approach would also be very costly and may still be uninformative due to poor quality primary studies? In this situation, a sensible decision may be to not go ahead with a systematic review and instead put resources toward one new, high quality primary study. For example, the poor quality of the neuroblastoma literature means that even if IPD were obtained for every study an evidence-based review may still not be able to produce clinically relevant or reliable results. Hence, in the long-run it may be better to devote time and resources toward the
initiation of one new, well-designed, clinically relevant, high quality primary study involving a large number of patients and focused directly on the clinical questions that need addressing. This is a realistic option that funders and leaders of evidence synthesis projects need to seriously consider, and cost-benefit or pay-back models are again a necessity for helping inform such decisions (see Section 10.4.3) [218]. Of course such models may also include the option to do neither an evidence synthesis nor a new primary study, and instead put the resources available into a different research area, maybe outside the prognostic marker field. This is another pertinent decision facing funding bodies on a regular basis, and further research is required to help them make the correct choices.

10.4.6 Initiation of prospectively planned pooled analyses

If one did opt for a new primary study rather than an evidence synthesis, then there may still be the dilemma of what should be clinically recommended if such a new study produces 'surprising' results which are inconsistent with the findings of previous studies. To help address this, when new primary prognostic marker studies are prioritised, the logical next step is to ensure such primary research fits into an evidence synthesis at a future time point. I have taken this approach in my reporting guidelines for new primary prognostic marker studies (see Figure 2.11 in Section 2.9), where alongside suggestions for improving the dissemination of results in the published article (e.g. present a HR and its confidence interval) I have also suggested how the study should seek to aid evidence-based reviews (e.g. by making IPD available). However, perhaps the most ideal approach is to go one step further and recommend the initiation of collaborative groups to work toward a number of high quality primary studies with the collective, long-term aim of pooling together the IPD from each study to formulate evidence-based results. Such a prospectively planned pooled analysis is the most exciting way forward for prognostic marker research (see Section 2.9.7), and may be important in other areas of epidemiological research [93].
In this situation, the primary studies themselves could be developed with a prospective meta-analysis in mind, so that it would be pre-specified which outcomes, markers, treatments, and other factors all the primary studies should assess. It would also allow authors of primary studies to know right from the beginning that their IPD would be needed, which should, one would hope, make them more amenable to recording and maintaining high quality IPD that they are also willing to make available at the end of their study. Of course, cost issues again need to be considered together with the time-scale of such a large project. One of the main benefits of a retrospective evidence-based review is that they can often produce results much more quickly than is possible from a new primary study. Clinicians obviously need to know answers as soon as possible, but where current prognostic marker studies are just not up to the necessary standards there may be little option but to prioritise new primary research. Therefore, to place primary research in the context of prospectively planned pooled analyses is perhaps an ideal way forward.

Research areas that are just developing should particularly aim for prospectively planned pooled analyses in order to avoid the problems inherent in prognostic marker research, where hundreds of poor quality and poorly targeted individual studies have been initiated separately from each other, making it difficult to produce evidence-based results. Studies in genetic epidemiology are one area that should particularly take heed from prognostic marker research. In the coming years, studies of the human genome are likely to identify hundreds of new individual and combinations of genes that are associated with one or more diseases [219;220]. In fact, many of these will have the potential to be new prognostic markers and so unless prospectively planned pooled analyses are considered from the outset the problems demonstrated in the neuroblastoma review of Chapter 2 are going to repeat themselves when evidence synthesis of genome studies are commonly desired in a
few years time. Indeed, meta-analysis of genome studies are already becoming prevalent in the literature [221-223].

10.5  **A possible strategy for future prognostic marker research**

As the future research priorities of Section 10.4 are met, those funders and leaders of prognostic marker projects will also need to adopt new research strategies that incorporate the improvements made and continue the move toward clinically relevant evidence-based results. As one possible suggestion, and to highlight the processes and decisions necessary, Figure 10.4 outlines a strategy for future prognostic marker research where clinically relevant evidence-based results are the ultimate objective. The strategy shown highlights the two possible paths toward evidence-based results; firstly, one could perform a formal systematic review and meta-analysis of previous studies, or alternatively one could perform a new high quality primary study and work alongside other such new studies toward prospectively planned pooled analyses. The choices taken at each stage should be justified, using formal cost-benefit models if and when these become available. Further development and evaluation of this and other such prognostic marker research strategies would make particularly relevant further work.
Consider the motivating clinical questions that need to be addressed by the prognostic marker research

Cost-benefit analysis to decide between:

- A formal evidence synthesis of previous studies
- A new primary study

Appropriate initiation of the approach chosen:

- An appropriate systematic review of the literature to identify all those published and unpublished studies of clinical relevance
- A well-designed and appropriately targeted study which is related to specific treatments and includes assessment of new markers against those currently used in practice (most likely a Phase III(b) study). The study should be designed and conducted alongside other such new studies, working toward prospectively planned pooled analyses

Cost-benefit analysis to decide how much IPD to obtain in addition to or instead of summary statistics

Appropriate statistical analyses followed by appropriate reporting in any subsequent publications. The IPD should then be made available for the prospectively planned pooled analysis and perhaps kept in a data repository.

Formation of the most appropriate evidence-based results:

(1) An appropriate meta-analysis to produce evidence-based results using one or more of:
   (i) just IPD
   (ii) just summary statistics
   (iii) both IPD and summary statistics
(2) One or more sensitivity analyses to assess the robustness of the results in (1) to any problems affecting the data available, such as dissemination bias, missing summary statistics, missing outcomes etc.

(1) An appropriate meta-analysis to produce evidence-based results using the IPD available from all the primary studies as planned
(2) One or more sensitivity analyses to assess the robustness of the results from (1) to any problems with the IPD, such as missing patient data, missing outcomes, missing marker levels, dissemination bias etc.

N.B. the dotted line shows the option for initiating a new primary study even after originally choosing the systematic review approach; for instance, it may become apparent (e.g. from the poor quality literature identified, or from the cost-benefit analysis of obtaining IPD) that the most appropriate, clinically relevant prognostic marker results would be obtained by initiating a new primary study with the long-term aim of a prospectively planned pooled analysis.
10.6 Conclusion

Evidence synthesis is clearly a crucial part of healthcare evaluation and should, wherever possible, be a vital component in the formation of clinical practice and public health policies. However in this thesis, by performing an empirical investigation in neuroblastoma, I have identified and formally demonstrated the numerous methodological difficulties that limit clinically relevant evidence-based results for prognostic markers. Clinicians, statisticians and other researchers need to work together and collectively address this, because the evidence-based application of prognostic markers is the most appropriate way to use these tools in the management and treatment of patients in clinical practice [7;78;94].

To help initiate a drive to meet this need, the research in this thesis has sought to address a number of the methodological issues identified in order facilitate future evidence syntheses of prognostic marker studies. In particular, I have developed reporting guidelines for primary prognostic studies which complement a growing movement toward better quality primary prognostic marker research [86]. I have also introduced, developed and illustrated the use of bivariate meta-analysis models and sensitivity methods to specifically limit the serious problems of unreported outcomes, biased within-study reporting and dissemination bias [208]. Importantly, all these methods have been critically assessed during the thesis and they have been shown to potentially produce more reliable evidence-based results than is currently possible from standard meta-analysis methods.

The research in this thesis therefore makes a positive contribution toward more reliable and clinically relevant evidence-based results for prognostic markers in the future. However, numerous methodological issues remain and so this final chapter has outlined the further research priorities to help ensure the move toward an evidence-based use of prognostic markers continues.
Appendix A

Appendix A1: WinBUGS version 1.3 syntax for fitting a univariate random-effects meta-analysis (URMA)

The syntax is shown below and relates to equation (1.13) in Section 1.8.3

Model
{
  for(i in 1:n)
  {
    surv[i] ~ dnorm(theta[i],prec[i])
    prec[i]<- 1/variance[i]
    theta[i] ~ dnorm(beta,invtausq)
  }
  invtausq~dgamma(0.001, 0.001)
  tausq <- 1/invtausq
  beta ~ dnorm(0.0,1.0E-6)
  hr<-exp(beta)
}

Data
list(n=6 =0, variance=c(0.57, 0.31, 0.412, 0.405, 0.093, 0.581),
    surv=c(0.031, 1.21, 0.85, 1.308, 1.56, 1.592))

Inits
list(beta=0, invtausq=10, theta=c(0,0,0,0,0,0))

Appendix A2: Forest plots for the neuroblastoma meta-analyses in Section 2.4.2

(i) Homovanillic Acid (HVA) for overall survival

2.5 SD = 2.5 standard deviations from mean normal value
N.B. No DFS meta-analysis was possible.
N.B. The DFS estimate of triploid versus tetraploid from paper 410 was omitted because it was not classifying the same groups as the other estimates which compared diploid (DNA index = 1) versus others; see Riley et al. for further details [1].
(iii) Chromosome 1p (Ch1p)

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Paper</th>
<th>n</th>
<th>u/a</th>
<th>age</th>
<th>stage</th>
</tr>
</thead>
<tbody>
<tr>
<td>DFS</td>
<td>168</td>
<td>91</td>
<td>u</td>
<td>all</td>
<td>1,2,3</td>
</tr>
<tr>
<td>280</td>
<td></td>
<td>89</td>
<td>u</td>
<td>all</td>
<td>all</td>
</tr>
<tr>
<td>386</td>
<td></td>
<td>74</td>
<td>a</td>
<td>unknown</td>
<td>1,2,3</td>
</tr>
<tr>
<td>295</td>
<td></td>
<td>32</td>
<td>u</td>
<td>all</td>
<td>all</td>
</tr>
<tr>
<td>463</td>
<td></td>
<td>52</td>
<td>u</td>
<td>all</td>
<td>unknown</td>
</tr>
<tr>
<td>506</td>
<td></td>
<td>238</td>
<td>u</td>
<td>all</td>
<td>all</td>
</tr>
<tr>
<td>80</td>
<td></td>
<td>15</td>
<td>u</td>
<td>unknown</td>
<td>all</td>
</tr>
<tr>
<td>171</td>
<td></td>
<td>30</td>
<td>u</td>
<td>all</td>
<td>all</td>
</tr>
<tr>
<td>279</td>
<td></td>
<td>50</td>
<td>u</td>
<td>all</td>
<td>all</td>
</tr>
</tbody>
</table>

Pooled result (95% CI)

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Paper</th>
<th>n</th>
<th>u/a</th>
<th>age</th>
<th>stage</th>
</tr>
</thead>
<tbody>
<tr>
<td>OS</td>
<td>338</td>
<td>53</td>
<td>u</td>
<td>all</td>
<td>all</td>
</tr>
<tr>
<td>111</td>
<td></td>
<td>156</td>
<td>u</td>
<td>all</td>
<td>all</td>
</tr>
<tr>
<td>109</td>
<td></td>
<td>49</td>
<td>u</td>
<td>all</td>
<td>all</td>
</tr>
<tr>
<td>544</td>
<td></td>
<td>53</td>
<td>u</td>
<td>all</td>
<td>2,3,4,6</td>
</tr>
<tr>
<td>93</td>
<td></td>
<td>21</td>
<td>u</td>
<td>all</td>
<td>all</td>
</tr>
<tr>
<td>80</td>
<td></td>
<td>15</td>
<td>u</td>
<td>unknown</td>
<td>all</td>
</tr>
<tr>
<td>106</td>
<td></td>
<td>48</td>
<td>u</td>
<td>all</td>
<td>all</td>
</tr>
<tr>
<td>173</td>
<td></td>
<td>30</td>
<td>u</td>
<td>all</td>
<td>all</td>
</tr>
<tr>
<td>279</td>
<td></td>
<td>50</td>
<td>u</td>
<td>all</td>
<td>1,2,3,4</td>
</tr>
<tr>
<td>506</td>
<td></td>
<td>237</td>
<td>u</td>
<td>all</td>
<td>all</td>
</tr>
<tr>
<td>198</td>
<td></td>
<td>91</td>
<td>u</td>
<td>all</td>
<td>1,2,3</td>
</tr>
</tbody>
</table>

Pooled result (95% CI)

N.B. For all these estimates the groups were 1p deletion not present versus 1p deletion present.

(iv) Vanillylmandelic Acid (VMA) for overall survival (OS)

<table>
<thead>
<tr>
<th>paper</th>
<th>n</th>
<th>u/a</th>
<th>age</th>
<th>stage</th>
<th>cut-off</th>
</tr>
</thead>
<tbody>
<tr>
<td>297</td>
<td>269</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>2.5 SD</td>
</tr>
<tr>
<td>544</td>
<td>74</td>
<td>u</td>
<td>all</td>
<td>2,3,4,6</td>
<td>11</td>
</tr>
<tr>
<td>256</td>
<td>32</td>
<td>u</td>
<td>all</td>
<td>3,4</td>
<td>4.6</td>
</tr>
</tbody>
</table>

Pooled result (95% CI)

2.5 SD = 2.5 standard deviations from mean normal value

N.B. No DFS meta-analysis was possible.
(v) Lactate Dehydrogenase (LDH)

<table>
<thead>
<tr>
<th>outcome</th>
<th>paper</th>
<th>n</th>
<th>u/a</th>
<th>age</th>
<th>stage</th>
<th>cut-off</th>
</tr>
</thead>
<tbody>
<tr>
<td>DFS</td>
<td>280</td>
<td>97</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>unknown</td>
</tr>
<tr>
<td></td>
<td>496</td>
<td>122</td>
<td>a</td>
<td>all</td>
<td>all</td>
<td>1500</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>164</td>
<td>u</td>
<td>all</td>
<td>1,2,3</td>
<td>III</td>
</tr>
<tr>
<td></td>
<td>327</td>
<td>164</td>
<td>u</td>
<td>all</td>
<td>1,2,3</td>
<td>II</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>98</td>
<td>a</td>
<td>all</td>
<td>4</td>
<td>unknown</td>
</tr>
<tr>
<td></td>
<td>all</td>
<td>all</td>
<td>all</td>
<td>all</td>
<td>all</td>
<td>all</td>
</tr>
<tr>
<td>Pooled result (95% CI)</td>
<td>all</td>
<td>all</td>
<td>all</td>
<td>all</td>
<td>all</td>
<td>all</td>
</tr>
</tbody>
</table>

| OS      | 297   | 268| u   | all | all   | 1500   |
|         | 14    | 126| u   | all | 1,2,3,4| 1500  |
|         | 16    | 142| u   | all | 4     | 1500   |
|         | 199   | 134| u   | all | all   | 1500   |
|         | 46    | 92 | u   | all | all   | 1500   |
| Pooled result (95% CI) | all | all | all | all | all | all |

Hazard Ratio (log-scale)

(vii) CD44 for disease-free survival (DFS)

<table>
<thead>
<tr>
<th>paper</th>
<th>n</th>
<th>u/a</th>
<th>age</th>
<th>stage</th>
<th>cut-off</th>
</tr>
</thead>
<tbody>
<tr>
<td>276</td>
<td>52</td>
<td>u</td>
<td>unknown</td>
<td>all</td>
<td>neg or pos</td>
</tr>
<tr>
<td>315</td>
<td>121</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>neg or pos</td>
</tr>
<tr>
<td>384</td>
<td>52</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>neg or pos</td>
</tr>
<tr>
<td>Pooled result (95% CI)</td>
<td>all</td>
<td>all</td>
<td>all</td>
<td>all</td>
<td>all</td>
</tr>
</tbody>
</table>

Hazard Ratio (log-scale)

N.B. No overall survival (OS) analysis was possible.
2N = twice the normal range; N.B. The OS estimates from papers 185 and 259 were excluded from the meta-analysis because they compared different groups of NSE patients than the other estimates; see Riley et al. for further details [1].

(vii) Neuron-specific enolase (NSE)

(viii) TrkA

N.B. Of the 4 estimates from paper 276 for DFS the result for 1+ versus 2+ was chosen as this had the most no. of patients; see Riley et al. for further details [1].
(ix) Ferritin

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Paper</th>
<th>n/u</th>
<th>u/a</th>
<th>age</th>
<th>stage</th>
<th>cut-off</th>
</tr>
</thead>
<tbody>
<tr>
<td>DFS</td>
<td>280</td>
<td>78</td>
<td>u</td>
<td>all</td>
<td>1.4-4.0</td>
<td>142</td>
</tr>
<tr>
<td></td>
<td>317</td>
<td>79</td>
<td>u</td>
<td>all</td>
<td>1.4-4.0</td>
<td>150</td>
</tr>
<tr>
<td></td>
<td>180</td>
<td>110</td>
<td>u</td>
<td>all</td>
<td>1.4-4.0</td>
<td>141</td>
</tr>
<tr>
<td></td>
<td>356</td>
<td>228</td>
<td>u</td>
<td>all</td>
<td>1.4-4.0</td>
<td>142</td>
</tr>
<tr>
<td>Pooled result (95% CI)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.26 (2.42, 7.53)</td>
</tr>
<tr>
<td>OS</td>
<td>297</td>
<td>254</td>
<td>u</td>
<td>all</td>
<td>1.4-4.0</td>
<td>150</td>
</tr>
<tr>
<td></td>
<td>189</td>
<td>254</td>
<td>u</td>
<td>all</td>
<td>unknown</td>
<td></td>
</tr>
<tr>
<td></td>
<td>144</td>
<td>77</td>
<td>u</td>
<td>all</td>
<td>all</td>
<td>varies for age</td>
</tr>
<tr>
<td>Pooled result (95% CI)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.74 (1.92, 3.91)</td>
</tr>
</tbody>
</table>

Hazard Ratio (log-scale)

(x) Multi-Drug Resistance (MDR)

<table>
<thead>
<tr>
<th>paper</th>
<th>n/u</th>
<th>age</th>
<th>stage</th>
<th>cut-off</th>
</tr>
</thead>
</table>
| DFS   | 297 | 78  | all   | unknown
|       | 306 | 84  | all   | expression or age
|       | 144 | 77  | all   | expression or age
|       | 256 | 41  | all   | neg or pos
|       | 520 | 14  | all   | neg or pos
| Pooled result (95% CI) |       |     |     |       |
| OS    | 297 | 78  | all   | unknown
|       | 306 | 84  | all   | expression or age
|       | 144 | 77  | all   | expression or age
|       | 256 | 41  | all   | neg or pos
|       | 520 | 14  | all   | neg or pos
| Pooled result (95% CI) |       |     |     |       |

Hazard Ratio (log-scale)

N.B. For the two OS results available from paper 117 the one for a cut-off of 0 versus 0-20 was used; for the two DFS results from paper 389 the one for a cut-off of '80th percentile' was used; the DFS result from paper 122 was omitted because the worse prognostic group was unclear, and the OS estimates from papers 532 and 107 were omitted because they compared different groups of patients to the other estimates; see Riley et al. for further details [1].

R.D. Riley, Ph.D. Thesis
Appendix A

379
### VMA: HVA Ratio

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Papers</th>
<th>n</th>
<th>u/a</th>
<th>Age</th>
<th>Stage</th>
<th>Cut-off</th>
</tr>
</thead>
<tbody>
<tr>
<td>DFS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>230</td>
<td>97</td>
<td>unknown</td>
<td>1,2,3</td>
<td>0.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>246</td>
<td>32</td>
<td>all</td>
<td>all</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>337</td>
<td>52</td>
<td>all</td>
<td>3,4,4S</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

**Pooled result (95% CI):**

**Hazard Ratio (log-scale):**

- DFS: 0.35 (0.17, 0.72)
- OS: 0.44 (0.18, 1.06)

---

R.D. Riley, Ph.D. Thesis
Appendix A 380
Appendix A3: Publication in Clinical Cancer Research arising from the neuroblastoma meta-analysis results in Section 2.4.2


RR oversaw the entire review. RR developed the search strategy with advice from clinicians SB and DH. RR performed all the literature searching and classification, with samples checked by DH and SB. RR performed at least 90% of the data extraction and indirect estimation of the hazard ratios with KA, AS, PL and DJ performing the remainder. RR performed all the meta-analyses and dissemination bias assessments. SB and DH provided clinical advice throughout, and helped ascertain the clinical importance of the meta-analysis results alongside the statistical interpretation provided by RR. BY and AW provided advice about the psychosocial and economic results respectively on the small number of occasions these were identified in the neuroblastoma literature. RR wrote the paper, with comments and suggestions provided on various draft versions by all the other authors.
SPECIAL NOTE

ITEM SCANNED AS SUPPLIED
PAGINATION IS AS SEEN
A Systematic Review of Molecular and Biological Tumor Markers in Neuroblastoma

Richard D. Riley,1 David Heney,2 David R. Jones,1 Alex J. Sutton,1 Paul C. Lambert,4 Keith R. Abrams,1 Bridget Young,3 Alan J. Walloo,4 and Susan A. Burchill5

Departments of 1Health Sciences, 2Medical Education, University of Leicester, Leicester; 3Department of Psychology, University of Hull, Hull; 4School of Health and Related Research, University of Sheffield, Sheffield; and 5Cancer Research United Kingdom Clinical Centre, St. James's University Hospital, Leeds, United Kingdom

Abstract

Purpose: The aim of this study was to conduct a systematic review, and where possible meta-analyses, of molecular and biological tumor markers described in neuroblastoma, and to establish an evidence-based perspective on their clinical value for the screening, diagnosis, prognosis, and monitoring of patients.

Experimental Design: A well-defined, reproducible search strategy was used to identify the relevant literature from 1966 to February 2000.

Results: A total of 428 papers studying the use of 195 different tumor markers in neuroblastoma were identified. Small sample sizes, poor statistical reporting, large heterogeneity across studies (e.g., in cutoff levels), and publication bias limited meta-analysis to the area of prognosis only; MYCN, chromosome 1p, DNA index, vanillylmandelic acid: homovanillic acid ratio, CD44, Trk-A, neuron-specific enolase, lactate dehydrogenase, ferritin, and multidrug resistance were all identified as potentially important prognostic tools.

Conclusions: This systematic review forms a knowledge base of the tumor markers studied thus far in neuroblastoma, and has identified some of the most important prognostic markers, which should be considered in future research and treatment strategies. Importantly, the review has also highlighted some general problems across primary tumor marker studies, in particular poor and heterogeneous reporting. These need to be addressed to allow better clinical interpretation and enable more appropriate evidence-based reviews in the future. In particular, collaboration of cancer research groups is needed to enable bigger sample sizes, standardize methods of analysis and reporting, and facilitate the pooling of individual patient data.

Introduction

Neuroblastoma is a neuroblastic tumor of the primordial neural crest and is the most common extracranial solid tumor of childhood, comprising between 8 and 10% of all childhood cancers. It is an enigmatic tumor demonstrating diverse clinical and biological characteristics and behavior (1). Tumors may regress spontaneously, reflecting induction of apoptosis or differentiation, or they may exhibit extremely malignant behavior with very low cure rates. The spectrum of clinical behavior suggests that genetic, biological, and morphological features may be useful markers to stratify children with this disease for the most appropriate management. Knowledge of prognostic markers may also help understand the genesis of this disease.

Neuroblastoma is predominantly a disease of the first decade with ~80% of children presenting at <4 years old; median age is 22 months. The incidence in the United Kingdom and the United States is ~1 in 7000 live births, and there is slight sex predominance in most series with a male-to-female ratio of 1.2:1. The disease accounts for 15% of all childhood cancer deaths, indicating the poor prognosis of many of the tumors (2–4). Children with stage 1, 2, or 4 s disease, or presenting in the first year of life have a good prognosis. In contrast, children (≥1 year of age) with stage 3 and 4 disease have 3-year survival rates of 50% and 15%, respectively. Most children present over the age of 1 year with metastatic (stage 4) disease; this group has an overall survival of 10–20% (3, 4).

A number of genetic and biological features have been investigated in recent years in an effort to improve the understanding of the behavior of neuroblastoma and to identify tumor markers that would improve cure rates by facilitating the screening, diagnosis, prognosis, or monitoring of patients. In particular, many prognostic studies have identified many tumor markers associated with overall or disease-free survival, including MYCN copy number, ploidy, and deletion or loss of heterozygosity of chromosome 1p and gain of chromosome 17q. However, it has proved difficult to identify which prognostic markers are the most useful, reflecting the complex nature of the tumor and the lack of large prospective clinical outcome studies.

The aim of this study was to conduct a systematic review, and where possible meta-analyses, of molecular and biological tumor markers described in neuroblastoma, and to establish an evidence-based perspective on their clinical value for the screening, diagnosis, prognosis, and monitoring of patients. This should facilitate identification of the most useful tumor markers for clinical management and the development of future research strategies by: (a) establishing the importance of the markers studied; and (b) identifying the markers that warrant additional investigation.
### Materials and Methods

The systematic review followed the guidelines contained in NHS Centre for Reviews and Dissemination (1996), and its underlying philosophy was to maintain breadth, synthesize the evidence qualitatively, and then, only where appropriate, use quantitative methods, making procedures explicit and transparent throughout (5).

#### Search Strategy.

The three on-line bibliographic databases Medline, Embase, and Cancerlit were chosen as a basis for identifying the relevant literature from 1966 to February 2000. An iterative procedure was used to develop an optimal search strategy, which culminated in the use of three important sets of keywords in the strategy (Table 1). The keywords in {Neuroblastoma} related to the family of this disease, whereas those in {Tumor Marker} included the named markers thought a priori to be potentially important. The set {Clinical Area} included more specific terms for the clinical use of markers in children.

A paper was included if a word from {Neuroblastoma}, a word from {Tumor Marker}, and a word from {Clinical Area} were included anywhere in the paper.

Three investigators independently performed the assessment of the papers. All three had previous experience of identifying relevant tumor marker literature for a systematic review and subsequent meta-analysis (6, 7). Furthermore, the second investigator is a pediatric oncology consultant, with a special interest in neuroblastoma, and the third investigator is a translational scientific research fellow, with a special interest in small round cell cancers of childhood. Both of these investigators held regular meetings with the first investigator about the review process and the assessment of the literature.

The first investigator read the available abstract to classify each paper into one of three categories: "relevant," "uncertain," or "not relevant." The second and third investigators, who had more background clinical knowledge in the research area,
checked all of the abstracts where classification was uncertain, and \( \sim 10\% \) of the papers in each of the relevant and not relevant categories. Copies of all of the papers classified as relevant, together with all of the papers in which relevance remained unclear after assessment of abstracts by the three investigators, were obtained and then read thoroughly to make a final decision as to their inclusion.

**Inclusion.** To be included in the systematic review a paper had to provide a quantitative result or give tabulated individual patient data (IPD) evaluating the use of a tumor marker in neuroblastoma. The paper had to be based on a primary research study of humans relevant to the clinical area of screening, diagnosis, prognosis, or monitoring. There was no restriction on age of patients in the study, although \( \sim 90\% \) of papers included just 0–18-year-olds. The criteria for classifying the four clinical areas was that the paper had to present data in the form of summary statistics or IPD for: (a) screening: the use of tumor markers to screen an apparent healthy population; (b) diagnosis: tumor marker levels considered of diagnostic value; (c) prognosis: tumor marker levels at a measured point in time with relation to the outcome of patients at the end of a specific follow-up period; and (d) monitoring: tumor marker levels taken repeatedly during a follow-up period with relation to disease status over that period.

**Exclusion.** Papers that reported only laboratory work, methodology for identifying new markers, or results from animal studies were excluded. Review articles and foreign language papers were also excluded. Histological characteristics of tumors [such as the presence of differentiated ganglia in neuroblastoma (Shimada index)] were not included in the markers reviewed.

**Information Extracted.** From the included papers, information was extracted on the tumor marker used and to which clinical area it related: screening, diagnosis, prognosis, or monitoring. Among the covariate information extracted from each paper on prognosis was whether survival was overall (OS) or disease-free (DFS), the marker cutoff level if applicable and, if so, the total number of patients and deaths within each high and low subgroup. The age range and stages of neuroblastoma disease represented by the patients in each study were also recorded, as these were known *a priori* to be important prognostic clinical features (8).

**Meta-Analysis and Assessment of Publication Bias.** Meta-analysis was performed, where possible, to combine all of the relevant results found from the literature search (9). For each of the areas of screening, diagnosis, and monitoring, only those tumor markers on which 3 or more papers provided data were considered. For the area of prognosis, due to the many prognostic markers and prognostic studies identified, meta-analysis was limited to those reported in \( \geq 10 \) papers. Both fixed and random effects meta-analyses were used, with the latter preferred if there was evidence of heterogeneity. Meta-analysis for clinically relevant subgroups of patients (*e.g.*, age <1 and stage 4 disease) was also considered, but was only performed where sufficient data were available.

For the meta-analysis of data from the prognosis papers, the extraction of the log, (hazard ratio; HR) and its variance was the desired target. These statistics were chosen because they provide an important comparative estimate of the risk of death/disease recurrence between two groups of patients.

It was common for a paper to report more than one prognostic result by relating one or more markers to OS and/or DFS, and also by providing unadjusted and/or adjusted results (*e.g.*, adjusted for age and stage of disease). Estimates of the log, (HR) and its variance comparing two groups defined by a single marker level were sought from all of the OS and DFS reports using the methods described by Parmar *et al.* (10). An unadjusted estimate was preferred for each report, as adjusted results are likely to be highly inconsistent in the factors for which adjustment are made (11). An adjusted estimate was sought in the absence of an unadjusted result.

An assessment of the publication bias in the prognosis literature was made by application of two appropriate statistical tests to the *MYCN* results (12, 13). The Trim and Fill method was also used to assess the likely impact of publication bias on the pooled *MYCN* results (14).

**Economic and Psychosocial Effects of Tumor Markers.** A set of keywords was used to screen the abstracts of the papers classified as relevant for any economic or psychosocial results relating to the use of a tumor marker in neuroblastoma for any of the clinical areas (Table 1).

**Results**

**Literature Search Results.** We identified 3415 papers from the searches; 1536 were first identified in Medline, an additional 473 in Embase, and then an additional 1406 from Cancerlit. These were classified by the three investigators (Fig. 1). The second and third investigator agreed that 85.7% of a sample of the first investigator’s relevant papers were indeed relevant or uncertain (42 of 49). They also agreed that 193 (86.9%) of a sample of 222 not-relevant papers checked were indeed not relevant, and of the 29 others classified 8 papers as relevant and 21 papers as uncertain. After obtaining and reading the entire articles, 15 of these 21 uncertain papers were ultimately classified as not relevant. Thus, 208 of the first investigator’s 222 not-relevant papers sampled by the other investigators were correctly classified (93.7%). Overall, 428 papers were considered relevant and included in our review (Fig. 1; complete list of these references available on the internet).6

**Tumor Markers Identified Overall and Within Each Clinical Area.** A total of 195 different tumor markers were studied in these 428 papers in relation to the screening, diagnosis, prognosis, or monitoring of neuroblastoma (Table 2). There were 49 different papers on screening, 288 on diagnosis, 260 on prognosis, and 51 on monitoring; 201 of the 428 papers covered two or more clinical areas.

**Screening.** The review identified 49 papers that gave quantitative data relating to the use of tumor markers in screening and potentially considered the evaluation of a population-based screening program for neuroblastoma. These papers covered programs established in geographical regions of Austria, Canada, France, Germany, Japan, and the United Kingdom.

---

6 Internet address: http://www.prw.le.ac.uk/epidemio/personal/rd3/paed2.html.
The studies considered a variety of outcomes including: 
(a) feasibility/uptake rate; (b) the number of false-positive and false-negative cases; (c) incidence; (d) stage distribution; and (e) mortality. In terms of outcomes (c), (d), and (e), some studies have undertaken, or are designed to enable in the future, a comparison between screened and control (nonscreened) populations. Unfortunately, a quantitative synthesis of results was not feasible given the heterogeneity in how and which outcomes were reported. However, a qualitative assessment suggested that considerable uncertainty still surrounds whether population-based screening for neuroblastoma is cost-effective overall, and, if so, the optimal age at which to screen, and also the optimal screening strategy, i.e., one-stage or multistage. Recent studies have shown that early screening (before 6 months of age) is not informative (see "Discussion").

Diagnosis. It was not possible to perform a meta-analysis of the data from the diagnosis papers, because the results mostly only compared the number of neuroblastoma patients with high positive marker levels to those with low/negative levels respectively. Marker levels from patients with neuroblastoma were rarely compared with those from a sample of healthy controls in the diagnosis papers, e.g., none of the 22 papers reporting levels of serum lactate dehydrogenase at diagnosis compared patients with neuroblastoma to healthy controls. Current clinical practice suggests that urinary catecholamines are important for the differential diagnosis of neuroblastoma from other small round cell tumors; however, the poor quality of the published literature prevented a quantitative evaluation of this practice.

Prognosis. The 12 most commonly studied prognostic markers were each selected for an in-depth study to establish their individual value as a prognostic tool; each marker was studied in ≥10 prognosis papers (Table 3). The prognostic value of CD44 expression was also evaluated, because all of its 8 prognostic studies were contained within those papers of the
Table 2  List of tumor markers in neuroblastoma that were identified by the systematic review together with the number of papers overall and within each clinical area

<table>
<thead>
<tr>
<th>Tumor Marker</th>
<th>Overall</th>
<th>Screening</th>
<th>Diagnosis</th>
<th>Prognosis</th>
<th>Monitoring</th>
</tr>
</thead>
<tbody>
<tr>
<td>MYCN</td>
<td>201</td>
<td>7</td>
<td>148</td>
<td>151</td>
<td>9</td>
</tr>
<tr>
<td>VMA</td>
<td>125</td>
<td>44</td>
<td>78</td>
<td>45</td>
<td>18</td>
</tr>
<tr>
<td>HVA</td>
<td>105</td>
<td>38</td>
<td>64</td>
<td>35</td>
<td>16</td>
</tr>
<tr>
<td>DNA index/ploidy/diploidy/triploidy/aneuploid/hyperdiploidy</td>
<td>56</td>
<td>5</td>
<td>37</td>
<td>44</td>
<td>1</td>
</tr>
<tr>
<td>Chromosome 1p or chromosome 1p36</td>
<td>47</td>
<td>4</td>
<td>34</td>
<td>40</td>
<td>1</td>
</tr>
<tr>
<td>Ferritin or isoferritin</td>
<td>49</td>
<td>3</td>
<td>36</td>
<td>33</td>
<td>5</td>
</tr>
<tr>
<td>NSE</td>
<td>45</td>
<td>2</td>
<td>33</td>
<td>28</td>
<td>9</td>
</tr>
<tr>
<td>LDH</td>
<td>32</td>
<td>1</td>
<td>22</td>
<td>26</td>
<td>4</td>
</tr>
<tr>
<td>Dopamine</td>
<td>24</td>
<td>2</td>
<td>22</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>TrkA (nerve growth factor receptor)</td>
<td>25</td>
<td>0</td>
<td>16</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>Adrenaline/epinephrine</td>
<td>15</td>
<td>0</td>
<td>15</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Multidrug resistance/associated protein/p-glycoprotein</td>
<td>16</td>
<td>0</td>
<td>7</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>Nonadrenaline/noradrenaline/norepinephrine</td>
<td>13</td>
<td>0</td>
<td>13</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>CD44</td>
<td>10</td>
<td>0</td>
<td>7</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>Neuropeptide Y</td>
<td>12</td>
<td>0</td>
<td>10</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>Tyrosine hydroxylase</td>
<td>12</td>
<td>0</td>
<td>11</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Chromosome 17q</td>
<td>11</td>
<td>0</td>
<td>9</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>Ha-ras/P21/H-ras/c-ha-ras</td>
<td>11</td>
<td>0</td>
<td>8</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>Telomerase/Telomeric repeats</td>
<td>11</td>
<td>0</td>
<td>6</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>Chromosome 14q</td>
<td>8</td>
<td>0</td>
<td>6</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>GD2 ganglioside</td>
<td>8</td>
<td>0</td>
<td>7</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Si100</td>
<td>7</td>
<td>1</td>
<td>5</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Chromosome 11q</td>
<td>6</td>
<td>0</td>
<td>5</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>Low affinity nerve growth receptor (LNGFR)</td>
<td>6</td>
<td>0</td>
<td>3</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>Metanephrine</td>
<td>6</td>
<td>0</td>
<td>6</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>TrkC</td>
<td>6</td>
<td>0</td>
<td>3</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>3-methoxy-4-hydroxyphenyl glycol</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4-hydroxy-3-methoxymandelic acid</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Dihydroxyphenylalanine</td>
<td>5</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Dopamine β hydroxylase</td>
<td>5</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Proliferating cell nuclear antigen/proliferation index/Ki67/KiS5 protein</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

*other 12 markers to be evaluated. Hence, 13 markers overall were evaluated as a prognostic tool. These 13 markers were studied in 211 (81.2%) of all of the prognosis papers. Within these there were 575 reports of prognostic power assessment, where levels of 1 of these 13 tumor markers were related to OS or DFS by summary statistics or IPD. Weakness of reporting, analysis, and presentation of results meant that only 204 (35.5%) estimates of both the loge (HR) and its variance could be extracted. Meta-analyses were additionally restricted by large variability in both clinical and statistical factors relating to the 204 estimates. For example, for the marker *MYCN*, 94 estimates of the loge (HR) and variance were obtained but these involved 9 different cutoff points to dichotomize the marker, 9 different stage groups, 4 different age groups, 17 adjusted/77 unadjusted estimates, and 2 different outcomes (OS and DFS). Type of treatment and method of marker measurement were not recorded but both would have added additional heterogeneity to that already noted. A more in-depth evaluation of these reporting problems, together with recommendations for improvement, are provided elsewhere *(15)*.

Whereas acknowledging the problems implied by this large degree of heterogeneity, it remained important to use the data extracted for each marker and determine which were potentially the most important markers for future research. Therefore, meta-analysis was performed for each of the 13 markers separately for OS and DFS. The meta-analysis results for each marker are presented in Table 3 and have been classified into three marker groups: DNA/chromosome abnormalities, biological markers, and urinary catecholamines. For the vast majority of markers, there is a statistically significant difference (in terms of the HR) between the groups defined by the markers. For example, there was strong statistically significant evidence that amplification of the *MYCN* gene was associated with a worse OS and DFS. The risk of death was 5.48 times greater for patients with *MYCN* amplification compared with those that did not have amplification [HR = 5.48; 95% confidence interval (Cl), 4.30–6.97], and similarly for risk of disease recurrence (HR = 4.28; 95% CI, 3.34–5.49). All of the papers that could be included in the meta-analyses were published from 1985 onwards, with the majority after 1989. For example, of the 70 articles used for the *MYCN* OS and DFS meta-analyses none were published before 1985, only 8 were published from 1985 to 1989, and 62 were published from 1990 onwards. Hence, the prognostic studies we included have been reported after the improved method for staging and treatment of neuroblastoma that has improved survival for children with this disease over the last 20 years.

**Monitoring.** The review identified 51 papers that provided quantitative data evaluating the serial use of tumor markers to aid the clinical management of patients with neuroblastoma. However, there was considerable heterogeneity between the studies in terms of: *(a)* tumor markers considered (e.g., VMA, HVA, *MYCN*, ferritin, NSE, and lactate dehydrogenase); *(b)* outcome (OS and DFS); *(c)* statistical analyses undertaken
Table 3 Meta-analysis results for the 13 prognostic markers, grouped by tumor marker class, for overall survival and disease-free survival together with the number of prognosis papers identified overall and the number of estimates of the log (hazard ratio) and variance obtained for each outcome

<table>
<thead>
<tr>
<th>Marker type</th>
<th>Tumor marker (relationship with prognosis)</th>
<th>No. prognosis papers</th>
<th>No. estimates obtained</th>
<th>Pooled hazard ratio</th>
<th>95% Confidence interval</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA or chromosome abnormalities</td>
<td>MYCN (amplification poor outcome)</td>
<td>151</td>
<td>OS 48</td>
<td>4.28</td>
<td>3.34 to 5.49</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>DNA index (diploidy poor outcome)</td>
<td>44</td>
<td>OS 11</td>
<td>3.12</td>
<td>1.95 to 4.98</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>Chromosome lp (deletion poor outcome)</td>
<td>40</td>
<td>OS 11</td>
<td>3.93</td>
<td>2.31 to 6.68</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Urinary catecholamines</td>
<td>VMA (elevated poor outcome)</td>
<td>36</td>
<td>OS 3</td>
<td>Not possible</td>
<td>Not possible</td>
<td>Not possible</td>
</tr>
<tr>
<td></td>
<td>HVA (elevated poor outcome)</td>
<td>26</td>
<td>OS 2</td>
<td>1.14</td>
<td>0.65 to 1.98</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>VMA:HVA (small ratio e.g. &lt;1) poor outcome</td>
<td>20</td>
<td>OS 2</td>
<td>0.44</td>
<td>0.18 to 1.06</td>
<td>0.068</td>
</tr>
<tr>
<td>Biological markers</td>
<td>CD44 (high expression good outcome)</td>
<td>10</td>
<td>OS 1</td>
<td>Not possible</td>
<td>Not possible</td>
<td>Not possible</td>
</tr>
<tr>
<td></td>
<td>TrkA (high expression good outcome)</td>
<td>8</td>
<td>OS 0</td>
<td>0.06</td>
<td>0.02 to 0.21</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>NSE (high serum levels poor outcome)</td>
<td>16</td>
<td>OS 4</td>
<td>Not possible</td>
<td>Not possible</td>
<td>Not possible</td>
</tr>
<tr>
<td></td>
<td>LDH (high serum levels poor outcome)</td>
<td>28</td>
<td>OS 4</td>
<td>Not possible</td>
<td>Not possible</td>
<td>Not possible</td>
</tr>
<tr>
<td></td>
<td>Ferritin (high serum levels poor outcome)</td>
<td>26</td>
<td>OS 5</td>
<td>3.20</td>
<td>2.06 to 4.98</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>MRP (high expression poor outcome)</td>
<td>33</td>
<td>OS 3</td>
<td>Not possible</td>
<td>Not possible</td>
<td>Not possible</td>
</tr>
<tr>
<td></td>
<td></td>
<td>16</td>
<td>OS 9</td>
<td>Not possible</td>
<td>Not possible</td>
<td>Not possible</td>
</tr>
</tbody>
</table>

*DFS, disease-free survival; OS, overall survival.

and and reporting of results; (d) length of follow-up, e.g., treatment phase, long-term follow-up; (e) number of patients; (f) age/stage distribution; and (g) number of serial measurements.

The combination of these problems with the relatively few papers identified meant that any meta-analysis using these studies was not worthwhile, statistically or clinically.

Publication Bias. The estimates of the log (HR) and its variance obtained for MYCN were used to assess the possibility of publication bias. Both Begg and Egger tests produced Ps < 0.001, and therefore publication bias was strongly suspected to be a problem (12, 13). In fact, using the Trim and Fill method, 17 studies with smaller HRs were estimated as missing from our results, in addition to the 45 studies included in the analysis. Hence, it appears likely that the effect size from the original meta-analyses may be biased upwards, i.e., they may overestimate the true underlying log, (HR) for MYCN. It is highly likely that this problem is also true for the other markers.

Economic and Psychosocial Results. No papers made an economic evaluation of the use of tumor markers in neuroblastoma, but 2 papers reported cost data in relation to screening (16, 17). They are both somewhat dated and contain few details about cost calculations, which made it difficult to assess the accuracy of the claims made or the relevance of the findings to current practice. Furthermore, none of the 428 relevant papers reported any data on the psychosocial consequences for children and their families of using tumor markers clinically in neuroblastoma, even in the monitoring papers.

Discussion

This is the first systematic review of tumor markers for neuroblastoma that has been reported, and it forms a knowledge base for future research. A systematic review is the preferred means of identifying and combining existing evidence (18), and it is particularly important for evidence-based evaluations of tumor markers in neuroblastoma, and indeed other rare diseases, because (sometimes conflicting) evidence relating to markers is published across a number of studies, many of which involve small numbers of patients. Systematic reviews are also important because they can highlight underlying problems across individual studies and help identify future research needs (19). Indeed, both of these aspects are demonstrated throughout our review and form the most important messages of this paper, which we hope will help improve tumor marker studies in the future.

Appraisal of the Systematic Review. During the systematic review we classified 3415 papers overall and ultimately identified 428 relevant papers, which showed diversity in primary interest, methodology, analysis of data, and quality of reporting. The search strategy used is likely to have identified the majority of the available literature, targeting in particular the
databases specializing in scientific and clinical reporting, although we acknowledge the possibility that the review may not be fully comprehensive, reflecting publication and reporting bias. Our initial search strategy included the names of those markers known *a priori* to be potentially important but, in light of the many markers identified during the review, this list was certainly not exhaustive, although we did include more general terms such as "marker" that will have limited the number of markers missed.

Initially our strategy had been to check the references quoted in all of the selected papers, to identify any other papers missing from the three databases we searched. This was not feasible given the large literature base, nor was it possible to systematically check for duplicates of patients across papers. We did not include foreign language papers because of the difficulties in translation, and this may have introduced bias if statistically or clinically significant studies were more likely to be (re)written for publication in an English language journal (20).

**Clinical Interpretation.** Neuroblastoma is a multifaceted disease. The expansion in biological and cytogenetic markers is an indication that the cancer process is complex with multiple changes taking place with a neuroblastoma cell. Research studies have tried to identify which of these markers are initiating factors contributing to the cancer phenotype and which may be regarded as secondary. Initial studies for a particular marker have inevitably sought to link that one marker with survival, to establish its importance. However, it is now becoming clear that a number of key biological markers are likely to be interlinked and must be evaluated in combination with one another and not in isolation.

The poor and heterogeneous reporting restricted any quantitative synthesis in the areas of screening, diagnosis, and monitoring, and therefore any overall clinical evaluation. Recent papers, published since the start of our review, suggest that there is currently insufficient evidence to support a screening program for infants up to 6 months of age, and the majority of authors conclude that it should be discontinued (21). Clinical and histopathological features, which have not been evaluated in our review, are the most informative for the diagnosis of neuroblastoma, and these include age at diagnosis, tumor histology, and primary tumor site. However, the detection of catecholamine metabolites in urine is also used for the differential diagnosis of neuroblastoma from other small round cell tumors of childhood. For the use of molecular and biological markers in diagnosis, the few studies comparing a healthy control group to patients with neuroblastoma was particularly disappointing, and future studies need to address this. Similarly, monitoring studies need to report the differences in serial marker measurements between those who develop a recurrence of disease and those who remain disease-free, preferably for a many patients over a long follow-up period. Where possible, research groups need to collaborate and pool together resources to enable bigger sample sizes and achieve consistency across studies, which should be targeted to address the important issues. Only then will the benefits of using tumor markers for screening, diagnosis, and monitoring be properly ascertainable.

This systematic review did produce an evaluation of the most commonly reported individual markers for prognosis; *MYCN*, chromosome 1p, DNA index, VMA:HVA ratio, CD44, Trk-A, NSE, lactate dehydrogenase, ferritin, and multidrug resistance were all identified as potentially important prognostic tools. However, the pooled results must be treated with caution given the reporting problems and large heterogeneity of clinical/statistical factors across studies. Recent studies have also indicated that chromosome 17q gains also have important prognostic significance (22–28). Unfortunately, many of these studies were published after the start of our review, and consequently chromosome 17q was not among the prognostic markers we selected for the in-depth evaluation above. However, in light of this current knowledge, we have subsequently extracted, wherever possible, HR results from each of the 8 prognosis papers in the review for this marker. Meta-analysis of these suggests that patients with gain of chromosome 17q have a significantly worse DFS (from 3 studies: HR = 4.16; 95% CI, 2.56–6.77) and OS (from 3 studies: HR = 4.30; 95% CI, 2.70–6.86) compared with those who did not. However, these results are again subject to the problems of poor reporting and heterogeneity.

It was not possible to compare subgroups of patients, individual prognostic markers, or assess the benefits of using any of the markers in combination. Availability of full IPD, including all of the exact values of all markers assessed and subgroup information (e.g., age and stage of disease), from each paper would facilitate such assessments in the future (15). Ideally, large multicenter studies should also be initiated to assess the benefits of using chromosome 17q and the other important prognostic markers, both individually and in combination, to improve strategies for the stratification of children with neuroblastoma for therapy. Results from such studies would enable clinicians to clearly see which are the most appropriate individual and combinations of prognostic markers to use.

**Reporting Problems.** Importantly, our review has highlighted problems in how individual prognostic marker studies are designed and reported. Many study reports do not allow for adequate extraction of data in order for comparisons to be made, and the large heterogeneity in cutoff levels and other measures limit evaluations of markers across studies and within specific subgroups of patients (e.g., age <1). There is also the threat of publication bias across the literature. For those researchers studying prognostic tumor markers, Altman and Lyman (29) have proposed important guidelines for both conducting and evaluating prognostic factor studies that should be considered. We have also provided guidelines for improved statistical reporting of primary studies to facilitate the evaluation and comparison of individual markers, identify the additional benefits of using combinations of markers, and ultimately allow important evidence-based reviews to be made (15). In particular, presentation of the HR with some measure of precision (e.g., 95% CI) and availability of IPD, either in the paper or on the internet (30), are both highly desirable. The availability of IPD has allowed important evidence-based reviews to be made in other cancer settings (31, 32).

**Psychosocial and Economic Issues.** The clinical implications of the results must also be considered together with psychosocial and economic aspects of tumor markers, for example in monitoring and screening, respectively. Our search
found no published evidence on the psychosocial consequences for children and their families of using tumor markers clinically in neuroblastoma, but this probably reflects the few papers studying the use of tumor markers for monitoring patients. Psychosocial evaluations of tumor markers are clearly important and have been performed for other disease settings (33). Future research on the use of tumor markers in pediatric oncology should seek to include an assessment of the psychosocial outcomes of using markers, particularly for prognostic and monitoring purposes.

We also only identified two studies that included an economic evaluation of using tumor markers, and so the cost-effectiveness of individual markers could not be evaluated. A clear prescription for future trials and studies is therefore to include an economic evaluation element wherever possible, either in terms of a cost-effectiveness study or the identification of resource use that would permit decision models to be developed.

Summary. This systematic review has emphasized the uncertainty on the clinical value of the studied tumor markers in neuroblastoma, reflecting the small size of many studies and poor statistical reporting. We have managed to assess the significance of the most commonly reported prognostic markers: MYCN, chromosome 1p, DNA index, VMA/HVA ratio, CD44, Trk-A. NSE, lactate dehydrogenase, ferritin, and multidrug resistance were all identified as potentially important prognostic tools. Each of these, and the more recently identified changes in chromosome 17q, should be considered in the development of future research strategies and for the stratification of children with neuroblastoma for different treatment strategies.

The multiplicity and complexity of tumor markers in neuroblastoma underlines the need for studies to be coordinated, through the cooperation of cancer research groups, using multiple laboratories and standardizing methods of analysis and reporting. In particular, collaboration is needed to consider prospectively planned pooled analyses and facilitate the pooling of IPD, a strategy that would adequately ensure a quantitative synthesis (rather than merely a qualitative synthesis) to address the questions of interest, such as which combinations of markers provide the best prognostic tool, or whether monitoring patients with neuroblastoma using markers is cost-effective. Large, multicenter collaborative studies would require agreement as to which markers to measure, and we have provided a base that highlights the ones reported in greatest detail thus far.

Acknowledgments
We thank Suzy Paisley at the School of Health and Related Research (SCHARR) in Sheffield for advice on systematic reviews and literature searching.

References


Appendix A4: Publication in the British Journal of Cancer arising from the reporting guidelines developed for primary prognostic marker studies in Section 2.9

Riley RD, Abrams KR, Sutton AJ, Lambert PC, Jones DR, Heney D, Burchill SA:

RR oversaw the entire review. RR developed the search strategy with advice from clinicians SB and DH. RR performed all the literature searching and classification, with samples checked by DH and SB. RR performed at least 90% of the data extraction and indirect estimation of the hazard ratios with KA, AS, PL and DJ performing the remainder. RR developed the reporting guidelines, with slight amendments and additional suggestions made by the other authors. SB and DH provided clinical advice throughout.
RR wrote the paper, with comments and suggestions provided on various draft versions by all the other authors.
SPECIAL NOTE

ITEM SCANNED AS SUPPLIED
PAGINATION IS AS SEEN
Prognostic markers help to stratify patients for treatment by identifying patients with different risks of outcome (e.g., recurrence of disease), and are important tools in the management of cancer and many other diseases. Systematic review and meta-analytical approaches to identifying the most valuable prognostic markers are needed because (sometimes conflicting) evidence relating to markers is often published across a number of studies. To investigate the practicality of this approach, an empirical investigation of a systematic review of tumour markers for neuroblastoma was performed. 260 studies of prognostic markers were identified, which considered 130 different markers.

The reporting of these studies was often inadequate, in terms of both statistical analysis and presentation, and there was considerable heterogeneity for many important clinical/statistical factors. These problems restricted both the extraction of data and the meta-analysis of results from the primary studies, limiting feasibility of the evidence-based approach.

Guidelines for reporting the results of primary prognostic marker studies in cancer, and other diseases, are given in order to facilitate both the interpretation of individual studies and the undertaking of systematic reviews, meta-analysis and, ultimately, evidence-based practice. General availability of full individual patient data is a necessary step forward and would overcome the majority of problems encountered, including poorly reported summary statistics and variability in cutoff level, outcome assessed and adjustment factors used. It would also limit the problem of reporting bias, although publication bias will remain a concern until studies are prospectively registered. Such changes in practice would help important evidence-based reviews to be conducted in order to establish the most appropriate prognostic markers for clinical use, which should ultimately improve patient care.

In this paper, we use a recently performed systematic review of tumour markers studied in neuroblastoma to demonstrate the problems encountered when using this approach, and highlight how they limit evidence-based practice. We then generalise the problems to other areas of oncology, and indeed other disease settings, and ultimately provide specific guidelines for reporting primary prognostic marker studies.

**METHODS**

Neuroblastoma is a neuroblastic tumour of the primordial neural crest and is the most common extracranial solid tumour of childhood. The study of prognostic markers for this disease forms an active research area within which a large body of evidence exists. This makes it an appropriate area for an empirical
investigation, and as such the problems identified in this study are likely to generalise to other disease settings. A brief description of the systematic review strategy adopted is now given.

Search strategy
The three on-line bibliographic databases Medline, Embase and Cancerlit were chosen as a basis for identifying the relevant literature from 1966 to February 2000. Papers written in a non-English language were excluded. A full description of the search strategy and inclusion/exclusion criteria is provided in Riley et al (2003b). One investigator performed the assessment of the papers, with second and third investigators independently checking a sample of them. To be included in the review, a paper had to provide a quantitative result or give tabulated individual patient data evaluating the use of a tumour marker in neuroblastoma from a primary research study of humans. To be classified as relevant to prognosis, a paper had to present data, in the form of summary statistics or individual patient data, relating tumour marker levels at a measured point in time to the outcome of patients at the end of a specific follow-up period. Owing to the large number of potential markers, we focused on genetic/biological markers rather than histological markers.

Data extraction for meta-analysis
From each of the papers included, information was extracted on the tumour markers studied. Meta-analysis of those markers on which eight or more papers provided data was considered. The loge(hazard ratio) and its variance were the essential information required from each study as they provide an important comparative estimate of the risk of death/disease recurrence between two groups of patients. Furthermore, there are several indirect estimation methods available when these statistics are not directly reported (Parmar et al, 1998), and the loge(hazard ratio) has an approximate normal distribution for large samples, making it particularly amenable to meta-analysis techniques. We make the assumption of proportional hazards throughout this paper.

It was common for a paper to report more than one prognostic result by relating one or more markers to overall survival and/or disease-free survival, and also by providing unadjusted and/or adjusted results (e.g. adjusted for age, stage of disease). Estimates of the loge(hazard ratio) and its variance comparing two groups of patients. Furthermore, there are several indirect estimation methods available when these statistics are not directly reported (Parmar et al, 1998), and the loge(hazard ratio) has an approximate normal distribution for large samples, making it particularly amenable to meta-analysis techniques. We make the assumption of proportional hazards throughout this paper.

Methods and results at each stage of the sequential process used to obtain a single direct or indirect estimate of the loge(hazard ratio) and its variance for each of the reports where one of the 13 tumour markers was related to overall survival or disease-free survival by summary statistics or individual patient data across the literature. Five steps were used, with unadjusted estimates sought primarily in each unless only an adjusted result was available or otherwise stated.

**Figure 1** Methods and results at each stage of the sequential process used to obtain a single direct or indirect estimate of the loge(hazard ratio) and its variance for each of the reports where one of the 13 tumour markers was related to overall survival or disease-free survival by summary statistics or individual patient data across the literature. Five steps were used, with unadjusted estimates sought primarily in each unless only an adjusted result was available or otherwise stated.
Table 1 Description of the methods used to obtain estimates of the log$_{10}$ (hazard ratio) and its variance

<table>
<thead>
<tr>
<th>Method</th>
<th>Summary statistics or data required</th>
<th>Step 1</th>
<th>Step 2</th>
<th>Step 3</th>
<th>Step 4</th>
<th>Step 5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>HR or log$_{10}$(HR) and V</td>
<td>—</td>
<td>3</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>Individual patient data</td>
<td>—</td>
<td>41</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>41</td>
</tr>
<tr>
<td>3</td>
<td>log$_{10}$(HR) and CI</td>
<td>0</td>
<td>—</td>
<td>—</td>
<td>1</td>
<td>—</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>HR and CI</td>
<td>30</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>30</td>
</tr>
<tr>
<td>5</td>
<td>log$_{10}$(HR) and P-value</td>
<td>—</td>
<td>2</td>
<td>0</td>
<td>—</td>
<td>—</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>HR and P-value</td>
<td>10</td>
<td>2</td>
<td>2</td>
<td>—</td>
<td>—</td>
<td>14</td>
</tr>
<tr>
<td>7</td>
<td>HR, group numbers and total events</td>
<td>2</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>Z$^2$ statistic, group numbers and total events</td>
<td>10</td>
<td>—</td>
<td>4</td>
<td>0</td>
<td>—</td>
<td>14</td>
</tr>
<tr>
<td>9</td>
<td>P-value, group numbers and total events</td>
<td>67</td>
<td>2</td>
<td>—</td>
<td>0</td>
<td>—</td>
<td>88</td>
</tr>
<tr>
<td>10</td>
<td>Survival curve</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>124</td>
<td>41</td>
<td>27</td>
<td>8</td>
<td>4</td>
<td>204</td>
</tr>
</tbody>
</table>

The summary statistics required for each method are shown together with the number of times each method was successfully used in Steps 1–5 of the extraction process. Methods 3–10 were used in order of preference shown. HR=hazard ratio; V=variance of the log$_{10}$(HR); CI=confidence interval.

Table 2 Names of the 13 markers grouped by tumour marker class, with the number of prognosis papers identified for each, the number of reports when it was related to either overall or disease-free survival by summary statistics or individual patient data, and the number of successful estimates made of the log$_{10}$(hazard ratio) and variance; evidence of heterogeneity is shown for outcome, cutoff levels, age, stage and adjustment factors

<table>
<thead>
<tr>
<th>Marker class</th>
<th>Marker name</th>
<th>Papers*</th>
<th>OS and DFS reports*</th>
<th>Total successful estimates (p)*</th>
<th>OS/DFS successes*</th>
<th>Different cutoff groups*</th>
<th>Different stage groups*</th>
<th>Different age groups*</th>
<th>U/A*</th>
<th>Different sets of adjustment factors*</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA or chromosome abnormalities</td>
<td>MIT-N</td>
<td>151</td>
<td>194</td>
<td>94</td>
<td>48/46</td>
<td>9</td>
<td>9</td>
<td>4</td>
<td>77/17</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>DNA index</td>
<td>44</td>
<td>62</td>
<td>19</td>
<td>11/8</td>
<td>8</td>
<td>3</td>
<td>3</td>
<td>18/1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Chromosome 1p</td>
<td>40</td>
<td>49</td>
<td>20</td>
<td>11/9</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>18/2</td>
<td>2</td>
</tr>
<tr>
<td>Urinary catecholamines</td>
<td>VMA</td>
<td>36</td>
<td>40</td>
<td>4</td>
<td>3/1</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>4/0</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>HVA</td>
<td>26</td>
<td>29</td>
<td>2</td>
<td>2/0</td>
<td>2</td>
<td>1</td>
<td>2/0</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>VMAH-VHA</td>
<td>20</td>
<td>28</td>
<td>5</td>
<td>2/3</td>
<td>3</td>
<td>4</td>
<td>2/0</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Dopamine</td>
<td>10</td>
<td>11</td>
<td>2</td>
<td>1/1</td>
<td>2</td>
<td>2</td>
<td>2/0</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Biological markers</td>
<td>CD44</td>
<td>8</td>
<td>8</td>
<td>3</td>
<td>0/3</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3/0</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>TrkA</td>
<td>16</td>
<td>21</td>
<td>11</td>
<td>4/7</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>9/2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>NSE</td>
<td>28</td>
<td>39</td>
<td>9</td>
<td>4/5</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>8/1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>LDH</td>
<td>26</td>
<td>30</td>
<td>12</td>
<td>5/7</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>8/4</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Ferritin</td>
<td>33</td>
<td>41</td>
<td>7</td>
<td>3/4</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>6/1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>MDR</td>
<td>16</td>
<td>30</td>
<td>16</td>
<td>9/7</td>
<td>8</td>
<td>3</td>
<td>3</td>
<td>13/3</td>
<td>2</td>
</tr>
</tbody>
</table>

Problems limiting meta-analysis

Problems limiting meta-analysis

**Poor reporting of primary studies**

Primary studies of prognostic tumour markers are clearly essential and we observed many important results across the literature that have implications for clinical practice. However, the general standard of reporting primary studies was inadequate, and it was disappointing that we only managed to obtain 35.5% of the estimates required despite the intensive, time-consuming extraction procedure (Figure 1). This hindered the use and interpretation of meta-analysis because we could not incorporate the majority of results reported in the literature and consequently introduced a strong potential for bias. Among the 371 reports that did not enable estimates to be made, there were five common reporting problems, most of which can be simply addressed (Figure 2). Encouragingly, there was some evidence that the reporting of prognostic markers has improved over the last 10 years because all the papers that did provide a hazard ratio or log$_{10}$(hazard ratio) were published after 1990. However, these papers still only represented approximately 17% of the total literature identified over this period (i.e. only 26 out of 157 papers published after 1990 reported a hazard ratio).
Clinical markers and evidence-based practice
RD Riley et al

Key problems in the reporting of prognostic marker studies

We studied 13 tumour markers in 211 papers, and identified 575 reports (including summary statistics or IPD) that assessed their prognostic value. On trying to extract the log(hazard ratio) and variance from these reports, we found five main problems:

1. No appropriate statistical analysis performed or reported.
   In 133 (23.1%) reports a paper reported prognostic data (e.g. the number of patients who had an event in each group) but no results from a Cox regression analysis or log-rank/Wilcoxon test, often because no such analyses had been performed. Hence, Methods 1–10 could not be used.

2. Hazard ratio not calculated or not reported.
   In only 57 of the 575 reports (9.9%) were direct estimates of the hazard ratio or log(hazard ratio) provided. For these, a variance (or a standard error) was given three times and a confidence interval 34 times. In the other 20, at most a P-value was given.

3. Inexact P-values provided.
   Overall, 273 of the 575 reports presented a P-value from a Cox regression analysis or log-rank/Wilcoxon test. In 126 of these the P-value was stated as P<0.05 or 'Psignificant', and in 13 reports the P-value was stated as P>0.05 or 'Pnot significant'. This again shows inappropriate emphasis placed on the P-value for a statistically significant result.

4. Group numbers and group events not given.
   There were 210 reports where a P-value/log(ratio) statistic from a Cox regression analysis or log-rank/Wilcoxon test was presented but without a hazard ratio or log(hazard ratio). From the 194 of these that had a sample size >25, only 104 indirect estimates were obtained because the group numbers and/or group events were not reported and could not be estimated from figures or tables. The number of patients and events in groups defined by marker levels are often smaller than the overall numbers because of missing or incomplete patient data. Hence, it is important to report numbers for the groups themselves.

5. Marker studies too small.
   In Steps 2–5, estimation methods were only considered appropriate if the sample size was greater than 25; as it was in only 196 of 318 reports otherwise suitable for these steps.

When necessary and possible, research groups need to collaborate to achieve larger sample sizes and thus increase statistical power.

Figure 2 Description of the key reporting problems that prevented estimation of the log(hazard ratio) and its variance in 371 (64.5%) of the reports.

Heterogeneity of clinical and statistical factors

The synthesis of our estimates was also restricted by the large variability in both clinical and statistical factors. For each estimate of the log(hazard ratio) and its variance obtained, the cutoff level used to dichotomise the continuous markers, stage of disease, age of patients and outcome (overall or disease-free survival) were recorded, and also whether the estimate was unadjusted or adjusted and, if so, what adjustment factors were used. There was great diversity in these features (Table 2). For example, for the marker MYC-N there were 94 estimates of the log(hazard ratio) and variance obtained but these involved nine different cutoff points, nine different stage groups, four different age groups, 77 unadjusted/17 adjusted estimates and two different outcomes (Table 3). Furthermore, of the 17 estimates that were adjusted for other prognostic markers or clinical features (using a Cox regression model) only two were adjusted for exactly the same set of factors, and these were from the same article (Maris et al, 2000).

This inconsistent and variable reporting was reflected equally in the estimates obtained for the other 12 markers (Table 2). The type of treatment of patients and the method of measuring the markers were not recorded, but both would have added further heterogeneity to that observed.

Publication bias and reporting bias

The common problem of publication bias, and other reporting biases, may still affect our data extraction; some results that do not generate formal statistically significant or clinically valuable findings may not have been published, because of a reluctance of journals to report or of researchers to present negative findings. Such problems severely limit the conclusions that can be drawn from meta-analyses because not all the available evidence can be included, and therefore the pooled results are likely to be biased. We investigated the estimates obtained for MYC-N and indeed there did appear to be evidence of publication bias, with a number of studies with smaller hazard ratios considered to be missing (Riley et al, 2003). This problem is likely to be closely related to the

Table 3 Heterogeneity in the 94 estimates of the log(hazard ratio) and its variance obtained for marker MYC-N

<table>
<thead>
<tr>
<th>Outcome</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>DFS</td>
<td>46</td>
</tr>
<tr>
<td>OS</td>
<td>48</td>
</tr>
<tr>
<td>Result type</td>
<td></td>
</tr>
<tr>
<td>Unadjusted</td>
<td>77</td>
</tr>
<tr>
<td>Adjusted</td>
<td>17</td>
</tr>
<tr>
<td>Stage groups</td>
<td></td>
</tr>
<tr>
<td>No stage</td>
<td>68</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>1, 2, 3</td>
<td>1</td>
</tr>
<tr>
<td>1, 2, 3, 4</td>
<td>5</td>
</tr>
<tr>
<td>2, 3, 4, 5</td>
<td>2</td>
</tr>
<tr>
<td>3, 4</td>
<td>1</td>
</tr>
<tr>
<td>Unknown</td>
<td>5</td>
</tr>
</tbody>
</table>

OS=overall survival, DFS=disease-free survival
Should we proceed with meta-analysis?

The poor reporting, potential for publication bias and, in particular, the large heterogeneity across studies meant it was practically impossible to perform reliable meta-analyses that would determine the clinical importance of each marker studied. Even the analysis of subgroups of estimates was not considered realistic because it was virtually impossible to obtain subgroups that reflected patients with similar features. For example, for marker MYC-N there were 48 overall survival estimates obtained, of which 41 were unadjusted, and 30 related to 'all ages' and 'all stages'. Furthermore, only eight of these 30 estimates related to the most commonly used cutoff level of '1 copy number', and there is then the additional problem of heterogeneity for treatment used and method of measuring the marker, not to mention the potential impact of publication/reporting bias. The subgroup numbers were even smaller for the other, less-studied markers; for example, lactate dehydrogenase had only two unadjusted overall survival estimates relating to the most common cutoff level (1500 U/l⁻¹) and patients of 'all ages' and 'all stages'.

The only possible benefit of meta-analysis using the estimates that we extracted is to highlight the results of previous studies and help prioritise which markers should be studied in the future. We take such an approach elsewhere (Riley et al., 2003b), but for the purposes of this feasibility study it is clear that no firm clinical policy decisions can be made from our evidence-based review.

DISCUSSION

Appraisal of the systematic review and data extraction

During the systematic review, we evaluated 3415 papers overall and identified 260 with results from studies assessing the prognostic power of tumour markers. This will have identified the majority of the English-language literature, but inevitably some papers will have been excluded unintentionally. However, it seems plausible that the reporting in such papers, and equally non-English papers, would be equally poor and heterogeneous.

We used the indirect methods suggested by Parmar et al. (1998) to increase the number of occasions an estimate of the log(hazard ratio) and its variance could be obtained. However, the estimates they provide are only approximate and simply make the best possible use of the results presented. Questions still exist about how best to combine indirect estimates with direct estimates. For this reason, we did not use other indirect methods. For example, given further assumptions, we could have used estimates of the proportion surviving to 2, 3, 5 or 10 years to obtain estimates of log(hazard ratio) and its variance (Vale et al., 2002). However, the papers were equally inadequate at presenting these survival statistics. For example, in the 26 prognostics papers for the serum marker lactate dehydrogenase, only 12 gave actuarial estimates of the proportion surviving, and only six of these also gave a confidence interval or standard error. They were also heterogeneous—five estimates were for overall survival, six were for disease-free survival and one was unspecified; estimates were made at 2, 3, 4 or 5 years. Further, very few reported numbers at risk explicitly, as required for reliable estimation.

Generalisations to other prognostic markers

Although these reporting problems were observed for tumour markers within the neuroblastoma literature, they have also limited reviews in other paediatric cancers (Riley et al., 2003a), and it seems plausible that the reporting will be equally poor for prognostic markers in other areas of oncology, and indeed other disease areas. Altman (2001, pp 228–247) discusses the potential problems involved in systematic reviews of prognostic markers, in particular that of poor and heterogeneous reporting of primary survival. Cutoff points are frequently used to dichotomise continuous markers and define groups, while different outcomes, adjustment factors and groups of patients are common features across prognostic studies. Inadequate reporting and presentation of survival data has been shown to be a concern in the cancer literature (Altman et al., 1995).

Reliable and clinically useful meta-analyses of observational and nonrandomised studies, such as the majority of prognostic marker studies, are generally difficult to perform (Fleiss and Gross, 1991). Other recent systematic reviews of prognostic markers have encountered similar problems to the ones we identified. Parker et al. (2001) performed a systematic review in prostate cancer to establish whether age is a prognostic marker, but the incomplete and heterogeneous nature of the reports prohibited any quantitative overview. Similarly, a systematic review of prognostic laboratory variables in patients with unresected colorectal liver metastases was limited by the heterogeneity and poor quality of individual studies (Friedburg et al., 2001). Zandbergen et al. (2001) performed a systematic review of biochemical markers of brain damage for identifying poor outcome in anoxic-ischaemic coma, but conclusions were limited by small sample sizes and different cutoffs and/or laboratory techniques.

Meta-analyses of prognostic markers have been facilitated when individual patient data were available (Look et al., 2002), in particular to determine a consistent cutoff level (Sakamoto et al., 1996). For those investigators currently interested in performing a quantitative review of prognostic markers, we recommend that they consider asking authors for individual patient data and/or the extra information they require, such as the log(hazard ratio) and its variance, as this approach is likely to be the most productive.

Towards guidelines for improved reporting of prognostic markers

It is clearly important that the quality of primary studies, and the reporting of their results improve if clear conclusions and policy recommendations are to be formed about prognostic markers. Altman and Lyman (1998) have proposed important guidelines for both conducting and evaluating prognostic marker studies, including the need for prospective registration of studies. Alongside these, we have developed simple guidelines on how to report results to facilitate both interpretation of individual studies and the undertaking of systematic reviews, meta-analysis and, ultimately, evidence-based practice (Figure 3). Collaboration of research groups is required to promote such practice and achieve both the consistency and standards required. Ideally, both summary data and individual patient data should be reported according to our guidelines. It is important that time to event is incorporated within prognostic marker analyses, and thus the hazard ratio is preferred to other measures of relative risk such as the odds ratio, which relates to a fixed time-point and ignores censoring. However, in addition authors may wish to present the more familiar actuarial % survival at n years preferably with a confidence interval and the number of patients at risk at that time in each group.

Benefits of individual patient data  Although improved reporting of summary statistics is very important, the availability of individual patient data is the most viable way forward in order to produce valid and clinically useful evidence-based reviews of prognostic markers. Subject to any restrictions imposed by data protection laws and guidelines, presentation or availability of full individual patient data using our guidelines would overcome...
Guidelines for reporting prognostic marker studies

Objective: To improve reporting of prognostic marker results and facilitate access to individual patient data for evidence-based reviews.

Results of all the marker analyses should be presented—both significant and nonsignificant results—and we recommend the following:

Essential to present:
1. The hazard ratio and its confidence interval, or the log(hazard ratio) and its variance. Markers that have a continuous function should be modelled as a continuous variable using appropriate methods. If there is a justifiable reason for using a cutoff level for a continuous marker, it should be specified at the start of the study and clearly reported.
2. The number of patients and number of events in total. For binary markers (and continuous markers if a cutoff level is used) also report the numbers within each group.
3. Both unadjusted and adjusted results for each marker. For adjusted results, clearly state what variables have been adjusted for. Ideally, a consistency in the set of adjustment factors used across studies should be sought through collaborative groups working towards prospectively planned pooled analyses.
4. Individual patient data in the paper or on the Internet, or make available with details clearly indicated within the paper. Data on markers that were not analysed should be included. Subject to any restrictions imposed by data protection laws and guidelines, include:
   - Exact initial marker level and how marker was measured.
   - Time of disease recurrence (if applicable).
   - Follow-up time.
   - Final disease status.
   - Levels of other existing prognostic markers of recognised and accepted importance for current clinical practice.
   - Patient subgroup information, for example, age, stage of disease, type of treatment received.
   - Details of inclusion/exclusion criteria would also be beneficial.

Highly desirable to present:
5. Exact P-values. Reporting of results as 'significant' or 'not significant' is insufficient. Very small P-values can be given as P < 0.0001, but in this case the exact χ²-statistic is also needed.
6. Survival curves showing the difference in survival over time between the groups, with clear step and censoring points; also the initial numbers in each group, and the number of events and remaining numbers at various time-points during follow-up are needed.
7. % survival for patients with a confidence interval using Kaplan–Meier or other methods that allow for censoring, together with the number of patients at risk at that time in each group.

Figure 3 Guidelines on how to report primary prognostic marker studies in order to improve current reporting standards and allow clinically useful evidence-based reviews to be made.

Variability in cutoff level, type of estimate (unadjusted or adjusted), outcome assessed (overall or disease-free survival) and adjustment factors; the study of markers in subgroups of patients (e.g. different ages, treatments) would also be easier. It would also eliminate the problem of extracting estimates when inexact P-values are presented, and would remove the need for arbitrary extraction decisions when an individual study presents a marker’s results for a range of cutoff values. Furthermore, if levels of all the prognostic markers measured (even those producing nonsignificant results) are provided, then the problem of reporting bias would be reduced. However, publication bias might still be a concern if some studies are not published and do not make IPD available; prospective registration of studies is therefore also important to counteract this.

Individual patient data would also enable direct estimates of the hazard ratio, and other statistics of interest, when data were available but not used, analysed or presented properly in the primary study. A total of 41 (20%) of the 204 estimates that we obtained in the neuroblastoma review were direct estimates calculated from individual patient data that would not have otherwise been possible. It is clearly important to include predominately direct estimates in any quantitative synthesis. In fact, the potential for substantial differences in meta-analysis of survival data when using results provided within the literature instead of individual patient data has recently been shown in the head and neck cancer literature (Duchateau et al, 2001). Individual patient data would also allow us to assume, for example, proportional hazards, to be checked as necessary, and enable the baseline survival function to be estimated.

Presentation or availability of individual patient data would permit more appropriate meta-analyses (Stewart and Parmar, 1993), and would further facilitate the identification of different publications whose results relate to the same or overlapping set of patients. It would also allow an evaluation of combinations of markers, which may produce more specific and accurate prognostic assessments. If it is not appropriate or feasible to provide individual patient data within a paper itself, then there is the opportunity to publish on the Internet (Hutchon, 2001). Of course, even making individual patient data available on the web is not without its problems, with the nonpermanency of individual web-pages, and so perhaps a central repository to collate and manage individual patient data is needed within each disease area. The United Kingdom Children’s Cancer Study Group have already initiated this type of approach within paediatric oncology (Mott et al, 1997). Authors may also wish to state in their paper that the IPD is available upon request (with contact details indicated) for those requiring it for evidence-based reviews.

We acknowledge that there are additional issues that arise when conducting individual patient data reviews (Stewart and Clarke, 1997), especially cost and time, but these have to be weighed against the substantial problems we encountered. Of course, even when prioritising the IPD approach, the meta-analyst will in practice end up with a mixture of estimates obtained from IPD and estimates obtained from summary statistics; hence, meta-analysis methods that take these different sources into account are needed.

Cutoff levels

The use of different cutoffs makes synthesis of results particularly difficult. Of added concern is the possibility that the choice of cutoff level in a report may be specifically chosen to optimise the difference between the groups and produce a result with the maximum statistical or clinical significance possible (Altman et al, 1994; Altman and Lyman, 1998). If there is good clinical reason to use a cutoff level, then it should be specified at the start of a study and clearly reported within the results (Figure 3). However, Altman (2001) suggests that continuous markers should not be dichotomised because, among other reasons, this approach discards potentially important quantitative information and considerably reduces the power to detect a real association between the marker and outcome. Hence, we encourage researchers to analyse and report results (e.g. hazard ratio) of continuous markers on their original continuous scale. Importantly, availability of individual patient data including exact marker levels would allow data to be reanalysed where cutoff levels were not consistent, and also where continuous marker results were desired but results using a cutoff level were given (or vice versa) (Figure 3). Indeed, the most appropriate analysis of continuous prognostic markers may require nonlinear modelling techniques, as highlighted by Sauerbrei et al (1999); consultancy with statisticians or others experienced with such techniques is recommended in this situation.

Adjustment factors

It is clear that once important prognostic markers have been identified, they need to be evaluated against, and also used in combination with, other known clinically useful prognostic factors, such as clinical characteristics (e.g. age, stage of
Prognostic markers and evidence-based practice

RD Riley et al

ACKNOWLEDGEMENTS

We wish to thank Doug Altman for his helpful comments and advice during the drafting of this paper. This work was funded by the NHS Health Technology Assessment (HTA) programme [Project no. 97/15/03].


Stewart LA, Parmar MKB (1993) Meta-analysis of the literature or of individual patient data: is there a difference? The Lancet 341: 418–422
APPENDIX B

Appendix B1: SAS PROC MIXED code to fit (i) Model A and (ii) Model A-reg

(i) To fit Model A I used SAS Proc MIXED, based on the code suggested by Van Houwelingen et al [148]. This example is for the original Berkey dataset introduced in Table 4.1 [151]:

```sas
# input the Berkey data (i.e. id no., estimate, variance of estimate, probing depth identifier
# attachment level identifier, constant, s.e. of estimate, observation no., year of publication)
data berkey;
input id diff var al pd cons se obsn year;
cards;
1 0.47 0.0075 0 1 0.08660254 1 0
1 -0.32 0.0077 1 0 0.087749644 2 0
2 0.2 0.0057 0 1 0.075498344 3 -1
2 -0.6 0.0008 1 0 0.028284271 4 -1
3 0.4 0.0021 0 1 0.045825757 5 -3
3 -0.12 0.0014 1 0 0.037416574 6 -3
4 0.26 0.0029 0 1 0.053851648 7 4
4 -0.31 0.0015 1 0 0.038729833 8 4
5 0.56 0.0148 0 1 0.121655251 9 5
5 -0.39 0.0304 1 0 0.174355958 10 5
;
run;
```

# estimate the parameters in Model A using restricted maximum likelihood (which is equivalent to
# restrictive generalised least squares)
proc mixed cl method=reml data=berkey;

# specify the categorical variables
class id obsn;

# specify the fixed-effects of the model, that there is no intercept and request the variance of
# the pooled estimates and their confidence interval
model diff=pd al/ noint s cl covb corrb ddf=4, 4;

# denote the between-study heterogeneity, and specify the Cholesky ('fa0(2)')
# parameterisation of the between-study covariance matrix (see Section 4.2)
random pd al / subjected g type=fa0(2);

# specify within-study heterogeneity, and specify the within-study covariance matrix in terms
# of the within-study correlations ('arh(1)'):
repeated / type=arh(1) subject=id group=id ;
# input the initial Cholesky parameter starting values (for $\alpha_{11}, \alpha_{12}, \alpha_{21}$)
 parms (1)
 (0.1)
 (0.1)

# input the known within-study covariance parameters (i.e. $s^2_i$ and $\rho_{W_i}$)
# study 1
 (0.0075)
 (0.0077)
 (0.385)

# study 2
 (0.0057)
 (0.0008)
 (0.421)

# study 3
 (0.0021)
 (0.0014)
 (0.408)

# study 4
 (0.0029)
 (0.0015)
 (0.432)

# study 5
 (0.0148)
 (0.0304)
 (0.339)

/ eqcons = 4 to 18;

# estimate the difference in pooled estimates
estimate 'pd-al' pd 1 al -1;
run;

(ii) Model A-reg extends Model A by including an additional regression covariate (e.g. for year of publication ('year'); see Section 6.7.4). The above SAS syntax for Model A can easily accommodate this by simply changing the 'model' statement to:

model diff=pd al pd*year al*year/ noint s cl covb corrb ddf=3, 3;
Appendix B2: Mathematical details behind the analytic solutions for Model A

The RIGLS solution minimises \((Y - X\beta)^T V^{-1} (Y - X\beta)\) w.r.t. \(\beta\) by differentiating with respect to \(\beta\) and setting the first derivative to zero; the solution is \(-2X^T V^{-1} (Y - X\beta) = 0\).

Thus rearranging obtains the RIGLS estimate \(\hat{\beta} = \left(X^T V^{-1} X\right)^{-1} X^T V^{-1} Y\). The covariance matrix of \(\hat{\beta}\) is \(\text{var}(\hat{\beta}) = \left(X^T V^{-1} X\right)^{-1}\), where \(V\) is the covariance matrix of \(Y\). I will now work out these formulae for Model A to obtain analytic expressions for \(\hat{\beta}\) and its variance.

For Model A, \(V\) will be block diagonal, as follows:

\[
V = \text{var}(Y) = \text{var}(X\beta + Z^{(1)} u + Z^{(0)} e) = \text{var}(X\beta) + \text{var}(Z u) + \text{var}(I_{2N} e) = 0 + Z \text{var}(u) Z^T + I_{2N} \text{var}(e) I_{2N}^T
= Z \Omega Z^T + I_{2N} \Omega I_{2N}^T
\]

Furthermore the inverse of \(V\) is:

\[
V^{-1} = \text{diag}
\left[
\begin{array}{cccc}
\frac{\tau_i^2}{\sigma_i^2} & \frac{s_{i1}^2}{\sigma_i^2} & \frac{s_{i2}^2}{\sigma_i^2} & \\
\frac{(\tau_i^2 + s_{i1}^2)(\tau_{i1}^2 + s_{i2}^2) - (\tau_{i2} + \lambda_1)^2}{(\tau_i^2 + s_{i1}^2)(\tau_{i1}^2 + s_{i2}^2) - (\tau_{i2} + \lambda_1)^2} & -\frac{(\tau_i^2 + s_{i1}^2)(\tau_{i1}^2 + s_{i2}^2) - (\tau_{i2} + \lambda_1)^2}{(\tau_i^2 + s_{i1}^2)(\tau_{i1}^2 + s_{i2}^2) - (\tau_{i2} + \lambda_1)^2} & \frac{(\tau_i^2 + s_{i1}^2)(\tau_{i1}^2 + s_{i2}^2) - (\tau_{i2} + \lambda_1)^2}{(\tau_i^2 + s_{i1}^2)(\tau_{i1}^2 + s_{i2}^2) - (\tau_{i2} + \lambda_1)^2} & \\
\frac{(\tau_i^2 + s_{i1}^2)(\tau_{i1}^2 + s_{i2}^2) - (\tau_{i2} + \lambda_1)^2}{(\tau_i^2 + s_{i1}^2)(\tau_{i1}^2 + s_{i2}^2) - (\tau_{i2} + \lambda_1)^2} & -\frac{(\tau_i^2 + s_{i1}^2)(\tau_{i1}^2 + s_{i2}^2) - (\tau_{i2} + \lambda_1)^2}{(\tau_i^2 + s_{i1}^2)(\tau_{i1}^2 + s_{i2}^2) - (\tau_{i2} + \lambda_1)^2} & \frac{(\tau_i^2 + s_{i1}^2)(\tau_{i1}^2 + s_{i2}^2) - (\tau_{i2} + \lambda_1)^2}{(\tau_i^2 + s_{i1}^2)(\tau_{i1}^2 + s_{i2}^2) - (\tau_{i2} + \lambda_1)^2} & \\
\frac{(\tau_i^2 + s_{i1}^2)(\tau_{i1}^2 + s_{i2}^2) - (\tau_{i2} + \lambda_1)^2}{(\tau_i^2 + s_{i1}^2)(\tau_{i1}^2 + s_{i2}^2) - (\tau_{i2} + \lambda_1)^2} & -\frac{(\tau_i^2 + s_{i1}^2)(\tau_{i1}^2 + s_{i2}^2) - (\tau_{i2} + \lambda_1)^2}{(\tau_i^2 + s_{i1}^2)(\tau_{i1}^2 + s_{i2}^2) - (\tau_{i2} + \lambda_1)^2} & \frac{(\tau_i^2 + s_{i1}^2)(\tau_{i1}^2 + s_{i2}^2) - (\tau_{i2} + \lambda_1)^2}{(\tau_i^2 + s_{i1}^2)(\tau_{i1}^2 + s_{i2}^2) - (\tau_{i2} + \lambda_1)^2} & \\
\end{array}
\right]
\]

I want to obtain an analytic expression for \(\hat{\beta}\) using \(\hat{\beta} = \left(X^T V^{-1} X\right)^{-1} X^T V^{-1} Y\). Firstly, I will calculate \(\left(X^T V^{-1} X\right)^{-1}\):
Therefore \((x'V^{-1}x)^{-1}\) is a 2 by 2 matrix, as follows:

\[
\begin{bmatrix}
\sum_{i=1}^{k} y_i^2 & \sum_{i=1}^{k} y_i \\
\sum_{i=1}^{k} y_i & k
\end{bmatrix}
\]

\[(x'V^{-1}x)^{-1} = \frac{1}{k\sum_{i=1}^{k} y_i^2 - \left(\sum_{i=1}^{k} y_i\right)^2 / k}
\]

which is also the \(\text{var}(\hat{\beta})\) is a 2 by 2 matrix, as follows:

\[
\begin{bmatrix}
\sum_{i=1}^{k} y_i^2 & \sum_{i=1}^{k} y_i \\
\sum_{i=1}^{k} y_i & k
\end{bmatrix}
\]

\[
\begin{bmatrix}
\frac{1}{k\sum_{i=1}^{k} y_i^2 - \left(\sum_{i=1}^{k} y_i\right)^2 / k} & 0 \\
0 & \frac{1}{k\sum_{i=1}^{k} y_i^2 - \left(\sum_{i=1}^{k} y_i\right)^2 / k}
\end{bmatrix}
\]
Now, using this information and the fact that \( \hat{\beta} = \left( \hat{\beta}_1 \right) = \left( \mathbf{X}^T \mathbf{V}^{-1} \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{y} \), the pooled estimate for \( j = 1 \) is:

\[
\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \left( \hat{y}_{1i} \left( x_{1i} \right)^2 \right) - \left( (x_{1i}) \right)^2}{\sum_{i=1}^{n} \left( (x_{1i}) \right)^2 - (\mathbf{X} \hat{\beta}_1)^2} \frac{\sum_{i=1}^{n} \left( x_{1i} \right)^2 \left( (x_{1i}) \right)^2 - (\mathbf{X} \hat{\beta}_1)^2}{\sum_{i=1}^{n} \left( (x_{1i}) \right)^2 - (\mathbf{X} \hat{\beta}_1)^2} \]

and simplifying gives:

\[
\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \left( (x_{1i}) \right)^2 - (\mathbf{X} \hat{\beta}_1)^2}{\sum_{i=1}^{n} \left( (x_{1i}) \right)^2 - (\mathbf{X} \hat{\beta}_1)^2} \]

Similarly, the GLS pooled estimate for outcome \( j = 2 \) is:

\[
\hat{\beta}_2 = \frac{\sum_{i=1}^{n} \left( (x_{2i}) \right)^2 - (\mathbf{X} \hat{\beta}_2)^2}{\sum_{i=1}^{n} \left( (x_{2i}) \right)^2 - (\mathbf{X} \hat{\beta}_2)^2} \]

Finally, \( \text{var} \left( \hat{\beta} \right) = \left( \mathbf{X}^T \mathbf{V}^{-1} \mathbf{X} \right)^{-1} \) and therefore:

\[
\text{var}(\hat{\beta}_1) = \frac{\sum_{i=1}^{n} \left( (x_{1i}) \right)^2 - (\mathbf{X} \hat{\beta}_1)^2}{\sum_{i=1}^{n} \left( (x_{1i}) \right)^2 - (\mathbf{X} \hat{\beta}_1)^2} \]

\[
\text{var}(\hat{\beta}_2) = \frac{\sum_{i=1}^{n} \left( (x_{2i}) \right)^2 - (\mathbf{X} \hat{\beta}_2)^2}{\sum_{i=1}^{n} \left( (x_{2i}) \right)^2 - (\mathbf{X} \hat{\beta}_2)^2} \]
and the $\text{cov}(\hat{\beta}_1, \hat{\beta}_2)$ is:

$$
\text{cov}(\hat{\beta}_1, \hat{\beta}_2) = \frac{\sum_{i=1}^{n} \frac{(r_{1i} + \lambda_i)}{(r_{1i}^2 + s_{1i}^2)} - (r_{1i} + \lambda_i)^2}{\sum_{i=1}^{n} \frac{r_{1i}^2 + s_{1i}^2}{(r_{1i}^2 + s_{1i}^2)} - (r_{12} + \lambda)^2} 
$$

**Appendix B3: Estimation of the between-study parameters in Model A**

(i) **Estimation of $\Omega_2$ using iterative generalised least squares (IGLS)**

The solutions shown in Appendix B2 are only part of the estimation procedure because, as well as estimating $\beta$, it is also necessary to estimate $\Omega_2$, i.e. $\tau_1^2$, $\tau_2^2$, and $\tau_{12}$. The IGLS estimation procedure iterates between estimating $\beta$ and $\Omega_2$ until the estimates for each parameter have converged to a pre-specified level (e.g. 6 decimal places). In order to estimate $\Omega_2$ in Model A the following is used, which incorporates the estimates of the level-2 (between-study) residuals: Let

$$
Y^* = \bar{y}y^T , \text{ where } \bar{y} = Y - X\beta - \Omega_1 \text{ are the between-study (level-2) residuals}
$$

$$
Y^* = \begin{pmatrix}
\bar{y}_{11} - s_{11} & \bar{y}_{12} - \lambda_1 & \ldots & \bar{y}_{1n} & \bar{y}_{1n} \\
\bar{y}_{12} - \lambda_1 & \bar{y}_{13} - s_{12} & \ldots & \bar{y}_{1n} & \bar{y}_{1n} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
\bar{y}_{1n} & \bar{y}_{1n} & \ldots & \bar{y}_{nn} - s_{1n} - \lambda_n & \bar{y}_{nn} \\
\bar{y}_{nn} - \lambda_n & \bar{y}_{nn} & \ldots & \bar{y}_{nn} - s_{nn} - \lambda_n \\
\end{pmatrix}
$$

and let

$$
Y^{**} = \begin{pmatrix}
\bar{y}_{11} - s_{11} \\
\bar{y}_{12} - \lambda_1 \\
\vdots \\
\bar{y}_{nn} - \lambda_n \\
\end{pmatrix}
\quad \psi = \begin{pmatrix}
\tau_1^2 \\
\tau_2^2 \\
\tau_{12}
\end{pmatrix}
$$

, where the $s_y^2$ and $\lambda_i$ terms are assumed known.
One needs to estimate $\psi$. Now, $E(Y^{**}) = Z^*\psi = \Omega_2$, where $Z^*$ is the design matrix for the variance components in $\Omega_2$. One can estimate $\psi$ by $\hat{\psi} = \left(Z^* V^{*\prime} Z\right)^{-1} Z^* V^{*\prime} Y^{**}$, where $V^*$ is the covariance matrix of $Y^{**}$ and is defined by $V^* = V \otimes V$ where $\otimes$ is the Kronecker product which multiplies every element of the left hand matrix by each element of the right hand matrix (N.B. $V$ is used in these equations because subtracting the $s_{ij}^2$s in $Y^*$ makes no difference on the var as the $s_{ij}^2$s are assumed known)[174]. The IGLS procedure iterates between calculating $\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y$ and

$$\hat{\psi} = \left(Z^* V^{*\prime} Z\right)^{-1} Z^* V^{*\prime} Y^{**}$$

until convergence is obtained, with the latest parameter values used at each iteration. Assuming multivariate normality for the errors it is also possible to obtain a covariance matrix for $\hat{\psi}$ using $\text{cov}(\hat{\psi}) = 2(ZV^{*\prime}Z)^{-1}$ [174].

Now it is extremely difficult to obtain analytic expressions for $\hat{\psi} = \left(Z^* V^{*\prime} Z\right)^{-1} Z^* V^{*\prime} Y^{**}$, because the matrices are so very large. To show the complexity of the equations I now consider the solutions when $n = 2$. The matrices for $n = 2$ are as follows:

\[
Y^{**} = \begin{pmatrix}
\bar{y}_{11} & \bar{y}_{12} & \bar{y}_{21} & \bar{y}_{22} \\
\bar{y}_{12} & \bar{y}_{22} & \bar{y}_{21} & \bar{y}_{11} \\
\bar{y}_{22} & \bar{y}_{21} & \bar{y}_{11} & \bar{y}_{12} \\
\bar{y}_{21} & \bar{y}_{11} & \bar{y}_{12} & \bar{y}_{22} \\
\bar{y}_{12} & \bar{y}_{22} & \bar{y}_{21} & \bar{y}_{11} \\
\bar{y}_{11} & \bar{y}_{22} & \bar{y}_{21} & \bar{y}_{12} \\
\bar{y}_{12} & \bar{y}_{21} & \bar{y}_{11} & \bar{y}_{22} \\
\bar{y}_{21} & \bar{y}_{11} & \bar{y}_{12} & \bar{y}_{22}
\end{pmatrix}
\]

where $E(Y^{**}) = Z^* \eta = \begin{pmatrix} \tau_1^2 \\ \tau_{12} \\ \tau_2^2 \\ \tau_{12} \\ \tau_1^2 \\ \tau_{12} \\ \tau_2^2 \\ \tau_{12} \end{pmatrix}$

and

\[
V = \begin{pmatrix}
\tau_1^2 + s_{11}^2 & \tau_{12} + \lambda_1 & 0 & 0 \\
\tau_{12} + \lambda_1 & \tau_2^2 + s_{12}^2 & 0 & 0 \\
0 & 0 & \tau_1^2 + s_{21}^2 & \tau_{12} + \lambda_2 \\
0 & 0 & \tau_{12} + \lambda_2 & \tau_2^2 + s_{22}^2
\end{pmatrix}
\]
Therefore V* is a 16 by 16 matrix:

2
(Tl

+ s J J 2 ) , ( xl 2+XI) ( xl 2 + s l l 2 ) ,

0, 0,

( t 12

+ X 1 ) ( t 12

+ 5 / / 2 ),

(xl2 + Xl ) ( x l 2 + s / / 2) , ( x l 2 + s U 2 ) ( x 2 2 + s 1 2 2 ) , 0, 0, (xl2 + Xl)2 ,

(x1 2 + X 1)2 ,

( t 12

0, 0, 0, 0, 0, 0, 0, 0, 0, 0

+ X ! ) ( x2 2 + s l 2 2) , G, 0, 0, 0, 0, 0, 0, 0, 0, 0

| 0 . 0 . ( x | 2 -*-:>7/2) ( x | 2 +.y2/2 ) , ( x l 2 + X 2 ) ( x | 2 + i 7 / 2) , 0 , 0 , ( x l 2 + X l ) ( x l 2 + i 2 / 2) , ( xl 2+XI ) ( xl 2 + X 2 ) , 0 , 0 , 0 , 0 , 0 , 0 , 0 , 0
0 , 0 , (i 12 +>.2)(x|2 + i 7 / 2 ) , ( x l 2 + i / / 2)(x22 + i / 2 2) , 0 , 0, ( T l 2 + XI )(x!2 + X2) ,(x!2 + XI) (x22 +*222) , 0 , 0 , 0 , 0 , 0 , 0 , 0 , 0
(x 12 + XI ) ( x l 2 + s / / 2), (xl2 + XI )2 , 0, 0, ( t 12 + * / / 2)( x22 + s / 2 2), ( xI 2+X1)( x2 2 + s / 2 2), 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
(xl2 + XI)2 , (xl2 + XI)(x2 2 + s I 2 2 ) , 0, 0, (xl2+

X I ) ( x2 2

+ j / 2 2),

( x2 2

2
+ s / 2 2) , 0, 0, 0, 0, 0, 0, 0, 0, 0, 0

0 . 0 , ( t 12 + X I ) ( x I 2 + s 2 / 2 ) , ( t I 2 + X I ) ( x 1 2 + X 2 ) , 0 , 0 , ( t 12 + j 2 / 2 ) ( t 2 2 + j / 2 2 ) , ( x 1 2 + X 2 ) ( x2 2 + 5 / 2 2 ) , 0 , 0 , 0 , 0 , 0 , 0 , 0 , 0

0 , 0 , ( x l 2 + X I ) ( x I 2 + X2),(xl2

+ X I ) ( x2 2 + * 2 2 2 ) , 0 , 0 , ( x I 2 + X 2 ) ( x2 2

+ 5 / 2 2 ) , ( x2 2

+ . s / 2 2 ) ( x 2 2 + . s2 2 2 ) , 0 , 0 , 0 , 0 , 0 , 0 , 0 , 0

0 . 0 . 0 , 0 , 0 , 0 , 0 , 0 . ( x l 2 + i / / 2) ( x l 2 + i-2/2).(xl2 + Xl )(xl 2 + 5 2/ 2) , 0 , 0 , ( x l 2 + X2)(xl2 + i / / 2),(xl 2 + Xl ) ( xl 2+X2) , 0. 0
0 . 0 , 0 . 0 . 0 . 0 , 0 . 0 , ( x l 2 + XI ) ( x l 2 + i 2 / 2) , ( x l 2 + i 2 / 2) ( xl 2 + i 7 2 2) , 0 , 0 , ( x l 2 + XI)(xl2 +X2),(xl2 +X2)(x22 + 5 2 / 2) , 0 , 0
2
0 , 0, 0. 0, 0, 0, 0. 0. 0. 0, ( xl 2 + s ’ / 2) , (x!2+X2)(xl 2 + s 2 l 2 ) , 0, 0, (xl2 +X2)(xl2 + s 2 1 2) , (xl2 + X2)2
0, 0. 0, 0. 0, 0, 0 , 0 . 0 , 0 , (xl 2+X2)( xl 2 + s 2 / 2), ( xl 2 + i 2 / 2)(x22 + *222) , 0 , 0 , (xl2 +X2)2 , (xl2 +X2)(x22 + a222)
0 . 0 , 0 , 0 , 0 . 0 , 0 , 0 , ( x l 2 +X2)(xl2 + 5 / / 2),(xl 2 + XI ) ( xl 2+ X2 ) , 0, 0 , ( x2 2 + s222) ( xl 2 + 5 / / 2 ) , ( x I 2 + X 1 ) ( x2 2 + s222), 0 , 0
0 , 0 , 0 , 0 , 0 , 0 , 0 , 0 , ( x I 2 + XI ) (xl 2+X2),(xl 2 + X2)(x22 + s / 2 2) , 0 , 0 , ( x l 2 + Xl)(x22 + s222) , ( x l 2 + W22) ( xl 2 + s222) , 0 , 0
0.

0, 0. 0, 0. 0, 0. 0. 0. 0, ( xl 2+X2) ( xl 2 + 5 2 / 2), (xI2 + X2)2 , 0, 0, ( x \ 2 + s 2 ! 2 ) ( x 2 2 + s 2 2 2 ) , (xl2 + X2)(x22 + s 2 2 2 )
2

L0. 0. 0, 0. 0. 0, 0, 0, 0, 0, (t12 -t-X2)2 . (x12 ^X2) ( x22 + s 2 2 2 ). 0, 0, (xl2 +X2)(x22 +.v222). (x22 +.v222 ) ^

It is a non-trivial exercise to obtain the inverse of V* (and this is for n = 2 remember!).
The Maths package Maple had not solved the inverse o f V* after 8 hours. As V* is block
diagonal (with two 8 x 8 blocks), I even tried to simplify matters by obtaining the inverses
for these two blocks separately, but Maple still had not found the analytic solution after a
considerable time. Hence, I could not obtain the subsequent analytic solutions for the
parameters in $ but needless to say they will be also extremely complex.

To help emphasise this complexity, I set A, = >l2 = 0 and Maple was then able to compute
(after a short time) the inverse of V* for n = 2. The solution for if was 10 pages long and
Maple could not simplify it further, despite extensive attempts. However, to help check

R.D. Riley, Ph.D. Thesis

Appendix B

390


that the solutions were correct, I set $\tau_{12} = 0$, and obtained the following solution from Maple for $\bar{\tau}_1^2$ (i.e. cell (1,1) in $\Psi$):

\[
- (-R_{21}^2 s1l^4 - R_{11}^2 \tau_1^4 - 2 R_{11}^2 s21^2 \tau_1^2 - R_{11}^2 s21^4 + S1l^2 s21^4 + S1l^2 \tau_1^4
- 2 R_{21}^2 \tau_1^2 s1l^2 + 2 S1l^2 s21^2 \tau_1^2 + 2 S1l^2 s1l^2 - R_{11}^2 \tau_1^4 + S21^2 \tau_1^4
+ S21^2 s1l^4) / (2 \tau_1^4 + 2 s21^2 \tau_1^2 + 2 \tau_1^2 s1l^2 + s1l^4 + s21^4))
\]

Factorising this gives:

\[
\bar{\tau}_1^2 = \sum_{i=1}^{2} \frac{R_{ii}^2 - s_{ii}^2}{\sum_{i=1}^{2} \frac{1}{(s_{ii}^2 + \tau_1^2)^2}}
\]

, where $\bar{\tau}_1^2$ on the right of the equation is the estimate from previous iteration and where $Y - X\hat{B} = (R_{11}, R_{12}, R_{21}, R_{22}, \ldots, R_{n1}, R_{n2})$. This formula concurs with the univariate solution for $\bar{\tau}_1^2$ shown in Section 1.7.2 and elsewhere [45].

**(ii) Estimation of $\Omega_2$ using restrictive iterative generalised least squares (RIGLS)**

An alternative and preferable estimation procedure to IGLS is RIGLS, which uses a slightly modified form of the $\hat{\Omega}_2$ equations for IGLS in order to obtain an unbiased estimate of $\Omega_2$ and therefore $\tau_1^2$, $\tau_2^2$, and $\tau_{12}$. To obtain the RIGLS solution for $\hat{\Omega}_2$,

$\bar{\gamma}\bar{y}^T$ in Appendix B3(i) above is replaced by $\bar{\gamma}\bar{y}^T + X(X^TV^{-1}X)^{-1}X^T$, where the second term can be thought of as a correction for bias [174]. N.B. The analytic equations for $\hat{\beta}$ and its variance, presented in Appendix B2, are not modified by RIGLS estimation.
Appendix B4: Weighting of each study toward the pooled estimates in Model A

Without loss of generality consider the pooled solution for \( j = 1 \). It is evident that each study \( i \) now contributes \( A_i \hat{Y}_{i1} + B_i \hat{Y}_{i2} \) to the estimate of \( \beta_1 \), where

\[
A_i = \frac{1}{\left( (\tilde{r}_i + s_i^2)(\tilde{r}_i + s_i^2) - (\tilde{r}_{12} + \lambda_i) \right)^2} \left[ \sum_{i=1}^{n} \frac{\left( \tilde{r}_i^2 + s_i^4 \right) (\tilde{r}_i^2 + s_i^4) - (\tilde{r}_{12} + \lambda_i)(\tilde{r}_{12} + \lambda_i)}{\left( (\tilde{r}_i + s_i^2)(\tilde{r}_i + s_i^2) - (\tilde{r}_{12} + \lambda_i) \right)} \right]
\]

\[
B_i = \frac{1}{\left( (\tilde{r}_i + s_i^2)(\tilde{r}_i + s_i^2) - (\tilde{r}_{12} + \lambda_i) \right)^2} \left[ \sum_{i=1}^{n} \frac{\left( \tilde{r}_i^2 + s_i^4 \right) + \lambda_i(\tilde{r}_i^2 + s_i^4) - \lambda_i^2(\tilde{r}_i^2 + s_i^4) \right]}{\left( (\tilde{r}_i + s_i^2)(\tilde{r}_i + s_i^2) - (\tilde{r}_{12} + \lambda_i) \right)} \right]
\]

So for example, for 3 studies, \( \hat{\beta}_1 = A_1 \hat{Y}_{11} + B_1 \hat{Y}_{12} + A_2 \hat{Y}_{21} + B_2 \hat{Y}_{22} + A_3 \hat{Y}_{31} + B_3 \hat{Y}_{32} \).

Furthermore, I will now show that \( \sum_{i=1}^{n} (A_i + B_i) = 1 \).

\( A_i \) and \( B_i \) have the same denominator for all \( i \), therefore:

\[
\sum_{i=1}^{n} (A_i + B_i) = \frac{1}{\left( (\tilde{r}_i + s_i^2)(\tilde{r}_i + s_i^2) - (\tilde{r}_{12} + \lambda_i) \right)^2} \left[ \sum_{i=1}^{n} \frac{\left( \tilde{r}_i^2 + s_i^4 \right) (\tilde{r}_i^2 + s_i^4) - (\tilde{r}_{12} + \lambda_i)(\tilde{r}_{12} + \lambda_i)}{\left( (\tilde{r}_i + s_i^2)(\tilde{r}_i + s_i^2) - (\tilde{r}_{12} + \lambda_i) \right)} \right]
\]

Consider now just the top line of this equation. This can be expanded out to give:

\[
\sum_{i=1}^{n} \left[ \frac{\left( \tilde{r}_i^2 + s_i^4 \right) (\tilde{r}_i^2 + s_i^4)}{(\tilde{r}_i + s_i^2)(\tilde{r}_i + s_i^2) - (\tilde{r}_{12} + \lambda_i)^2} \right] - \sum_{i=1}^{n} \left[ \frac{(\tilde{r}_i^2 + s_i^4)}{(\tilde{r}_i + s_i^2)(\tilde{r}_i + s_i^2) - (\tilde{r}_{12} + \lambda_i)^2} \right] \sum_{i=1}^{n} \left[ \frac{(\tilde{r}_i^2 + s_i^4)}{(\tilde{r}_i + s_i^2)(\tilde{r}_i + s_i^2) - (\tilde{r}_{12} + \lambda_i)^2} \right] \sum_{i=1}^{n} \left[ \frac{(\tilde{r}_i^2 + s_i^4)}{(\tilde{r}_i + s_i^2)(\tilde{r}_i + s_i^2) - (\tilde{r}_{12} + \lambda_i)^2} \right]
\]
Collating back together and reintroducing the bottom line gives:

$$\sum_{i=1}^{n} \sum_{k=1}^{n} \left( \tilde{r}_{j}^{2} + s_{i}^{2} \right) \left( \tilde{r}_{k}^{2} + s_{k}^{2} \right) - (\tilde{r}_{12} + \lambda_{i})(\tilde{r}_{12} + \lambda_{j})\left( \tilde{r}_{12}^{2} + s_{k}^{2} \right) - \left( \tilde{r}_{12} + \lambda_{i} \right)\left( \tilde{r}_{12}^{2} + s_{j}^{2} \right) - (\tilde{r}_{12} + \lambda_{k})^{2}$$

One can see that the top and bottom of this equation both have the same denominator.

Furthermore, one can actually remove the $$-(\tilde{r}_{12} + \lambda_{i})(\tilde{r}_{12} + \lambda_{j})\left( \tilde{r}_{12}^{2} + s_{k}^{2} \right) - \left( \tilde{r}_{12} + \lambda_{i} \right)\left( \tilde{r}_{12}^{2} + s_{j}^{2} \right) - (\tilde{r}_{12} + \lambda_{k})^{2}$$ term from the very top line because these cancel each other out over the entire sum as both $$i$$ and $$k$$ go from $$1$$ to $$n$$. Thus, there now remains the following:

$$\sum_{i=1}^{n} \sum_{k=1}^{n} \left( \tilde{r}_{j}^{2} + s_{i}^{2} \right) \left( \tilde{r}_{k}^{2} + s_{k}^{2} \right) - (\tilde{r}_{12} + \lambda_{i})(\tilde{r}_{12} + \lambda_{j})\left( \tilde{r}_{12}^{2} + s_{k}^{2} \right) - \left( \tilde{r}_{12} + \lambda_{i} \right)\left( \tilde{r}_{12}^{2} + s_{j}^{2} \right) - (\tilde{r}_{12} + \lambda_{k})^{2}$$

If one multiplies out and collates the summations in the denominator of the above term, one will find that that the denominator is actually identical to the numerator, i.e.:

$$\sum_{i=1}^{n} \sum_{k=1}^{n} \left( \tilde{r}_{j}^{2} + s_{i}^{2} \right) \left( \tilde{r}_{k}^{2} + s_{k}^{2} \right) - (\tilde{r}_{12} + \lambda_{i})(\tilde{r}_{12} + \lambda_{j})\left( \tilde{r}_{12}^{2} + s_{k}^{2} \right) - \left( \tilde{r}_{12} + \lambda_{i} \right)\left( \tilde{r}_{12}^{2} + s_{j}^{2} \right) - (\tilde{r}_{12} + \lambda_{k})^{2}$$

Hence, $$\sum_{i=1}^{n} (A_{i} + B_{i}) = 1$$.
Appendix B5: S-PLUS code to simulate data from Model A

The S-Plus code used to simulate data from Model A (see Section 5.2.1) is shown below, and the program is denoted 'bimeta':

```splus
bimeta <- function(nsim, nstud, beta, tau1, tau2, corr2, s1, s2, corrl)
{
  # nsim - number of simulations (scalar)
  # nstud - number of studies (scalar)
  # beta - true overall pooled HR (vector of length 2; i.e. OS pooled HR and DFS pooled HR)
  # tau1 - between study standard deviation for outcome 1 (scalar)
  # tau2 - between study standard deviation for outcome 2 (scalar)
  # corr2 - between study correlation (scalar)
  # s1 - standard error of the HR estimate for outcome 1 in each study (vector of length nstud)
  # s2 - standard error of the HR estimate for outcome 2 in each study (vector of length nstud)
  # corrl - within-study correlation between OS and DFS HR estimates in each study (vector of length nstud)
  
  # Generate simulation id and study id
  simid <- rep(1:nsim, rep(nstud, nsim))
  studyid <- rep(1:nstud, nsim)
  
  # Generate thetas (i.e. the \( \theta_j \)s, see Step 2 in Section 5.2.1)
  theta <- rmvnorm(n = nsim * nstud, mean = beta, sd = c(tau1, tau2), rho = corr2, d = 2)
  
  # Stack the \( s_j \)s and within-study correlations to enable one to generate the \( Y_{ij} \)s
  s1.repeat <- rep(s1, nsim)
  s2.repeat <- rep(s2, nsim)
  vector <- cbind(s1.repeat, s2.repeat)
  corrl.repeat <- rep(corrl, nsim)
  
  # Generate the data (i.e. the \( Y_{ij} \)s, see Step 2 in Section 5.2.1)
  y <- rmvnorm(n = nsim * nstud, mean = theta, sd = vector, rho = corrl.repeat, d = 2)
  
  # Return the data in a tabulated form
  ret <- cbind(simid, studyid, theta, y, vector, corrl.repeat)
  dimnames(ret)[[2]] <- c("simid", "studyid", "theta1", "theta2", "y1", "y2", "s1", "s2", "corrl")
  return(ret)
}
```

For example, consider 1000 simulations for \( n = 50 \) studies ('nstud') where \( \text{beta} = (0,2), \tau_1 = \tau_2 = 0.5, \) within and between correlation = 0.8, and \( s_j \)s are as specified below. To perform the simulation from the 'bimeta' program one needs to specify `answer<-bimeta(nsim, nstud, beta, tau1, tau2, corr2, s1, s2, corrl)`, as follows:

```splus
answer<-bimeta(1000, 50, c(0,2), 0.5, 0.5, 0.8, c(0.404723, 0.347741, 0.447079, 0.328599, 0.225539, 0.300663,
                              0.439057, 0.405453, 0.313644, 0.051227, 2.911511, 0.977952, 0.515025, 0.782173, 0.547235, 0.297652, 0.688934,
                              0.715069, 0.344505, 0.890901, 1.2258, 0.173675, 0.271026, 0.248307, 0.225556, 0.623, 0.670567, 0.842108, 4.032389,
                              0.415638, 0.572102, 0.475132, 0.197862, 0.58636, 1.731292, 0.649311, 1.057469, 0.796258, 0.827002, 0.49161,
                              1.667605, 0.295874, 1.324152, 0.655519, 0.427458, 1.708746, 0.495606, 2.36817, 0.372497, 0.219897, 0.375355,
                              0.243129, 0.482223, 0.861778, 0.40251, 0.325179, 1.205242, 0.418773, 2.796543, 0.45531, 1.572266, 0.49356, 0.40792,
                              0.35384, 0.913133, 1.020756, 0.548832, 0.918017, 0.424918, 0.575893, 0.421968, 1.049944, 0.668157, 0.402123,
                              1.27743, 0.283208, 0.733669, 1.235447, 0.424295, 1.318702, 0.443653, 1.207076, 0.576207, 0.295128, 3.747028,
                              0.411737, 0.370139, 0.774149, 0.278736, 0.687943, 0.601798, 0.449012, 1.505477, 0.384549, 0.230605, 0.927665,
                              1.007345, 2.237176, 0.277983, c(0.8,0.8,0.8,0.8,0.8,0.8,0.8,0.8,0.8,0.8,0.8,0.8,0.8,0.8,0.8,0.8,0.8,0.8,0.8,0.8,
                              0.8,0.8,0.8,0.8,0.8,0.8,0.8,0.8,0.8,0.8,0.8,0.8,0.8))
```

To export the simulated data one could specify:

```splus
write.table(answer, "c://extension")
```
**Appendix B6**: Simulation results from Model A for complete-case data from 25 studies

These are the simulation results from the univariate (URMA) and bivariate (Model A) random-effects meta-analysis models for complete-case data from 25 studies. A description of the simulation procedure is given in Section 5.2.1 (except here 25 studies were used and not 50), and the true parameter values to compare the results to are \( \beta_1 = 0, \beta_2 = 2, \tau_1^2 = 0.25, \tau_2^2 = 0.25 \). In Model A the within-study correlation (\( \rho_W \)) was known and the same for each study, whilst the between-study correlation (\( \rho_B \)) was estimated. The true values of \( \rho_W \) and \( \rho_B \) in each simulation are shown in the table.

| Meta-analysis model | \( \rho_W \) | \( \rho_B \) | Converged out of 1000 | Mean of \( \hat{\beta}_1 \) (s.e. of mean) | Mean s.e. of \( \hat{\beta}_1 \) | MSE of \( \hat{\beta}_1 \) | No. of 95% CIs for \( \hat{\beta}_1 \) including \( \beta_1 \) (%) | Mean of \( \hat{\beta}_2 \) (s.e. of \( \hat{\beta}_2 \)) | Mean s.e. of \( \hat{\beta}_2 \) | MSE of \( \hat{\beta}_2 \) | No. of 95% CIs for \( \hat{\beta}_2 \) including \( \beta_2 \) (%) | Mean of \( \hat{\tau}_1^2 \) (no. of \( \hat{\tau}_1^2 = 0 \)) | Mean of \( \hat{\tau}_2^2 \) (no. of \( \hat{\tau}_2^2 = 0 \)) | Mean of \( \hat{\rho}_B \) = \( \frac{\hat{\rho}_B - 1}{\hat{\rho}_B} \) (%) | No. of \( \frac{\hat{\rho}_B - 1}{\hat{\rho}_B} = 1 \) (%) |
|---------------------|--------------|--------------|----------------------|---------------------------------|---------------------|-----------------|----------------------|---------------------------------|---------------------|-----------------|----------------------|---------------------------------|----------------------|---------------------------------|----------------------|----------------------|
| URMA               | 0.8          | 0.8          | 1000                 | -0.00166 (0.137)              | 0.131               | 0.0187          | 945 (94.5%)          | 1.995 (0.147)              | 0.146               | 0.0216          | 949 (94.9%)          | 0.248 (2)                     | 0.253 (5)                      | -                               | -                                 |
| Model A             | 0.8          | 0.8          | 1000                 | -0.00206 (0.129)              | 0.125               | 0.0167          | 935 (93.5%)          | 1.995 (0.136)              | 0.134               | 0.0184          | 943 (94.3%)          | 0.250 (1)                     | 0.253 (0)                      | 0.780                           | 6 (99)                           |
**APPENDIX C**

**Appendix C1: WinBUGS version 1.3 syntax to fit Model A-Bayes**

The syntax used was as follows:

```
Model

# define the number of studies in the meta-analysis
{
  for(i in 1:n)
    # define the Known within-study covariance matrix and let WinBUGS calculate its inverse
    {
      covariance[1,1]<-variance1[i]+tau[1]
      covariance[i,2,2]<-variance2[i]+tau[2]
      covariance[i,1,2]<-cov[i]+rhob*sqrt(tau[1]*tau[2])
      covariance[i,2,1]<-covariance[i,1,2]
    for (k in 1 : 2) {
      for (j in 1 : 2) {
        prec[i,k,j] <- inverse(covariance[i,k,j])
      }
    }

    # specify the distribution of the data
    surv[i,1:2] ~ dmnorm(beta[1:2],prec[i,1:2, 1:2])
    }

    # specify the prior distribution of your choice for each of the unknown parameters above;
    # for example, I have specified vague prior distributions as follows
    tauinv[1]<-dgamma(0.001, 0.001)
    tauinv[2]<-dgamma(0.001, 0.001)
    beta[1]<dnorm(0.0, 1.0E-6)
    beta[2]<dnorm(0.0, 1.0E-6)
    rhob~dunif(-1,1)
    }

    # specify the data; for example, the original Berkey dataset is specified below.
    Data
    list(n=5, variance1=c(0.0075, 0.0057, 0.0021, 0.0029, 0.0148), variance2=c(0.0077, 0.0008, 0.0014, 0.0015, 0.0304), cov=c(0.003, 0.0009, 0.0007, 0.0009, 0.0072),
        surv=structure(.Data=c(0.47, -0.32, 0.2, -0.6, 0.4, -0.12, 0.26, -0.31, 0.56, -0.39), .Dim =c(5,2)))

    # specify the initial starting values of the parameters
    Inits
    list(tauinv=c(0.1,0.1), rhob=0.5, beta=c(0,0))
```

---

*R.D. Riley, Ph.D. Thesis Appendix C 396*
### Appendix C2: Additional Model A-zero simulation results for complete-case data

Simulation results from the univariate (URMA) and bivariate (Model A) random-effects meta-analysis for complete-case data of 50 and 5 studies for true parameter values of \( \beta_1 = 0 \), and \( \beta_2 = 2 \). The 95% confidence intervals (CIs) for the pooled estimates were calculated using a t-distribution with \((n-1)\) degrees of freedom. The within-study correlation (\( \rho_w \)) was known to be 0.8 and the same for each study, whilst the between-study correlation (\( \rho_b \)) was estimated. However, for Model A-zero \( \rho_w \) was wrongly set to zero to assess the impact of this.

1. \( \tau_1^2 = \tau_2^2 = 0.0025 \) (on average much smaller than the \( s_j^2 \)): Model A-zero would rarely converge
2. \( \tau_1^2 = \tau_2^2 = 1.5 \) (on average larger than the \( s_j^2 \))

| No. of studies \((n)\) | Meta-analysis model | \( \rho_w \) | \( \rho_B \) | Converged out of 1000 | Mean of \( \hat{\beta}_1 \) (s.e. of mean) | Mean s.e. of \( \hat{\beta}_1 \) | MSE of \( \hat{\beta}_1 \) (s.e.) | No. of 95% CIs including \( \beta_1 \) (%) | Mean of \( \hat{\beta}_2 \) (s.e.) | MSE of \( \hat{\beta}_2 \) (s.e.) | No. of 95% CIs including \( \beta_2 \) (%) | Mean of \( \tau_1^2 \) (no. = 0) | Mean of \( \tau_2^2 \) (no. = 0) | Mean of \( \hat{\rho}_B \) \( = \) \( -1 \) (%) | Mean of \( \hat{\rho}_B \) \( = \) \( 1 \) (%) |
|-------------------------|---------------------|---------------|---------------|-----------------------|-------------------------------------------|------------------------|-------------------|-----------------|------------------------|-------------------|-------------------|-----------------|------------------|-----------------|-----------------|-----------------|
| 50                      | URMA                | 0.8           | 0.8           | 1000                  | -0.001 (0.195)                           | 0.198                  | 0.0380            | 958 (95.8%)      | 1.995 (0.196)                          | 0.200              | 0.0383            | 949 (94.9%)      | 1.515 (0)        | 1.510 (0)       | -               | -                |
| 50                      | Model A             | 0.8           | 0.8           | 1000                  | 0.0010 (0.189)                           | 0.194                  | 0.0356            | 957 (95.7%)      | 1.995 (0.189)                          | 0.196              | 0.0358            | 956 (95.6%)      | 1.509 (0)        | 1.514 (0)       | 0.798 (0)      | 0 (0%)           |
| 50                      | Model A-zero        | 0.8           | 0.8           | 1000                  | 0.0011 (0.191)                           | 0.196                  | 0.0365            | 951 (95.1%)      | 1.995 (0.193)                          | 0.198              | 0.0372            | 955 (95.5%)      | 1.574 (0)        | 1.576 (0)       | 0.908 (0)      | 0 (0%)           |
| 5                       | URMA                | 0.8           | 0.8           | 1000                  | 0.0189 (0.579)                           | 0.542                  | 0.335             | 945 (94.5%)      | 2.011 (0.588)                          | 0.552              | 0.346             | 941 (94.1%)      | 1.512 (10)       | 1.525 (6)       | -               | -                |
| 5                       | Model A             | 0.8           | 0.8           | 1000                  | 0.0183 (0.579)                           | 0.538                  | 0.335             | 946 (94.6%)      | 2.012 (0.590)                          | 0.546              | 0.348             | 941 (94.1%)      | 1.511 (0)        | 1.522 (0)       | 0.745 (2.3%)    | 190 (19.0%)       |
| 5                       | Model A-zero        | 0.8           | 0.8           | 999                   | 0.0187 (0.578)                           | 0.545                  | 0.335             | 950 (95.1%)      | 2.103 (0.589)                          | 0.550              | 0.346             | 949 (95.0%)      | 1.529 (0)        | 1.550 (0)       | 0.848 (1.3%)    | 464 (46.4%)       |

N.B. \( \hat{\rho}_B \) was not applicable when one or both of \( \hat{\tau}_1^2 \) and \( \hat{\tau}_2^2 \) was zero. MSE = mean-square-error, CIs = confidence intervals, s.e. = standard error

**Additional results for \((\hat{\beta}_1 - \hat{\beta}_2)\) from (ii):**

<table>
<thead>
<tr>
<th>Meta-analysis model</th>
<th>( n )</th>
<th>Mean ((\hat{\beta}_1 - \hat{\beta}_2))</th>
<th>Mean s.e. ((\hat{\beta}_1 - \hat{\beta}_2))</th>
<th>Mean corr ((\hat{\beta}_1, \hat{\beta}_2))</th>
<th>Meta-analysis model</th>
<th>( n )</th>
<th>Mean ((\hat{\beta}_1 - \hat{\beta}_2))</th>
<th>Mean s.e. ((\hat{\beta}_1 - \hat{\beta}_2))</th>
<th>Mean corr ((\hat{\beta}_1, \hat{\beta}_2))</th>
</tr>
</thead>
<tbody>
<tr>
<td>URMA 50</td>
<td>-1.996</td>
<td>0.282</td>
<td>0.660</td>
<td>0</td>
<td>URMA 5</td>
<td>-1.995</td>
<td>0.380</td>
<td>0.710</td>
<td>0</td>
</tr>
<tr>
<td>Model A 50</td>
<td>-1.996</td>
<td>0.134</td>
<td>0.760</td>
<td>0</td>
<td>Model A 5</td>
<td>-1.996</td>
<td>0.752</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Model A-zero 50</td>
<td>-1.996</td>
<td>0.138</td>
<td>0.750</td>
<td>0</td>
<td>Model A-zero 5</td>
<td>-1.996</td>
<td>0.387</td>
<td>0.689</td>
<td>0</td>
</tr>
</tbody>
</table>
Appendix C3: Additional Model A-zero simulation results for missing data

**Part (a):** Simulation results from a univariate (URMA) and bivariate (Model A) random-effects meta-analysis for 50 and 10 studies, where outcome \( j = 1 \) was available from all studies but outcome \( j = 2 \) was *missing completely at random (MCAR)* from half the studies in each case. The true parameter values to compare the results to are \( \beta_1 = 0 \) and \( \beta_2 = 2 \). In Model A the within-study correlation (\( \rho_w \)) was known and the same for each study, whilst the between-study correlation (\( \rho_B \)) was estimated. In Model A-zero \( \rho_w \) was wrongly assumed zero to assess the impact of this.

1. \( \tau_1^2 = \tau_2^2 = 0.0025 \) (on average much smaller than the \( s_{ij}^2 \)'s) - Model A-zero would rarely converge
2. \( \tau_1^2 = \tau_2^2 = 1.5 \) (on average larger than the \( s_{ij}^2 \)'s)

| No. of studies | Meta-analysis model | \( \rho_w \) | \( \rho_B \) | Converged out of 1000 | Mean of \( \hat{\beta}_1 \) (s.e. of mean) | Mean s.e. of \( \hat{\beta}_1 \) | MSE of \( \hat{\beta}_1 \) | No. of 95% CIs for \( \hat{\beta}_1 \) including \( \beta_1 \) (%) | Mean of \( \hat{\beta}_2 \) (s.e.) | Mean s.e. of \( \hat{\beta}_2 \) | MSE of \( \hat{\beta}_2 \) | No. of 95% CIs for \( \hat{\beta}_2 \) including \( \beta_2 \) (%) | Mean of \( \hat{\tau}_1^2 \) (no. of \( \hat{\tau}_1^2 = 0 \)) | Mean of \( \hat{\tau}_2^2 \) (no. of \( \hat{\tau}_2^2 = 0 \)) | Mean of \( \rho_B \) | No. of \( \rho_B = 1 \) (%) | No. of \( \rho_B = -1 \) (%) |
|---------------|---------------------|-------------|-------------|---------------------|------------------------------------------|-----------------|-----------------|-----------------------------------------------|-----------------|-----------------|-----------------|-----------------------------------------------|-----------------|-----------------|-----------------|-----------------------------------------------|-----------------|-----------------|
| 50            | URMA                | 0.8         | 0.8         | 1000                | -0.0027 (0.0689)            | 0.0708          | 0.0047          | 953 (95.3%)                  | 1.999 (0.0974)  | 0.0996          | 0.0095          | 965 (96.5%)                  | 0.244 (0)       | 0.246 (0)       | -               | -               | -               |
| 50            | Model A             | 0.8         | 0.8         | 1000                | -0.0027 (0.0689)            | 0.0708          | 0.0048          | 954 (95.4%)                  | 2.000 (0.0822)  | 0.0824          | 0.0067          | 958 (95.8%)                  | 0.244 (0)       | 0.246 (0)       | 0.798 (0)       | 0               | 0               |
| 50            | Model A-zero        | 0.8         | 0.8         | 1000                | -0.0051 (0.100)             | 0.101           | 0.0101          | 952 (95.2%)                  | 1.995 (0.128)   | 0.131           | 0.0163          | 957 (95.7%)                  | 0.255 (0)       | 0.230           | 0.986 (0)       | 0               | 841 (0%         |
| 10            | URMA                | 0.8         | 0.8         | 1000                | 0.0003 (0.155)              | 0.155           | 0.0239          | 951 (95.1%)                  | 2.003 (0.226)   | 0.215           | 0.0509          | 952 (95.2%)                  | 0.248 (0)       | 0.258           | -               | -               | -               |
| 10            | Model A             | 0.8         | 0.8         | 1000                | 0.0003 (0.155)              | 0.155           | 0.0239          | 952 (95.2%)                  | 2.008 (0.194)   | 0.176           | 0.0375          | 951 (95.1%)                  | 0.248 (0)       | 0.263           | 0.763 (0)       | 0               | 10 (1%)         |
| 10            | Model A-zero        | 0.8         | 0.8         | 1000                | -0.0076 (0.2223)            | 0.223           | 0.0499          | 948 (94.8%)                  | 1.986 (0.268)   | 0.251           | 0.0721          | 980 (98.0%)                  | 0.273 (3)       | 0.275           | 0.902 (3.5%)    | 35 (87.5%)       | -               |

N.B. \( \rho_B \) was not applicable when one or both of \( \hat{\tau}_1^2 \) and \( \hat{\tau}_2^2 \) was zero. MSE = mean-square-error, CIs = confidence intervals, s.e. = standard error

The 95% CI for \( \hat{\beta}_j \) was calculated using a t-distribution with (\( n_j - 1 \)) degrees of freedom, where \( n_j \) is the number of studies for which outcome \( j \) was available. So for example, for \( n = 50 \) the degrees of freedom for \( \hat{\beta}_1 \) was 49 whereas the degrees of freedom for \( \hat{\beta}_2 \) was 24.
Part (b): Simulation results from a univariate (URMA) and bivariate (Model A) random-effects meta-analysis for 50 and 10 studies, where outcome $j = 2$ was available from all studies but outcome $j = 1$ was missing at random (MAR) if $Y_{ni}$ was negative. The true parameter values to compare the results to are $\beta_1 = 0$ and $\beta_2 = 2$. In Model A the within-study correlation $\rho_w$ was known and the same for each study, whilst the between-study correlation $\rho_B$ was estimated. The values of $\rho_w$ and $\rho_B$ in each simulation are shown in the table.

(i) $\tau_1^2 = \tau_2^2 = 0.0025$ (on average much smaller than the $s_{ij}^2$s) - Model A-zero would rarely converge

(ii) $\tau_1^2 = \tau_2^2 = 1.5$ (on average larger than the $s_{ij}^2$s) - Model A-zero results were very similar to those for Model A

(iii) $\tau_1^2 = \tau_2^2 = 0.0025$ (on average similar in size to the $s_{ij}^2$s):

<table>
<thead>
<tr>
<th>No. of studies (n)</th>
<th>Meta-analysis model</th>
<th>$\rho_w$</th>
<th>$\rho_B$</th>
<th>Converged out of 1000</th>
<th>Mean of $\hat{\beta}_1$ (s.e. of mean)</th>
<th>Mean s.e. of $95%$ CIs for $\hat{\beta}_1$ including $\beta_1$ (%)</th>
<th>MSE of $\hat{\beta}_1$ (s.e.)</th>
<th>No. of $95%$ CIs for $\hat{\beta}_1$ (no. of $\hat{\beta}_1^2=0$)</th>
<th>Mean of $\hat{\beta}_2$ (s.e. of mean)</th>
<th>Mean s.e. of $95%$ CIs for $\hat{\beta}_2$ including $\beta_2$ (%)</th>
<th>MSE of $\hat{\beta}_2$ (s.e.)</th>
<th>No. of $95%$ CIs for $\hat{\beta}_2$ (no. of $\hat{\beta}_2^2=0$)</th>
<th>Mean of $\hat{\rho}_B$ =</th>
<th>No. of $\hat{\rho}_B$ = %</th>
</tr>
</thead>
<tbody>
<tr>
<td>50 URMA</td>
<td>0.8</td>
<td>0.8</td>
<td>998</td>
<td></td>
<td>0.505 (0.105)</td>
<td>0.0928 (0.106)</td>
<td>0.266 (0.012)</td>
<td>1 (99.9%)</td>
<td>1.993 (0.107)</td>
<td>0.106 (0.012)</td>
<td>0.111 (94.5%)</td>
<td>943 (117)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>50 Model A</td>
<td>0.8</td>
<td>0.8</td>
<td>983*</td>
<td></td>
<td>0.314 (0.096)</td>
<td>0.0840 (0.108)</td>
<td>0.108 (0.012)</td>
<td>893 (90.8%)</td>
<td>1.979 (0.103)</td>
<td>0.101 (0.012)</td>
<td>0.111 (93.8%)</td>
<td>922 (0)</td>
<td>0.0759 (0)</td>
<td>0.245 (0)</td>
</tr>
<tr>
<td>50 Model A-zero</td>
<td>0.8</td>
<td>0.8</td>
<td>988</td>
<td></td>
<td>0.416 (0.0833)</td>
<td>0.0974 (0.107)</td>
<td>0.108 (0.6%)</td>
<td>6 (100)</td>
<td>1.997 (0.107)</td>
<td>0.106 (0.0114)</td>
<td>0.0114 (94.6%)</td>
<td>935 (1)</td>
<td>0.0744 (1)</td>
<td>0.259 (0)</td>
</tr>
</tbody>
</table>

* Convergence is quite difficult, even though there are 50 studies, because $\tau_2^2$ is too small (due to the missing data) and with the $s_{ij}^2$s much larger, $\tau_1^2$ and $\hat{\rho}_B$ are very poorly defined and struggle to converge. Hence the simulations were not performed for $n = 10$. The 95% CI for $\hat{\beta}_j$ was calculated using a t-distribution with $(n_i-1)$ degrees of freedom, where $n_i$ is the number of studies for which outcome $j$ was available.

Additional results for $(\hat{\beta}_1 - \hat{\beta}_2)$ from (iii):

<table>
<thead>
<tr>
<th>n</th>
<th>Meta-analysis model</th>
<th>Mean$(\hat{\beta}_1 - \hat{\beta}_2)$</th>
<th>Mean s.e. of$(\hat{\beta}_1 - \hat{\beta}_2)$</th>
<th>Mean$\text{corr} (\hat{\beta}_1, \hat{\beta}_2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>URMA</td>
<td>-1.487 (0.130)</td>
<td>0.142</td>
<td>-</td>
</tr>
<tr>
<td>50</td>
<td>Model A</td>
<td>-1.665 (0.113)</td>
<td>0.0956</td>
<td>0.481 (0.155)</td>
</tr>
<tr>
<td>50</td>
<td>Model A-zero</td>
<td>-1.581 (0.115)</td>
<td>0.126</td>
<td>0.237 (0.112)</td>
</tr>
</tbody>
</table>

R.D. Riley, Ph.D. Thesis  Appendix C  399
Appendix C4: Simulation results to show Model A-zero is similar to Model A for the Nam et al. dataset

The dataset used by Nam et al. has the majority of the $s^2_i$'s small relative to the $\tau^2_j$ on average and this explains why their results were not influenced by the prior distributions for $\rho_w$ and $\rho_B$ [147]. Of course this is the same situation in which Model A-zero closely approximates Model A (see Section 8.1). To highlight this I performed classical simulations for Model A as detailed in Section 5.2.1 but used the $s^2_i$'s reported for the 8 complete-case studies from Nam et al., and assumed $\tau^2_1 = 0.036$, $\tau^2_2 = 0.019$, $\beta_1 = 0.23$ and $\beta_2 = 0.25$, all based on the reported values within their paper.

Furthermore, the true values for $\rho_w$ and $\rho_B$ were chosen to be 0.5. The mean value of the $s^2_i$'s was 0.019, the median value was 0.010, and 13 of the 16 (81%) $s^2_i$'s were smaller than the smallest $\tau^2_j$ (i.e. 0.019). The results show that the Model A-zero results are close to those from Model A. The only parameter estimate that is affected is $\hat{\rho}_B$, which is slightly higher in Model A-zero in order to compensate for the understated $\rho_w$. The results are shown below:

<table>
<thead>
<tr>
<th>Meta-analysis model</th>
<th>Converged out of 1000</th>
<th>Mean of $\hat{\beta}_1$ (s.e. of mean)</th>
<th>Mean s.e. of $\hat{\beta}_1$</th>
<th>MSE of $\hat{\beta}_1$ (no. of 95% CIs including $\beta_1$ (%))</th>
<th>Mean of $\hat{\beta}_2$ (s.e.)</th>
<th>MSE of $\hat{\beta}_2$ (no. of 95% CIs including $\beta_2$ (%))</th>
<th>Mean $\tau^2_1$ (no. of $\hat{\tau}^2_1 = 0$)</th>
<th>Mean $\tau^2_2$ (no. of $\hat{\tau}^2_2 = 0$)</th>
<th>Mean of $\hat{\rho}_B$</th>
<th>Mean of $\hat{\rho}_B - 1$ (%)</th>
<th>Mean of $\hat{\rho}_B = 1$ (%)</th>
<th>No. of $P_B = -1$ (%)</th>
<th>No. of $P_B = 1$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>URMA</td>
<td>999</td>
<td>0.230 (0.0755)</td>
<td>0.0750</td>
<td>0.0060 (93.3%)</td>
<td>0.249 (0.0614)</td>
<td>0.0574 (91.7%)</td>
<td>0.0367 (7)</td>
<td>0.0398 (19)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Model A</td>
<td>999</td>
<td>0.230 (0.0777)</td>
<td>0.0747</td>
<td>0.0060 (94.4%)</td>
<td>0.249 (0.0616)</td>
<td>0.0567 (92.0%)</td>
<td>0.0370 (0)</td>
<td>0.0201 (0)</td>
<td>0.436 (87)</td>
<td>0.436 (15.4%)</td>
<td>-</td>
<td>154</td>
<td>(8.7%)</td>
</tr>
<tr>
<td>Model A-zero</td>
<td>1000</td>
<td>0.230 (0.0778)</td>
<td>0.0751</td>
<td>0.0060 (94.4%)</td>
<td>0.250 (0.0620)</td>
<td>0.0570 (92.0%)</td>
<td>0.0375 (0)</td>
<td>0.0207 (0)</td>
<td>0.624 (49)</td>
<td>0.624 (33.7%)</td>
<td>-</td>
<td>337</td>
<td>(4.9%)</td>
</tr>
</tbody>
</table>

R.D. Riley, Ph.D. Thesis

Appendix C
Appendix C5: STATA code for maximum likelihood estimation of Model B

The syntax used was as follows, where ‘bvar’ and ‘wvar’ are the between and within-study variances respectively, and L1, L2 and L3 specify the contribution of each study to the total log-likelihood when the study has complete-case data (i.e. \( \tilde{Y}_{ij} \) and \( \tilde{Y}_{i2} \), only \( j = 1 \) data (i.e. just \( \tilde{Y}_{i1} \)) and only \( j = 2 \) data (i.e. just \( \tilde{Y}_{i2} \)) respectively.

program define ML
args lnL betal beta2 lnbvar1 lnbvar2 rho
tempvar bvar1 bvar2 v1 v2 cov L1 L2 L3
quietly {
    gen double `bvar1' = exp(`lnbvar1')
    gen double `bvar2' = exp(`lnbvar2')
    gen double `v1' = `bvar1' + 1 / `wvar1'
    gen double `v2' = `bvar2' + 1 / `wvar2'
    gen double `cov' = `rho' * sqrt(`v1' * `v2')
    gen double `L3' = log(`v1') + (y1-'betal')^2 / `v1'
    gen double `L2' = log(`v2') + (y2-'beta2')^2 / `v2'
    gen double `L1' = ((y1-'betal')^2 / `v1' /*
    -2 * `rho' * (y1-'betal') * (y2-'beta2') / sqrt(`v1' * `v2') /*
    + (y2-'beta2')^2 / `v2' / (1-`rho'^2) /*/
    + log(`v2'^2 * `v1'^2 * `cov'^2))
    replace `lnL' = -0.5 * `L1' if type == 1
    replace `lnL' = -0.5 * `L2' if type == 2
    replace `lnL' = -0.5 * `L3' if type == 3
}
end

ml model lf ML ( ) ( ) ( ) ( ), technique(nr)
ml maximize , difficult iter(200)
Appendix C6: Restricted maximum likelihood estimation of Model B in STATA

The syntax used was as follows, where ‘bvarj’ and ‘wvarj’ are the between- and within-study variances respectively for outcome $j = 1$ or $2$:

```stata
program define REML
args InL betal beta2 Inbvar1 Inbvar2 rho
tempvar bvar1 bvar2 v1 v2 cov L1 L2 L3 tmp det
quietly {
    gen double 'bvar1' = exp('Inbvar1')
    gen double 'bvar2' = exp('Inbvar2')
    gen double 'v1' = 'bvar1' + 1 / 'wvar1'
    gen double 'v2' = 'bvar2' + 1 / 'wvar2'
    gen double 'cov' = 'rho' * sqrt('v1' * 'v2')
    gen double L3' = log('v1') + (.y1-'beta1')^2 / 'v1'
    count if type==3
    local n3=r(N)
    gen double 'tmp' = 1 / 'v1'
    sum 'tmp' if type==3
    local xSx = r(sum)
    replace 'lnL' = -0.5 * (L3'+log('xSx')/n3') if type == 3
    gen double 'L2' = log('v2') + (.y2-'beta2')^2 / 'v2'
    count if type==2
    local n2=r(N)
    replace 'tmp' = 1 / 'v2'
    sum 'tmp' if type==2
    local xSx = r(sum)
    replace 'lnL' = -0.5 * (L2'+log('xSx')/n2') if type == 2
    gen double 'det' = 'v2' * 'v1' - 'cov'^2
    gen double 'L1' = ((.y1-'beta1')^2 / 'v1' -
        */-2 * 'rho' * (.y1-'beta1') * (.y2-'beta2') / sqrt('v1' * 'v2') /*
        */+(.y2-'beta2')^2 / 'v2' / (1-'rho'^2) /*
        */+ log('det')
    count if type==1
    local n1=r(N)
    replace 'tmp' = 'v2'/ 'det'
    sum 'tmp' if type==1
    local xSx11 = r(sum)
    replace 'lnL' = -0.5 * (L1'+log('xSx11'*xSx12'^2)/n1') if type == 1
}
end

ml model If REML () () () () (), technique(nr)
ml maximize , difficult iter(200)

N.B. The ‘replace ‘lnL’ = …’ lines show the addition of an $n$th of $\log(X'V^{-1}X)$ to each study’s contribution to the REML log-likelihood (see Section 8.3).
Appendix C7: Model B complete-case data simulation results for \( n = 5 \) studies

The Model B results are shown below for those simulations (\( nsim \)) that met the criteria specified in Section 8.3.3, with comparison to those results from Model A, Model A-zero and a URMA.

(i) \( n = 5, \rho_w = 0.8, \ \rho_b = 0.8, \ \tau_1^2 = \tau_2^2 = 0.25 \) (on average similar to the \( s_i^2 \) s)

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>Meta-analysis Model</th>
<th>nsim</th>
<th>Mean Estimate</th>
<th>No. 95% Cls including true value (%)</th>
<th>Mean-Square Error</th>
<th>Mean s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL URMA</td>
<td>604</td>
<td>-0.0144</td>
<td>1.961</td>
<td>596 (98.7%)</td>
<td>559 (92.5%)</td>
<td>0.0869</td>
</tr>
<tr>
<td>Model A</td>
<td>604</td>
<td>-0.0163</td>
<td>1.978</td>
<td>594 (98.3%)</td>
<td>559 (92.5%)</td>
<td>0.0854</td>
</tr>
<tr>
<td>Model B</td>
<td>604</td>
<td>-0.0119</td>
<td>1.960</td>
<td>594 (98.3%)</td>
<td>560 (92.7%)</td>
<td>0.0893</td>
</tr>
<tr>
<td>Model A-zero</td>
<td>604</td>
<td>-0.0141</td>
<td>1.978</td>
<td>599 (99.2%)</td>
<td>579 (95.5%)</td>
<td>0.0860</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ALL where ( \rho_B = 1 ) in Model A</th>
<th>Meta-analysis Model</th>
<th>nsim</th>
<th>Mean Estimate</th>
<th>No. 95% Cls including true value (%)</th>
<th>Mean-Square Error</th>
<th>Mean s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>URMA</td>
<td>233</td>
<td>-0.0176</td>
<td>1.966</td>
<td>230 (98.7%)</td>
<td>209 (89.7%)</td>
<td>0.0930</td>
</tr>
<tr>
<td>Model A</td>
<td>233</td>
<td>-0.0174</td>
<td>1.966</td>
<td>228 (97.9%)</td>
<td>209 (89.7%)</td>
<td>0.0889</td>
</tr>
<tr>
<td>Model B</td>
<td>233</td>
<td>-0.0071</td>
<td>1.967</td>
<td>228 (97.9%)</td>
<td>208 (89.3%)</td>
<td>0.0979</td>
</tr>
<tr>
<td>Model A-zero</td>
<td>233</td>
<td>-0.0169</td>
<td>1.962</td>
<td>231 (99.1%)</td>
<td>227 (97.4%)</td>
<td>0.0913</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ALL where ( \rho_B = -1 ) in Model A</th>
<th>Meta-analysis Model</th>
<th>nsim</th>
<th>Mean Estimate</th>
<th>No. 95% Cls including true value (%)</th>
<th>Mean-Square Error</th>
<th>Mean s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>URMA</td>
<td>58</td>
<td>-0.0263</td>
<td>1.971</td>
<td>57 (99.3%)</td>
<td>49 (84.5%)</td>
<td>0.0880</td>
</tr>
<tr>
<td>Model A</td>
<td>58</td>
<td>-0.0305</td>
<td>1.971</td>
<td>56 (96.6%)</td>
<td>49 (84.5%)</td>
<td>0.0870</td>
</tr>
<tr>
<td>Model B</td>
<td>58</td>
<td>-0.0293</td>
<td>1.974</td>
<td>56 (96.6%)</td>
<td>49 (84.5%)</td>
<td>0.0912</td>
</tr>
<tr>
<td>Model A-zero</td>
<td>58</td>
<td>-0.0258</td>
<td>1.974</td>
<td>57 (99.3%)</td>
<td>50 (86.2%)</td>
<td>0.0877</td>
</tr>
</tbody>
</table>

(ii) \( n = 5, \rho_w = 0.8, \ \rho_b = 0.8, \ \tau_1^2 = \tau_2^2 = 1.5 \) (on average larger than the \( s_i^2 \) s)

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>Meta-analysis Model</th>
<th>nsim</th>
<th>Mean Estimate</th>
<th>No. 95% Cls including true value (%)</th>
<th>Mean-Square Error</th>
<th>Mean s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL URMA</td>
<td>787</td>
<td>0.0173</td>
<td>2.012</td>
<td>743 (94.4%)</td>
<td>734 (93.3%)</td>
<td>0.340</td>
</tr>
<tr>
<td>Model A</td>
<td>787</td>
<td>0.0162</td>
<td>2.0126</td>
<td>744 (94.5%)</td>
<td>735 (93.4%)</td>
<td>0.340</td>
</tr>
<tr>
<td>Model B</td>
<td>787</td>
<td>0.0182</td>
<td>2.0119</td>
<td>748 (95.0%)</td>
<td>739 (93.9%)</td>
<td>0.345</td>
</tr>
<tr>
<td>Model A-zero</td>
<td>787</td>
<td>0.0169</td>
<td>2.013</td>
<td>747 (94.9%)</td>
<td>741 (94.2%)</td>
<td>0.340</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ALL where ( \rho_B = -1 ) in Model A</th>
<th>Meta-analysis Model</th>
<th>nsim</th>
<th>Mean Estimate</th>
<th>No. 95% Cls including true value (%)</th>
<th>Mean-Square Error</th>
<th>Mean s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>URMA</td>
<td>67</td>
<td>0.0166</td>
<td>1.933</td>
<td>58 (86.6%)</td>
<td>54 (80.6%)</td>
<td>0.308</td>
</tr>
<tr>
<td>Model A</td>
<td>67</td>
<td>0.00373</td>
<td>1.932</td>
<td>57 (85.1%)</td>
<td>54 (80.6%)</td>
<td>0.310</td>
</tr>
<tr>
<td>Model B</td>
<td>67</td>
<td>-0.0034</td>
<td>1.926</td>
<td>57 (85.1%)</td>
<td>54 (80.6%)</td>
<td>0.330</td>
</tr>
<tr>
<td>Model A-zero</td>
<td>67</td>
<td>0.0157</td>
<td>1.937</td>
<td>59 (88.1%)</td>
<td>57 (85.1%)</td>
<td>0.310</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ALL where ( \rho_B = 1 ) in Model A</th>
<th>Meta-analysis Model</th>
<th>nsim</th>
<th>Mean Estimate</th>
<th>No. 95% Cls including true value (%)</th>
<th>Mean-Square Error</th>
<th>Mean s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>URMA</td>
<td>16</td>
<td>0.0784</td>
<td>2.154</td>
<td>11 (68.8%)</td>
<td>10 (62.5%)</td>
<td>0.422</td>
</tr>
<tr>
<td>Model A</td>
<td>16</td>
<td>0.0963</td>
<td>2.148</td>
<td>11 (68.8%)</td>
<td>10 (62.5%)</td>
<td>0.431</td>
</tr>
<tr>
<td>Model B</td>
<td>16</td>
<td>0.0810</td>
<td>2.145</td>
<td>11 (68.8%)</td>
<td>11 (68.8%)</td>
<td>0.442</td>
</tr>
<tr>
<td>Model A-zero</td>
<td>16</td>
<td>0.0774</td>
<td>2.141</td>
<td>12 (75.0%)</td>
<td>10 (62.5%)</td>
<td>0.426</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ALL where ( \rho_B = -1 ) in Model A</th>
<th>Meta-analysis Model</th>
<th>nsim</th>
<th>Mean Estimate</th>
<th>No. 95% Cls including true value (%)</th>
<th>Mean-Square Error</th>
<th>Mean s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>URMA</td>
<td>704</td>
<td>0.0160</td>
<td>2.0158</td>
<td>674 (95.7%)</td>
<td>669 (95.0%)</td>
<td>0.342</td>
</tr>
<tr>
<td>Model A</td>
<td>704</td>
<td>0.0155</td>
<td>2.017</td>
<td>676 (96.0%)</td>
<td>671 (95.3%)</td>
<td>0.341</td>
</tr>
<tr>
<td>Model B</td>
<td>704</td>
<td>0.0188</td>
<td>2.017</td>
<td>680 (96.6%)</td>
<td>674 (95.7%)</td>
<td>0.344</td>
</tr>
<tr>
<td>Model A-zero</td>
<td>704</td>
<td>0.0155</td>
<td>2.0176</td>
<td>676 (96.0%)</td>
<td>674 (95.7%)</td>
<td>0.341</td>
</tr>
</tbody>
</table>
Appendix C8: Model A versus Model B simulation results for complete-case data

(a) Figures showing the Model A versus Model B results from Appendix C7(i) for \( n = 5, \rho_w = 0.8, \rho_B = 0.8, \tau^2_1 = \tau^2_2 = 0.25 \) (blue line shows line of equality, dotted line shows linear regression line through the points)

(i) Where \( \hat{\rho}_B = 1 \) in Model A and \( \hat{\rho} \) was between -0.95 and 0.95 in Model B:

(ii) Where \( \hat{\rho}_B = -1 \) in Model A and \( \hat{\rho} \) was between -0.95 and 0.95 in Model B:
(b) Figures showing Model A versus Model B for those results in Appendix C7(ii) for \( n = 5 \), \( \rho_w = 0.8 \), \( \rho_B = 0.8 \), \( t_1^2 = t_2^2 = 1.5 \) (blue line shows line of equality, dotted line shows linear regression line through the points)

(i) Where \( \hat{\rho}_B \) does not equal 1 or -1 in Model A and \( \hat{\rho} \) was between -0.95 and 0.95 in Model B:

(ii) Where \( \hat{\rho}_B = 1 \) in Model A and \( \hat{\rho} \) was between -0.95 and 0.95 in Model B:

(iii) Where \( \hat{\rho}_B = -1 \) in Model A and \( \hat{\rho} \) was between -0.95 and 0.95 in Model B:

R.D. Riley, Ph.D. Thesis  Appendix C 405
Appendix C9: Missing completely at random (MCAR) simulations for Model B

The Model B results are shown below for those simulations (nsim) that met the criteria specified in Section 8.3.3, with comparison to those results from Model A, Model A-zero and a URMA.

(i) $n = 50$ for $j = 1$, $n = 25$ for $j = 2$: $\rho_w = 0.8$, $\rho_B = 0.8$, $\tau_1^2 = \tau_2^2 = 0.25$ (on average similar in size to the $s_j^2$'s)

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>Meta-analysis Model</th>
<th>nsim</th>
<th>Mean Estimate</th>
<th>No. 95% CIs including true value (%)</th>
<th>Mean-Square Error</th>
<th>Mean s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>URMA</td>
<td>988</td>
<td>$\hat{\beta}_1$</td>
<td>934 (94.5%) 937 (94.8%)</td>
<td>0.0103</td>
<td>0.0218</td>
</tr>
<tr>
<td></td>
<td>Model A</td>
<td>988</td>
<td>$\hat{\beta}_2$</td>
<td>934 (94.5%) 940 (95.1%)</td>
<td>0.00959</td>
<td>0.0151</td>
</tr>
<tr>
<td></td>
<td>Model B</td>
<td>988</td>
<td>$\hat{\beta}_3$</td>
<td>940 (95.1%) 945 (95.6%)</td>
<td>0.0101</td>
<td>0.0165</td>
</tr>
<tr>
<td></td>
<td>Model A-zero</td>
<td>988</td>
<td>$\hat{\beta}_4$</td>
<td>940 (95.1%) 945 (95.6%)</td>
<td>0.0101</td>
<td>0.0165</td>
</tr>
</tbody>
</table>

(ii) $n = 10$ for $j = 1$, $n = 5$ for $j = 2$: $\rho_w = 0.8$, $\rho_B = 0.8$, $\tau_1^2 = \tau_2^2 = 0.25$ (on average similar in size to the $s_j^2$'s)

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>Meta-analysis Model</th>
<th>nsim</th>
<th>Mean Estimate</th>
<th>No. 95% CIs including true value (%)</th>
<th>Mean-Square Error</th>
<th>Mean s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>URMA</td>
<td>680</td>
<td>$\hat{\beta}_1$</td>
<td>633 (93.1%) 623 (91.6%)</td>
<td>0.0524</td>
<td>0.1000</td>
</tr>
<tr>
<td></td>
<td>Model A</td>
<td>680</td>
<td>$\hat{\beta}_2$</td>
<td>623 (91.6%) 628 (92.4%)</td>
<td>0.0524</td>
<td>0.0761</td>
</tr>
<tr>
<td></td>
<td>Model B</td>
<td>680</td>
<td>$\hat{\beta}_3$</td>
<td>642 (94.4%) 646 (95.1%)</td>
<td>0.0546</td>
<td>0.0768</td>
</tr>
<tr>
<td></td>
<td>Model A-zero</td>
<td>680</td>
<td>$\hat{\beta}_4$</td>
<td>663 (97.5%) 666 (97.8%)</td>
<td>0.0521</td>
<td>0.253</td>
</tr>
<tr>
<td>ALL where $\hat{\rho}_B \neq 1$</td>
<td>URMA</td>
<td>90</td>
<td>$\hat{\beta}_1$</td>
<td>85 (94.4%) 82 (91.1%)</td>
<td>0.0117</td>
<td>0.0276</td>
</tr>
<tr>
<td></td>
<td>Model A</td>
<td>90</td>
<td>$\hat{\beta}_2$</td>
<td>85 (94.4%) 178 (198.6%)</td>
<td>0.0118</td>
<td>0.0182</td>
</tr>
<tr>
<td></td>
<td>Model B</td>
<td>90</td>
<td>$\hat{\beta}_3$</td>
<td>80 (80.9%) 80 (80.9%)</td>
<td>0.0139</td>
<td>0.0187</td>
</tr>
<tr>
<td></td>
<td>Model A-zero</td>
<td>90</td>
<td>$\hat{\beta}_4$</td>
<td>86 (95.6%) 84 (93.3%)</td>
<td>0.0119</td>
<td>0.0194</td>
</tr>
</tbody>
</table>

R.D. Riley, Ph.D. Thesis Appendix C 406
Appendix C10: Model A versus Model B simulation results for data MCAR

(a) Figures showing Model A versus Model B for the MCAR simulation results presented in Appendix C9(i) where \( n = 50 \) for \( j = 1 \), \( n = 25 \) for \( j = 2 \); \( \rho_w = 0.8 \), \( \rho_B = 0.8 \), \( \tau_1^2 = \tau_2^2 = 0.25 \) (blue line shows line of equality, dotted line shows linear regression line through the points)

(i) \( \hat{\rho}_B \) does not equal 1 or -1 in Model A and \( \hat{\rho} \) was between -0.95 and 0.95 in Model B:

(ii) \( \hat{\rho}_B = 1 \) in Model A and \( \hat{\rho} \) was between -0.95 and 0.95 in Model B:
(b) Figures showing Model A versus Model B for the MCAR simulation results presented in Appendix C9(ii) where $n = 10$ for $j = 1$, $n = 5$ for $j = 2$; $ho_W = 0.8$, $\rho_B = 0.8$, $\tau_1^2 = \tau_2^2 = 0.25$ (blue line shows line of equality, dotted line shows linear regression line through the points).

(i) $\hat{\rho}_B$ does not equal 1 or −1 in Model A and $\hat{\rho}$ was between −0.95 and 0.95 in Model B:

(ii) $\hat{\rho}_B = 1$ in Model A and $\hat{\rho}$ was between −0.95 and 0.95 in Model B:

(iii) $\hat{\rho}_B = -1$ Model A and $\hat{\rho}$ was between −0.95 and 0.95 in Model B:
Appendix C11: WinBUGS version 1.3 syntax to fit Model B-Bayes

The syntax is shown below and relates to equation (8.10) in Section 8.6, and the data used in this code is the original Berkey data (see Table 4.1).

Model
{
for(i in 1:n)
{

covariance[i,1,1]<-variance1[i]+bsvar[1]
covariance[i,2,2]<-variance2[i]+bsvar[2]
covariance[i,1,2]<rho*sqrt((bsvar[1]+variance1[i])*(bsvar[2]+variance2[i]))
covariance[i,2,1]<covariance[i,1,2]
for (k in 1:2) {
    for (j in 1:2) {
        prec[i,k,j] <- inverse(covariance[ i , ], k, j)
    }
}
}
for(i in 1:n)
{
    surv[i,1:2] ~ dmnorm(beta[i, 1:2],prec[i,1:2,1:2])
    beta[i,1]<beta1
    beta[i,2]<beta2
}
bsvarinv[1]-dgamma(0.001, 0.001)
bsvarinv[2]-dgamma(0.001,0.001)
beta1 ~ dnorm(0.0,1.0E-6)
beta2 ~ dnorm(0.0,1.0E-6)
rho~dunif(-1,1)
}

Data
list(n=5, variance1=c(0.0075, 0.0057, 0.0021, 0.0029, 0.0148),
    variance2=c(0.0077, 0.0008, 0.0014, 0.0015, 0.0304),
    surv=structure(Data=c(0.47, -0.32, 0.2, -0.6, 0.4, -0.12, 0.26, -0.31, 0.56 , -0.39), .Dim=c(5,2)))

Inits
list(beta1=0, beta2=0, bsvarinv=c(0.1,0.1), rho=0.5)
Appendix D1: Publication in the Journal of Clinical Epidemiology arising from the sensitivity analyses for dissemination bias that were developed in Chapter 9


RR oversaw the entire review upon which the MYCN estimates came from (see Appendix A3 and Appendix A4). RR conceived the study and performed all the meta-analyses and all the other statistical analyses, with advice at various stages provided by the other authors. RR developed the ‘three-step approach’ (see Section 9.3.1), again with suggestions from the other authors throughout. RR wrote the paper, with comments provided on various draft versions by the other three authors.
SPECIAL NOTE

ITEM SCANNED AS SUPPLIED
PAGINATION IS AS SEEN
Sensitivity analyses allowed more appropriate and reliable meta-analysis conclusions for multiple outcomes when missing data was present

Richard D. Riley*, Alex J. Sutton, Keith R. Abrams, Paul C. Lambert

Centre for Biostatistics and Genetic Epidemiology, Department of Health Sciences, University of Leicester,
Princess Road West, Leicester LE1 6TF, United Kingdom

Accepted 10 January 2004

Abstract

Objective: A major problem for meta-analysis of multiple outcomes is the unavailability of some estimates from published and unpublished studies. Dissemination bias, in how and what outcomes are reported or published, may be causing this incompleteness. This article illustrates these problems and presents possible sensitivity analyses to allow the most reliable conclusions.

Study Design and Setting: In a systematic review of prognostic marker MYC-N in neuroblastoma, meta-analysis for overall survival (OS) and disease-free survival (DFS) was of interest. Only 17 published studies enabled extraction of both outcome estimates, 25 enabled only DFS, 39 enabled only OS, and 70 enabled neither outcome. Unidentified unpublished studies may also exist. We assessed the robustness of the pooled estimates to the problem of missing information. Because OS and DFS estimates seemed to be related, we used the known outcome estimates to predict estimates known to be missing, and combined this approach with existing methods for assessing dissemination bias.

Results: The results of the sensitivity analyses suggested that the original meta-analysis results were likely to be an overestimate of the true OS and DFS effect-sizes but strengthened the belief that MYC-N is a potentially important prognostic marker in neuroblastoma.

Conclusion: Sensitivity analyses in meta-analysis allow more appropriate and reliable conclusions when problems such as unavailable estimates and dissemination bias are present. © 2004 Elsevier Inc. All rights reserved.

Keywords: Meta-analysis; Publication bias; Prognosis; Multiple outcomes; Sensitivity analysis

1. Introduction

Systematic reviews and meta-analyses are potentially important for synthesizing current evidence and providing evidence-based conclusions [1]. However, reliable and clinically useful quantitative syntheses are generally difficult to perform because individual studies are often of poor quality and susceptible to several sources of bias that can invalidate their results [2]. Two major problems that can affect the results of meta-analyses are poor reporting of primary studies and dissemination bias, which refers to the various ways the reporting and publishing of an individual study can be influenced by the nature and direction of its results [3,4]. A full assessment of the potential impact of these problems is important when performing and presenting meta-analyses so that the most appropriate and reliable conclusions can be made. This is especially important for meta-analysis of observational studies (e.g., prognostic studies), where these problems are likely to be much greater than in randomized clinical trials [2].

Publication bias is the most recognized form of dissemination bias, existing because studies that do not generate statistically significant or clinically valuable findings are less likely to be published. Other types of dissemination bias include outcome reporting bias, subgroup reporting bias, time lag bias, language bias, citation bias, and duplicate (multiple) publication bias [5]. Numerous methods have been developed to help assess dissemination bias, the majority of which are motivated by the so-called funnel plot of individual study estimates against their standard error or another measure of uncertainty [6]. For example, Duval and Tweedie [7] present a "Trim and Fill" method of testing and adjusting for dissemination bias based on funnel plot asymmetry, whereas Egger et al. [8] suggest a hypothesis test to help detect whether
dissemination bias is likely to exist. Poor statistical reporting is another related problem for meta-analysts because it restricts evidence-based research by limiting the extraction of desired estimates and may lead to a set of extracted estimates that do not reflect the overall evidence base [9]. It may contribute to the overall dissemination bias if standards are worse for the less significant or less interesting results.

The first aim of this article is to illustrate how these problems easily affect meta-analysis of multiple outcomes, where the meta-analyst desires an estimate for more than one outcome from each identified study, and to encourage researchers to consider whether similar problems are affecting their own data sets and reviews. There are several possibilities for a given outcome in a given published study:

(i) It is mentioned, and the results are reported in sufficient detail to allow extraction of the desired estimate.
(ii) It is mentioned, but the results are not reported in sufficient detail to allow extraction of the desired estimate (e.g., a P value may be given but no estimate of effect size).
(iii) It is mentioned, but the results are not reported [10].
(iv) It is not mentioned.

Only those in (i) can be included in the meta-analysis; those in (ii) and (iii) cannot because of the incomplete reporting. Dissemination bias is a concern for those in (ii), (iii), and (iv) because the reporting of an outcome may be related to the significance or direction of its result; similarly, the possibility of unidentified unpublished studies needs to be considered. In this situation, the meta-analyst is able to obtain estimates for a complete set of outcomes in some studies, for a partial set of outcomes in others, and for none of the outcomes in the remaining studies. This leads to meta-analysis results more vulnerable to bias than when all outcome estimates are available from all the studies.

Guidelines to improve the design and reporting of primary observational studies have been advocated and should, in the long-term, help reduce those and other related problems to facilitate evidence-based practice [9,11]. Until such standards are established, and until individual patient data (IPD) is generally made available, those contemplating meta-analysis have to consider one or more of the following: (1) not performing meta-analysis because coherent and worthwhile pooled results will not be possible, (2) contacting authors for the estimates or IPD needed, and (3) undertaking sensitivity analyses alongside meta-analysis to assess the robustness of the pooled results to the problems. When meta-analysis is performed, it is likely that not all the available information will be obtained even after considering (2), and thus sensitivity analyses are important for the majority of meta-analyses.

The second aim of this paper is to consider possible sensitivity analyses given unavailable outcome estimates and the threat of dissemination bias. For this we use a recent systematic review of prognostic marker, MYC-N, in neuroblastoma where meta-analysis for overall survival (OS) and disease-free survival (DFS) was of interest but was limited by unavailability of these outcomes in some studies [3]. For our sensitivity assessments, we focus on methods for imputing missing outcome estimates, with the aim not to obtain "adjusted" pooled estimates as the corrected answer but rather to help assess the robustness of the original results to the problems [12].

2. Partially and completely unavailable outcomes: the example of prognostic marker MYC-N in neuroblastoma

Prognostic markers help to stratify patients for treatment by identifying patients with different risks of outcome and are important tools in the management of many diseases. Systematic reviews and meta-analysis can help clarify the most important prognostic markers for a disease area and thus facilitate the most appropriate treatment strategies. We have recently undertaken a systematic review of tumor markers in neuroblastoma, the most common extracranial solid tumor of childhood (for full details of the review, including inclusion/exclusion criteria, see [3]). During the review we encountered problems in performing reliable and clinically useful meta-analyses; in particular, we found the quantitative synthesis of prognostic marker results to be limited by poor reporting of primary studies and the likely presence of dissemination bias [9].

For example, for MYC-N, the most commonly studied marker, we were interested in its relationship with OS and DFS outcomes and wished to perform meta-analysis of appropriate estimates for each outcome. We tried to obtain an estimate of the log-hazard ratio [log(HR)] and its standard error from each of 194 reports across 151 published studies where summary statistics or IPD for OS or DFS were presented for this marker. Despite using an extensive range of direct and indirect methods, based on those described by Parmar et al. [13], we managed to extract a total of 107 estimates (55.2%) (Fig. 1). Ninety-four estimates were originally extracted during our review [3], but we have since extracted, where possible, estimates from all studies with a sample size <25 (some of which were originally not located at because of time constraints during the original review) and obtained 13 more; hence, in total we have 107. All estimates were based on dichotomized MYC-N levels using a cut-off level, although the choice of cut-off was inconsistent across studies and was often not specified in the published paper. The problem of different cut-off levels and its impact on meta-analysis is considered in more detail elsewhere [9,14].

Because it was felt appropriate to use only one estimate per outcome per study in the meta-analyses, 98 of the 107 estimates obtained overall for MYC-N were considered. Of these, 42 estimates were for DFS and 56 were for OS (Fig. 1, Table 1). One consequence of the limited reporting was that OS and DFS estimates were not always available from an
Identifying the literature
A comprehensive search strategy was used to identify the relevant literature [3].

Screening the results
Papers were screened to identify the number of overall survival (OS) and disease-free survival (DFS) reports where summary statistics or individual patient data (IPD) were presented relating this marker to patient outcome (N.B. individual studies sometimes had more than one report, e.g. for both OS & DFS, different cut-off levels etc.; hence, number of overall reports exceeds overall number of studies)

Data Extraction
From each OS and DFS report, an estimate of the loge(hazard ratio) and its standard error was sought using an extensive extraction procedure using the methods of Parmar and colleagues [13] and IPD. For further details see [3].

Number of estimates suitable for meta-analysis in each outcome
For meta-analysis, OS and DFS outcomes were treated separately. Some estimates were excluded at this stage as only one estimate was used per outcome per study*.

From which published studies do these estimates come from?
Of the 151 studies overall, 70 did not allow either an OS or DFS estimate to be extracted; of the other 81 studies:

* Only one estimate per outcome per study was used for the meta-analysis. If two or more estimates were available for an outcome from a study (usually because it presented results for a range of cut-off levels or different subgroups) then the one relating to the biggest sample size and most similar cut-off level to other papers was chosen; e.g. DFS had 48 estimates extracted, but 3 studies provided their estimates twice for the same data, but using different cut-offs (hence 3 estimates were excluded here), and another 3 studies provided estimates for two different stage of disease patients (hence another 3 estimates were excluded here).

Fig. 1. Description of how the overall and disease-free survival MYC-N estimates of the loge(HR) and its standard error were obtained.

individual study; there were three different studies to be considered from our review of MYC-N (Table 1):

1. Studies with a complete set of extracted outcome estimates (i.e., those from which OS and DFS estimates were extracted) (17 studies).
2. Studies with a partial set of extracted outcome estimates (i.e., those from which only OS estimates were extracted) (39 studies) and those from which only DFS estimates were extracted (25 studies).
3. Studies with no outcome estimates extracted (i.e., known studies that reported prognostic results or IPD but not sufficiently to allow OS or DFS estimates to be extracted) (70 studies) and unidentified studies that have not been published (unknown number of studies).

Apart from unintentionally poor statistical reporting, there could be a variety of reasons why only an OS or only a DFS estimate was available from those studies in item 2.
### Table 1

<table>
<thead>
<tr>
<th>Study ID</th>
<th>DFS (log(HR)) (SE)</th>
<th>OS (log(HR)) (SE)</th>
<th>Study ID</th>
<th>DFS (log(HR)) (SE)</th>
<th>OS (log(HR)) (SE)</th>
<th>DFS predicted* (log(HR)) (SE)</th>
<th>OS predicted* (log(HR)) (SE)</th>
<th>Study ID</th>
<th>DFS (log(HR)) (SE)</th>
<th>OS (log(HR)) (SE)</th>
<th>DFS predicted* (log(HR)) (SE)</th>
<th>OS predicted* (log(HR)) (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.11 (0.67)</td>
<td>0.14 (0.81)</td>
<td>18</td>
<td>0.25 (0.29)</td>
<td>NA</td>
<td>0.55 (0.38)</td>
<td>43</td>
<td>NA</td>
<td>0.84 (0.85)</td>
<td>0.69 (0.76)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.30 (0.26)</td>
<td>0.43 (0.81)</td>
<td>19</td>
<td>0.29 (0.59)</td>
<td>NA</td>
<td>0.59 (0.66)</td>
<td>44</td>
<td>NA</td>
<td>0.05 (0.40)</td>
<td>0.08 (0.37)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.41 (0.82)</td>
<td>0.67 (0.29)</td>
<td>20</td>
<td>0.52 (0.41)</td>
<td>NA</td>
<td>0.80 (0.50)</td>
<td>45</td>
<td>NA</td>
<td>0.73 (0.71)</td>
<td>0.68 (0.64)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.47 (0.53)</td>
<td>0.70 (0.56)</td>
<td>21</td>
<td>0.55 (0.38)</td>
<td>NA</td>
<td>0.83 (0.47)</td>
<td>46</td>
<td>NA</td>
<td>0.76 (0.20)</td>
<td>0.70 (0.19)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.76 (0.49)</td>
<td>0.71 (0.63)</td>
<td>22</td>
<td>0.84 (0.26)</td>
<td>NA</td>
<td>1.10 (0.35)</td>
<td>47</td>
<td>NA</td>
<td>0.91 (0.66)</td>
<td>0.83 (0.59)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1.06 (0.54)</td>
<td>1.32 (0.51)</td>
<td>23</td>
<td>0.93 (0.32)</td>
<td>NA</td>
<td>1.18 (0.41)</td>
<td>48</td>
<td>NA</td>
<td>0.93 (0.27)</td>
<td>0.95 (0.26)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1.46 (0.41)</td>
<td>1.38 (0.37)</td>
<td>24</td>
<td>1.18 (0.57)</td>
<td>NA</td>
<td>1.41 (0.65)</td>
<td>49</td>
<td>NA</td>
<td>0.96 (0.47)</td>
<td>0.87 (0.43)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1.64 (0.64)</td>
<td>1.51 (0.48)</td>
<td>25</td>
<td>1.34 (0.51)</td>
<td>NA</td>
<td>1.56 (0.58)</td>
<td>50</td>
<td>NA</td>
<td>1.05 (0.86)</td>
<td>0.95 (0.77)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>1.64 (0.64)</td>
<td>1.54 (0.52)</td>
<td>26</td>
<td>1.43 (0.37)</td>
<td>NA</td>
<td>1.64 (0.45)</td>
<td>51</td>
<td>NA</td>
<td>1.16 (1.18)</td>
<td>1.04 (1.05)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1.64 (0.51)</td>
<td>1.82 (0.71)</td>
<td>27</td>
<td>1.44 (1.17)</td>
<td>NA</td>
<td>1.64 (1.21)</td>
<td>52</td>
<td>NA</td>
<td>1.22 (0.22)</td>
<td>1.10 (0.21)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>1.70 (0.39)</td>
<td>1.83 (0.47)</td>
<td>28</td>
<td>1.45 (0.57)</td>
<td>NA</td>
<td>1.66 (0.65)</td>
<td>53</td>
<td>NA</td>
<td>1.26 (0.49)</td>
<td>1.13 (0.45)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>1.85 (0.66)</td>
<td>2.08 (0.67)</td>
<td>29</td>
<td>1.52 (0.35)</td>
<td>NA</td>
<td>1.73 (0.44)</td>
<td>54</td>
<td>NA</td>
<td>1.26 (0.38)</td>
<td>1.13 (0.36)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>1.90 (0.46)</td>
<td>2.59 (1.04)</td>
<td>30</td>
<td>1.60 (0.49)</td>
<td>NA</td>
<td>1.80 (0.57)</td>
<td>55</td>
<td>NA</td>
<td>1.27 (1.28)</td>
<td>1.14 (1.14)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>1.90 (0.88)</td>
<td>2.75 (1.10)</td>
<td>31</td>
<td>1.62 (0.42)</td>
<td>NA</td>
<td>1.81 (0.50)</td>
<td>56</td>
<td>NA</td>
<td>1.31 (0.82)</td>
<td>1.17 (0.74)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>2.19 (0.42)</td>
<td>2.90 (1.10)</td>
<td>32</td>
<td>1.77 (0.46)</td>
<td>NA</td>
<td>1.96 (0.54)</td>
<td>57</td>
<td>NA</td>
<td>1.52 (0.46)</td>
<td>1.36 (0.43)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>2.95 (1.08)</td>
<td>2.99 (0.51)</td>
<td>33</td>
<td>1.90 (0.58)</td>
<td>NA</td>
<td>2.08 (0.65)</td>
<td>58</td>
<td>NA</td>
<td>1.54 (0.55)</td>
<td>1.38 (0.50)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>5.70 (1.73)</td>
<td>5.70 (1.73)</td>
<td>34</td>
<td>1.92 (0.34)</td>
<td>NA</td>
<td>2.09 (0.43)</td>
<td>59</td>
<td>NA</td>
<td>1.55 (0.70)</td>
<td>1.39 (0.63)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>2.04 (0.62)</td>
<td>NA</td>
<td>36</td>
<td>2.19 (0.35)</td>
<td>NA</td>
<td>2.34 (0.44)</td>
<td>61</td>
<td>NA</td>
<td>1.67 (1.13)</td>
<td>1.49 (1.01)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>2.37 (1.00)</td>
<td>NA</td>
<td>38</td>
<td>2.39 (0.73)</td>
<td>NA</td>
<td>2.51 (1.05)</td>
<td>62</td>
<td>NA</td>
<td>1.72 (0.67)</td>
<td>1.53 (0.60)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>2.50 (0.76)</td>
<td>NA</td>
<td>40</td>
<td>2.56 (0.55)</td>
<td>NA</td>
<td>2.53 (0.79)</td>
<td>63</td>
<td>NA</td>
<td>1.74 (0.45)</td>
<td>1.55 (0.41)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>2.98 (0.58)</td>
<td>NA</td>
<td>42</td>
<td>3.29 (0.50)</td>
<td>NA</td>
<td>3.07 (0.65)</td>
<td>66</td>
<td>NA</td>
<td>1.87 (0.57)</td>
<td>1.66 (0.52)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Abbreviations:** DFS: disease-free survival; OS: overall survival; HR: hazard ratio; SE: standard error; NA: not available.

* Predicted using Approach 3.

Some authors may believe that OS is misleading, especially when there are other possible causes of death, whereas some may not report DFS because of the subjectivity in determining recurrence (and time of recurrence). Furthermore, because OS and DFS results can often be similar, reporting both may be deemed unnecessary. Of most concern is the possibility that dissemination bias is the reason behind studies in item 2 and that the outcome estimates missing were less significant or less interesting than the other outcome.

Dissemination bias may also have influenced those studies in item 3 because the decision not to publish the study or not to report results in sufficient detail may have been influenced by the nonsignificant or noninteresting findings, although the effect might not have been overwhelming. Hence, meta-analysis for OS and DFS was clouded by the possibility that any pooled results would be sensitive to the problems affecting our available data set. In particular, some outcome estimates extracted might not form an accurate representation of the overall evidence-base, and thus any subsequent meta-analysis results and conclusions about MYC-N might be misleading.

### 2.1 Meta-analysis despite these problems

For clinicians who treat patients with neuroblastoma, there is uncertainty as to which are the most appropriate prognostic markers to use, and this is reflected in the available data set.
large number of prognostic marker studies published [3]. Hence, despite the concerns, we considered meta-analysis important for our systematic review to help priorities those markers for future research.

3. Possible approaches for meta-analysis in this situation

In our original report we performed meta-analysis for MYC-N using the estimates known in each outcome and gave an assessment of dissemination bias [3]. We now consider meta-analysis and sensitivity analyses considering the problem of partial sets of outcome estimates alongside the threat of dissemination bias. We consider four possible approaches: meta-analysis without sensitivity-analysis of the problems (Approach 1), meta-analysis with sensitivity analysis for just dissemination bias (Approach 2) or for just the partial sets of available outcomes (Approach 3), and meta-analysis with a combined sensitivity analysis for both these problems (Approach 4).

The sensitivity analyses in Approaches 2, 3, and 4 are only assessing the robustness of the results in Approach 1 and are not trying to obtain efficiency gains in our estimation. Hence, although we calculate “adjusted” estimates in Approaches 2, 3, and 4, they are used only to help assess how the standard meta-analysis results might be affected by the missing information described in Section 2.

3.1. Approach 1: random-effects meta-analysis for each outcome

Random-effect meta-analysis was chosen because there was considerable evidence of between-study heterogeneity in the estimates for OS and DFS (P < 0.001). This was not surprising given the heterogeneity in different cut-offs, method of measurement, stage groups, age of patients, and other factors observed across studies. High levels of MYC-N were associated with a statistically significant increased risk of death (56 OS studies: pooled log(HR) = 1.63, 95% CI 1.39–1.87, P < 0.001) and risk of death or recurrence of disease (42 DFS studies: pooled log(HR) = 1.48, 95% CI 1.23–1.74, P < 0.001) (Table 2).

These pooled results suggest that MYC-N is a potentially important prognostic marker for neuroblastoma. However, such conclusions based on these results are open to criticism given unavailable outcome estimates and the underlying threat of dissemination bias.

3.2. Approach 2: sensitivity analysis for the impact of dissemination bias

To assess dissemination bias and its potential impact on these meta-analysis results, we used the “Trim and Fill” method [7] alongside a visual inspection of the funnel plot and the hypothesis test proposed by Egger et al. [8]. We also considered whether study characteristics, and not dissemination bias, could explain the funnel plot asymmetry by looking at subgroups of patients [15].

3.2.1. Visual inspection of the funnel plots

Funnel plots of the DFS and OS estimates against their standard error are shown in Figs. 2A and 2B. The assumption is that these plots should form a funnel shape if there is no dissemination bias present because estimates from smaller studies will be more widely spread about the mean effect due to larger standard errors. However, the plots for DFS and OS were not indicative of a funnel-like shape, and asymmetry was apparent, with a gap in the bottom right of each plot. Hence, it seemed that some studies with less positive results than those obtained were potentially missing from our analyses (Fig. 2).

3.2.2. Egger’s test for the presence of dissemination bias

This method uses a linear regression approach in which the standard normal deviates (study effect size estimate b[9]/standard error of estimate s[1]) are regressed against precision (1/s). This approach corresponds to a weighted regression of effect size [e.g., log(HR)] on standard error (θ = α + βs), where the weights are inversely proportional to the variance (s^2) of the effect size. The degree of asymmetry in the funnel plot, and therefore the potential for dissemination bias, is indicated by the magnitude and significance of the intercept

Table 2
Overall and disease-free survival meta-analysis and sensitivity analysis results for Approaches 1–4

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Approach</th>
<th>Description of the approach</th>
<th>Number of known estimates</th>
<th>Number of predicted estimates</th>
<th>Number of till studies</th>
<th>Random-effects meta-analysis for log(HR) 95% CI</th>
<th>Between-study variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>DFS</td>
<td>1</td>
<td>Just known estimates</td>
<td>42</td>
<td>—</td>
<td>8</td>
<td>1.48 (1.23–1.74)</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Trim and Fill method for Approach 1</td>
<td>42</td>
<td>—</td>
<td>8</td>
<td>1.27 (1.01–1.53)</td>
<td>0.57</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Known plus predicted estimates</td>
<td>42</td>
<td>39</td>
<td>19</td>
<td>1.49 (1.31–1.66)</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Trim and Fill method for Approach 3</td>
<td>42</td>
<td>39</td>
<td>19</td>
<td>1.22 (1.04–1.40)</td>
<td>0.52</td>
</tr>
<tr>
<td>OS</td>
<td>1</td>
<td>Just known estimates</td>
<td>56</td>
<td>—</td>
<td>12</td>
<td>1.63 (1.40–1.87)</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Trim and Fill method for Approach 1</td>
<td>56</td>
<td>—</td>
<td>12</td>
<td>1.37 (1.13–1.62)</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Known plus predicted estimates</td>
<td>56</td>
<td>25</td>
<td>18</td>
<td>1.67 (1.48–1.86)</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Trim and Fill method for Approach 3</td>
<td>56</td>
<td>25</td>
<td>18</td>
<td>1.41 (1.21–1.61)</td>
<td>0.56</td>
</tr>
</tbody>
</table>

Abbreviation: HR, hazard ratio

* predicted using Approach 3.

Note: Approaches 2, 3, and 4 are sensitivity analyses to assess the robustness of results in Approach 1.
coefficient $\alpha$. Applying this test to the estimates in each outcome produced statistically significant evidence of asymmetry for DFS ($P = .003$) but not for OS ($P = .45$) (Table 2). When the extreme study in the top right of the funnel plots in Fig. 2 was removed, Egger’s test for DFS and OS produced similar $P$ values to before.

3.2.3. The Trim and Fill method

The Trim and Fill method imputes studies considered missing based on funnel plot asymmetry [7]. An iterative algorithm is used to estimate the number of studies considered missing from the funnel plot; these studies can broadly be considered to be those with no counterpart on the opposite side of the funnel. This number of studies is “trimmed” from the asymmetric outlying part of the funnel, leaving a symmetric remainder, which is used to produce an “adjusted” estimate of the “true center” of the funnel using standard meta-analysis techniques. The trimmed studies are replaced, and their missing counterparts are imputed, or “filled,” as mirror images of the trimmed studies, with the mirror axis placed along the adjusted pooled estimate. This last stage is necessary to calculate an adjusted standard error and an adjusted confidence interval for the adjusted pooled estimate.

We applied this method using a random-effects model to our estimates in each outcome; the estimator $L_0$ was used in the iterative procedure to estimate the number of missing studies as previously described [7]. There were eight DFS studies and 12 OS studies, with values of log$_e$(HR) close to and less than zero, estimated as missing (Fig. 2C, D). When these missing studies were imputed and incorporated into random-effects meta-analyses, the pooled estimates were considerably less than those originally calculated for OS (original pooled log$_e$(HR) = 1.63, adjusted pooled log$_e$(HR) = 1.37, 95% CI 1.12–1.62, $P < .0001$) and DFS (original pooled log$_e$(HR) = 1.48, adjusted pooled log$_e$(HR) = 1.27, 95% CI 1.01–1.53, $P < .0001$), although they were still indicative of MYC-N being a potentially important prognostic marker (Table 2).
The Trim and Fill method emphasizes the findings from the visual inspection of the funnel plots, and also from Egger’s test for DFS, that the OS and DFS funnel plots are asymmetric. Assuming this is caused by dissemination bias, it is likely that a number of studies with a small or negative log(HR) are potentially missing from our analyses and that our original pooled estimates from Approach 1 of the log(HR) are likely to be an overestimate of the true effect for OS and DFS. However, given that the pooled results were especially large from the original analysis and also from the Trim and Fill sensitivity analysis, MYC-N seems to be a potentially important prognostic marker.

3.2.4. Other possible causes of funnel plot asymmetry beside dissemination bias

There may be other explanations for the funnel plot asymmetry other than dissemination bias. Sterne et al. [15] recommend using meta-regression to investigate possible factors causing the between-study heterogeneity of effect estimates because these factors could also be causing the observed asymmetry [15]. For example, the bigger estimates with larger standard errors could be related to a particular subgroup of patients, and by removing this subgroup the funnel plot may become more symmetric. However, Lambert et al. [16] have recently shown that meta-regression using summary patient-level covariates may produce misleading and inaccurate conclusions. Because we did not have IPD available for the majority of studies, we could only consider summary patient-level covariates for a meta-regression of our data, and subsequently our analyses were fraught with the problems highlighted by Lambert et al. [16]. For example, most studies involved patients above and below 1 year old at diagnosis, although five involved only patients older than 1 year. Hence, when considering age as a covariate, the only two groups we could consider for age were (i) studies involving only patients older than 1 year, versus (ii) all other studies. However, the majority of studies in (ii) also involved patients older than 1 year old, so the meta-regression results could be misleading and misinterpreted. For these reasons, we do not present meta-regression analyses in this paper.

We considered funnel plots for subgroups of the MYC-N studies to assess whether asymmetry was different for each subgroup of estimates. We looked at subgroups formed by the number of patients in each study (e.g., those with >50), cut-off level used (e.g., 10 copies of MYC-N used), and year of study publication (e.g., those published since 1990). These were the only study characteristics for which sufficient information was available from the literature and whose subgroups were distinct (i.e., not overlapping), unlike age and stage. There was evidence of OS and DFS funnel plot asymmetry in all the subgroups considered, and practically the same answers were found from Egger’s test and the Trim and Fill method as before (Fig. 3). Egger’s Test gave P < .005 for both DFS and OS. Of the 42 original DFS studies included, 19 studies involved less than 50 patients and one study was omitted because the number of patients was unclear. Of the 56 original OS studies included, 38 studies involved less than 50 patients. Hence, although we could consider only a few study characteristics, it seemed that dissemination bias was the most likely cause of funnel plot asymmetry and needed to be investigated further.

To best guide clinical policy, one would want to consider separate meta-analyses and separate funnel plots for the different subgroups of factors in this situation (e.g., age > 1, stage 4, n > 50). However, given the number of possible subgroups, the potential for small numbers of studies in each, and the amount of missing study information, we kept all the study estimates together for the work in this article and continue to assess dissemination bias collectively. Because the majority of studies included all types of patients (e.g., all ages, all stages), the pooled OS and DFS estimates presented here should be considered to relate to an “average” individual with neuroblastoma.

3.3. Approach 3

3.3.1. Sensitivity analysis for the problem of partially unavailable outcomes

A concern of using Egger’s test and the Trim and Fill method in Approach 2 is that we have not specifically taken into account that only partial sets of outcome estimates were available for some studies. For example, the Trim and Fill method estimated that there were eight DFS and 12 OS studies missing based on funnel plot asymmetry; however, there are at least 95 studies missing for OS and 109 studies missing for DFS because 70 studies provided neither outcome, 39 provided OS only, 25 provided DFS only, and other unidentified and unpublished studies may exist. Our meta-analysis results from Approach 1 and dissemination bias assessments from Approach 2 may have been different if more of the missing estimates had been available. For example, some of the studies missing from the lower right-hand side of the OS funnel plot may be the known studies where only DFS was available. Hence, it seems appropriate to first perform sensitivity analyses for the problem of missing estimates from known studies before applying the methods in Approach 2.

3.3.2. Using the observed relationship between OS and DFS estimates

From our data set, 17 studies provided OS and DFS estimates, 39 provided OS only, and 25 provided DFS only (Table 1). It makes sense that OS and DFS estimates should be related. If we assume that the relationship between OS and DFS estimates from the 17 studies providing both is the same for studies where only one outcome is available, then we can potentially use any observed relationship to predict the missing outcome in these studies.

From the 17 studies from which both outcomes were available, OS and DFS estimates seem to be similar for MYC-N levels in neuroblastoma. This similarity between outcomes was cited in three articles as the reason why only
one outcome was reported [18–20]. Furthermore, there was an observed linear relationship between the log_{HR} outcome estimates and between the outcome estimates of the standard error of the log_{HR} (Fig. 4). Using simple linear regression to model log_{HR} for OS against log_{HR} for DFS (and vice-versa) suggested a linear relationship was plausible (adjusted $R^2 = 0.79$). Similarly, a linear regression between the standard error of the log_{HR} for OS against that for DFS (and vice-versa) suggested a linear relationship was plausible (adjusted $R^2 = 0.81$) (Table 3). We used the standard error rather than log_{HR}(standard error) here because it made little difference to the regression models and because we wanted to keep the approach as clear as possible; in practice, the residuals from a regression of log_{HR}(standard error) are more likely to meet the required assumption of normality.

### 3.3.3. Predicting OS estimates where only DFS was available, and vice-versa

For the 64 studies where only one outcome was available (25 only DFS, 39 only OS), we used this observed relationship to predict the outcome currently unavailable, an
Table 3
Linear regression equations for the relationship between overall and disease-free survival estimates, with $R^2$ model-fit statistics.

<table>
<thead>
<tr>
<th>Model</th>
<th>Predicted estimate = $\alpha + \beta \times$ covariate</th>
<th>SE($\alpha$)</th>
<th>SE($\beta$)</th>
<th>$R^2$</th>
<th>Adjusted $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i)</td>
<td>$\log\text{HR}(\text{OS})$ = 0.32 + 0.93 $\times \log\text{HR}(\text{DFS})$</td>
<td>0.24</td>
<td>0.12</td>
<td>0.81</td>
<td>0.79</td>
</tr>
<tr>
<td>(ii)</td>
<td>$\log\text{HR}(\text{DFS})$ = 0.35 + 0.87 $\times \log\text{HR}(\text{OS})$</td>
<td>0.11</td>
<td>0.11</td>
<td>0.84</td>
<td>0.81</td>
</tr>
<tr>
<td>(iii)</td>
<td>$\text{SE}(\log\text{HR}(\text{OS})) = 0.11 + 0.94 \times \text{SE}(\log\text{HR}(\text{DFS}))$</td>
<td>0.082</td>
<td>0.11</td>
<td>0.84</td>
<td>0.81</td>
</tr>
<tr>
<td>(iv)</td>
<td>$\text{SE}(\log\text{HR}(\text{DFS})) = 0.021 + 0.88 \times \text{SE}(\log\text{HR}(\text{OS}))$</td>
<td>0.084</td>
<td>0.10</td>
<td>0.84</td>
<td>0.81</td>
</tr>
</tbody>
</table>

Abbreviations: HR, hazard ratio; OS, overall survival; DFS, disease-free survival; SE, standard error.

approach suggested by Pigott [17]. Using the linear regression equations (i) through (iv) in Table 3, we predicted the log$_t$(HR) for OS when only DFS was available (i), and vice-versa (ii), and similarly the standard error of the log$_t$(HR) for OS when only DFS was available (iii), and vice-versa (iv) (see Table 1 for predicted estimates). Hence, 39 DFS estimates and 25 OS estimates, previously unavailable, were predicted. For example, for study 18 only DFS was available, but we predicted for OS a log$_t$(HR) of 0.55 (using (i)) with standard error 0.38 (using (iii)).

3.3.4. Sensitivity analysis using the predicted values

On the assumption that there is an underlying linear relationship between estimates, the predicted values formed our best estimate of the log$_t$(HR) and its standard error for the missing outcome. For the purposes of sensitivity analysis, we treated these predicted values as known and imputed them into our data set and updated our meta-analyses. We did not take the uncertainty of the predicted estimates into account because for the sensitivity assessment we wanted to gauge how the original results would change if the predicted values were correct; incorporating all the uncertainty would defeat this purpose because the predicted estimates would have little weight in the meta-analysis (this point is considered further in the Discussion).

Meta-analysis for OS now included estimates from 81 studies (56 known plus 25 predicted), as did DFS (42 known plus 39 predicted). The random-effects pooled estimate was similar to the original pooled estimate in Approach 1 for OS (pooled log$_t$(HR) = 1.67, 95% CI 1.48–1.86, $P < .0001$) and DFS (pooled log$_t$(HR) = 1.49, 95% CI 1.31–1.66, $P < .0001$) (Table 2). This indicated that the predicted estimates were similar to the known estimates already in the data set (Fig. 5C and D). Conversely, the known estimates from the studies providing only one outcome must also have been similar to those from studies providing both. These points were further emphasized by reasonably similar meta-analysis results to Approaches 1 and 3 for just the subset of predicted estimates and just the subset of known values from studies reporting only one outcome and from studies reporting both outcomes (Table 4).

On the assumption that the relationship between OS and DFS estimates is the same in studies where only one outcome is available as in those where both are available, this approach suggests that the missing values are likely to be similar to the known estimates already in the data set, and thus the impact of partial sets of outcome estimates on our original pooled estimates is likely to be small. All the results obtained during this approach are in accordance with MYC-N being a potentially important prognostic marker.

3.4. Approach 4: follow Approach 3 with an assessment of dissemination bias

The sensitivity analyses in Approach 3 predicted estimates for only some of the studies known to be missing. An additional concern is the possible impact of those 70 studies where neither OS nor DFS outcomes were available for whom it is more difficult to predict the values of the missing estimates. One also needs to consider the possible impact of any unidentified, unpublished studies. The worst-case scenario is that the outcome estimates from all these studies are not available due to dissemination bias and that they are less positive or less significant than the estimates available. It is therefore important to follow-up Approach 3 with an assessment of the existence and potential impact of dissemination bias, especially because our original assessment of this problem in Approach 2 may now be different given the inclusion of the extra 64 predicted estimates.

The visual inspection of the extended funnel plots, which included the predicted estimates from Approach 3, indicated that asymmetry was still a concern (Fig. 5C, D). Applying Egger’s test to the extended data sets suggested that asymmetry was still a problem for DFS ($P = .046$) and possibly also for OS, with Egger’s $P$ value for OS ($P = .16$) now considerably smaller than the previous value in Approach 2 ($P = .45$). Furthermore, applying the Trim and Fill method using a random-effects model estimated 19 DFS and 18 OS studies to be missing (Fig. 5E, F). This was a greater number of studies than estimated in Approach 2, although the proportion of “filled” to the overall number studies was similar; for example, in Approach 2, eight DFS studies were filled in addition to the 42 already known (19%), whereas in Approach 3, 19 studies were filled in addition to the 81 known (23%). The asymmetry in the funnel plots could not be explained by any study characteristics, so dissemination bias was deemed the most likely cause. It is likely that the remaining 70 known studies not included in Approach 3 for which neither OS nor DFS was available combined with some unknown and unpublished studies are contributing to the dissemination bias problem.
To consider the possible impact of dissemination bias on the meta-analysis results from Approach 3, we incorporated the missing study estimates from Trim and Fill into the meta-analyses alongside the known and predicted values. The random-effects pooled estimates of the log$_e$(HR) for DFS (100 studies: pooled log(HR) = 1.22, 95% CI 1.03–1.40, $P < .0001$) and OS (99 studies: pooled log(HR) = 1.41, 95% CI 1.21–1.61, $P < .0001$) were similar to the Trim and Fill sensitivity results from Approach 2 (Table 2). Therefore, our conclusion echoes that from Approach 2 that the true OS...
...and DFS pooled-effect is likely to be somewhat lower than that obtained in Approach 3, although still indicative of MYC-N being of prognostic value.

4. Discussion

Systematic reviews should seek to incorporate all the published and unpublished evidence [1]; however, our example of MYC-N in neuroblastoma has illustrated the difficulty of achieving this because of dissemination bias and poor statistical reporting in primary studies. In this type of situation, when presented with an incomplete evidence-base, the meta-analyst will be torn between wishing to utilize the evidence available and not wishing to produce biased or misleading conclusions. In many situations it may be sensible not to proceed with meta-analysis, and researchers should not be afraid to make such conclusions and to recommend further studies or collection of IPD. For our review, given the reasons outlined earlier, we undertook meta-analyses. However, in retrospect we would first seek to obtain the IPD from each study rather than try to extract estimates from the published articles, and we recommend others with a similar decision opt for the IPD approach because this is likely to be the most fruitful [9]. Extracting estimates can be a lengthy process and may itself introduce heterogeneity and bias.

Even when using the IPD approach some outcome estimates will be unavailable, data extraction will be required for some studies, and dissemination bias will be a concern. Therefore, whenever meta-analyses are undertaken for multiple outcomes, we have highlighted the opportunity for sensitivity analyses that consider the effect poor reporting and dissemination bias might have on the results and conclusions. However, sensitivity analyses are not easy to perform, and their pooled results and confidence intervals should not be treated as the correct answer but only for assessing the robustness of original meta-analysis results. Interpreting our sensitivity analyses in this manner, Approaches 2, 3, and 4 have strengthened our original conclusions from Approach 1, even though the original pooled estimate is likely to be an overestimate of the true effect. Hence, despite the problems of unavailable estimates and dissemination bias, MYC-N seems to be a potentially important prognostic marker, and future research in this field should prioritize the study of how best to implement MYC-N into the management and treatment of patients with neuroblastoma. This conclusion is perhaps not surprising given the large estimates of the log(HR) originally obtained for MYC-N in Approach 1. Similar sensitivity analyses in other less clear-cut situations, where pooled effect sizes are much smaller, are likely to be even more crucial because conclusions are probably less robust to the various problems of bias.

4.1. Current methods used to assess dissemination bias and poor reporting

There are a variety of methods available for assessing dissemination bias, including hypothesis tests, selection models, and those that produce adjusted pooled estimates [6]. All such approaches can be a helpful assessment tool, but they must be considered only as a guide to the problem because they are vulnerable to incorrect underlying assumptions and may have low power of detecting any problem [21]. One reason for being cautious is that funnel plot asymmetry may not be caused by dissemination bias but rather by the true heterogeneity present, poor design of small studies, or even chance [15]. This has caused the Trim and Fill method to be criticized because it can detect missing studies even in the absence of bias [22]. However, to assess heterogeneity and funnel plot asymmetry properly, IPD is required from the majority of studies [16]. IPD was not available for the majority of studies in our data set, although funnel plot asymmetry was not related to any of the study characteristics we were able to assess, and dissemination bias was still the likely cause.

Poor statistical reporting limits systematic reviews [9, 23], although there is little development of methods to assess its impact on meta-analysis. In this article, we have demonstrated how poor reporting in individual MYC-N studies can cause unavailable OS and DFS estimates, which makes one concerned that the known estimates available might not be an accurate representation of the overall evidence base. This problem is likely to generalize to studies of other prognostic markers and other disease settings [9]. Furthermore, although observational studies are likely to be the most poorly affected, those conducting systematic
reviews of randomized trials are likely to encounter similar problems relating to unavailable outcomes and dissemination bias. The important question is, if we had the additional data from the missing studies, is it possible that the pooled estimates and our conclusions would change?

4.2. Performing the most informative sensitivity analyses given both problems

The problem is thus posed: How to perform worthwhile sensitivity analyses given the problem of unavailable outcome estimates alongside the threat of dissemination bias? There are no simple answers, and possible approaches depend on the type and amount of information available and require assumptions to be made. However, when multiple related outcomes are of interest, it may be possible to develop methods that use the known outcome estimates. The idea of borrowing strength between related outcome estimates has been proposed previously in the form of a multivariate meta-analysis model [24]. Our Approach 3 also uses the known outcome estimates available, but it allows the consideration of any dissemination bias (Approach 4), which is more difficult after the application of a multivariate (multilevel) model.

The appeal of Approach 3 is that we are using the information extracted to specifically predict outcome estimates for studies that were known to be missing, rather than using funnel plot asymmetry to estimate studies possibly missing, as is done in the Trim and Fill method. Furthermore, it is possible to follow this approach with a subsequent assessment of dissemination bias (Approach 4), and so the robustness of the original results can be assessed for these two major problems individually and collectively.

4.3. Critical discussion of Approach 3

To predict estimates in Approach 3, we used the observed close relationship between OS and DFS estimates in the 17 studies where both were available, which is likely to be a reflection of the poor outcome for patients experiencing a recurrence of neuroblastoma [18]. Our predicted estimates were noticeably similar to those already available. In fact, the predicted estimates would have differed only if the known outcome estimates used in the prediction calculation were themselves at the extremes of the available data set. To determine if this is likely, one might consider assessing subsets of the known estimates in each outcome, as can be seen in Table 4. If the known estimates used for predicting others are mostly extreme observations, there is then the additional concern of making predictions beyond the range of the observed data, and one would have to assume here that the observed relationship between outcomes continued above and below the observed range. Of the predicted estimates we calculated for MYC-N, only one was outside the range of those used to estimate the regression model (study 43).

There are additional concerns about Approach 3. For instance, the prediction model used is sensitive to extreme observations, due to the small number of studies providing both OS and DFS; for example, when study 17 is removed, equation (ii) changes considerably (α from 0.038 to 0.34, β from 0.87 to 0.65, adj-R² from 0.79 to 0.54). However, when applying Approach 3 with prediction models formed without study 17, our meta-analysis results changed little (e.g., DFS pooled log_e(HR) changes from 1.49 to 1.45). In general, meta-analyses of prognostic or other studies involve a considerably smaller number of overall studies than we had available, thus limiting the opportunity to observe and then model relationships between outcomes as we have.

4.4. Ignoring the uncertainty in the predicted values

The fact that we are ignoring the uncertainty of the predicted values and treating them as known is an idea perhaps contrary to statistical thinking. For example, the predicted value of log_e(HR) for OS in study 18 was 0.55, but this value is not known and has uncertainty because α and β in equation (i) and log_e(HR) for DFS are estimated (Table 1, Table 3). Because our aim was for sensitivity analyses of our original results, we wanted to see the most likely impact of the missing estimates given our model. If all the uncertainty were taken into account, the predicted values would have little weight in the updated meta-analysis (because the standard error of the log_e(HR) would be large), and the pooled results would therefore be similar to before, defeating the need to assess the robustness of the original results given the problem of unavailable OS and DFS outcomes. It would also mean that the Trim and Fill method could not be used in Approach 4 because this method would impute unrealistic missing studies based on the increased asymmetry in the funnel plots caused by extremely large standard errors in the predicted estimates. Multiple imputation was not considered for Approach 3 because this method takes into account all of the uncertainty in the predicted values and therefore would have limited value in a sensitivity assessment.

4.5. Possible extensions to our work

Ideally, sensitivity methods to assess dissemination bias would use as much of the known information as possible. However, it is difficult to relate the known information to that unknown due to bias, which is why the majority of current methods use the information relating to funnel plot asymmetry. Our assumption that the observed relationship between OS and DFS estimates can be applied to studies where only one outcome was available (i.e., essentially a missing at random assumption) may be inappropriate if the missing outcomes were absent due to dissemination bias; in this situation the relationship between outcomes may be very different. For example, some studies where only DFS was available may have deliberately not reported OS because it was not as interesting or statistically significant.
Sensitivity methods that predict missing estimates assuming a specific biased selection procedure could therefore also be important. For example, in a worst-case scenario, one could estimate the missing data if the meta-analysis results change in each case. A similar approach is taken by Hahn et al. [25], who impute odds ratio estimates for missing subgroups by fitting constraints on the meta-analysis results using the information known. Hutton et al. [26] also use selection models to adjust meta-analysis results for biased outcome reporting. However, neither of these two methods considers the impact of completely unavailable sets of estimates and the possibility of any remaining dissemination bias. Approaches that differentiate between the different dissemination bias problems may enable stronger conclusions than when looking at the overall problem. As the Trim and Fill method and use of funnel plots may be inappropriate when heterogeneity is present [22], selection models and alternative methods that impute missing studies using the information known may become increasingly important.

Some other possible extensions to our work include using other covariate information to provide more reliable prediction models and enable estimates from studies where neither OS nor DFS was available. For example, it is common for studies of prognostic markers to report the results of more than one marker [3], and so one could assess the relationship of estimates across studies between different markers and use it to predict missing marker estimates from those known. This also opens the question of whether the choice of which prognostic markers to report is also affected by dissemination bias. Furthermore, what if there was evidence of biased choice of cut-off levels across studies, and how could sensitivity analyses account for this problem? This was hard to ascertain in our review but may affect the methods and results used if it was a large problem. Application of Bayesian methods may prove useful for all these situations, and would allow the inclusion of other pertinent information alongside the estimates known [27].

4.6. Future research needs

Sensitivity analyses can investigate but do not overcome the impact of reporting problems on meta-analysis. The primary need is to prevent such problems from occurring, and there is currently a drive to improve the conduct, reporting, and prospective registration of primary studies, including those of prognostic markers [9,11,14]. Availability of full IPD is the most promising way forward [9,28]. However, there are additional concerns for IPD reviews, especially cost and time [29], and one cannot assume enough data has been collected for calculation of the outcomes desired. It is therefore likely that sensitivity analyses will continually play an important role in systematic reviews and meta-analysis, and therefore further research into the most appropriate sensitivity methods is required. In particular, for those conducting meta-analysis of multiple related outcomes, sensitivity methods that use the known outcome estimates may be an especially valuable and informative approach to consider.

Acknowledgments

We thank the NHS Health Technology Assessment (HTA) Programme, who originally funded the systematic review of prognostic markers in pediatric oncology. We thank Susan A. Burchill and David Heney for their important contributions to this systematic review, and in particular David R. Jones, who helped extract the estimates we required from the literature. Finally, we thank the two referees whose comments helped us to improve an original draft of this paper.

References


128. Stewart LA, Parmar MK: Meta-analysis of the literature or of individual patient data: is there a difference? Lancet 1993; 341: 418-22.


R.D. Riley, Ph.D. Thesis Bibliography 425


