ESTIMATING THE HETEROGENEITY VARIANCE IN A RANDOM-EFFECTS META-ANALYSIS

Two Volumes

Volume I of II

Dean Langan

Doctor of Philosophy

University of York

Health Sciences

November 2015

Abstract

In a meta-analysis, differences in the design and conduct of studies may cause variation in effects beyond what is expected from chance alone. This additional variation is commonly known as heterogeneity, which is incorporated into a random-effects model. The heterogeneity variance parameter in this model is commonly estimated by the DerSimonian-Laird method, despite being shown to produce negatively biased estimates in simulated data. Many other methods have been proposed, but there has been less research into their properties.

This thesis compares all methods to estimate the heterogeneity variance in both empirical and simulated meta-analysis data. First, methods are compared in 12,894 empirical meta-analyses from the *Cochrane Database of Systematic Reviews* (CDSR). These results showed high discordance in estimates of the heterogeneity variance between methods, so investigating their properties in simulated meta-analysis data is worthwhile. A systematic review of relevant simulation studies was then conducted and identified 12 studies, but there was little consensus between them and conclusions could only be considered tentative.

A new simulation study was conducted in collaboration with other statisticians. Results confirmed that the DerSimonian-Laird method is negatively biased in scenarios where within-study variances are imprecise and/or biased. On the basis of these results, the REML approach to heterogeneity variance estimation is recommended. A secondary analysis combines simulated and empirical meta-analysis data and shows all methods usually have poor properties in practice; only marginal improvements are possible using REML.

In conclusion, caution is advised when interpreting estimates of the heterogeneity variance and confidence intervals should always be presented to express its uncertainty. More promisingly, the Hartung-Knapp confidence interval method is robust to poor heterogeneity variance estimates, so sensitivity analysis is not usually required for inference on the mean effect.

Contents

Abstract iii

List of Figures xvii

List of Tables xxi

Acknowledgements xxiv

Author's declaration xxv

1 Introduction 1

1.1 Principles of systematic reviews and meta-analysis 2

1.2 The Cochrane Database of Systematic Reviews (CDSR) 3

1.3 Types of study-level data 4

1.3.1 Summarising continuous outcome data 4

1.3.1.1 Mean difference (MD) 5

1.3.1.2 Standardised mean difference (SMD) 5

1.3.2 Summarising binary outcome data 6

1.3.2.1 Relative risk (RR) 6

1.3.2.2 Odds ratio (OR) 7

1.4 Heterogeneity of true study effects 8

1.5 Meta-analysis models 9

1.5.1 The standard fixed-effect model 10

1.5.2 The standard random-effects model 10

1.5.3 Modelling with covariates 11

1.5.4 Bayesian models 11

1.6 The inverse-variance method 12

1.7 Quantifying heterogeneity 13

1.7.1 Estimating the heterogeneity variance 14

1.7.2 The I^2 statistic 14

1.8 Prediction intervals 15

1.9 Fixed effect vs. random-effects models 16

1.10 An overview of the thesis 17

1.10.1 An overview of problems associated with heterogeneity variance estimation 17

1.10.2 Focus of the thesis 18

1.10.3 Aims 19

1.10.4 Structure of the thesis 20

2 Methods for estimating the heterogeneity variance **21**

2.1 Introduction 22

2.2 Method of moments approach 23

2.2.1 DerSimonian-Laird (DL) 25

2.2.2 Cochran's ANOVA (CA) 26

2.2.3 Non-parametric Bootstrap DerSimonian-Laird (DL_B) 26

2.2.4 Paule-Mandel (PM) 27

2.2.5 Two-step versions of Paule-Mandel (PM_{CA} & PM_{DL}) 28

2.3 Non-truncated moments-based approaches 29

2.3.1 Hartung-Makambi (HM) 29

2.3.2 Sidik-Jonkman (SJ) 30

2.4 Hunter-Schmidt (HS) 31

2.5 Maximum likelihood approach 32

2.5.1 Maximum likelihood (ML) 33

2.5.2 Restricted maximum likelihood (REML) 34

2.5.3 Approximate restricted maximum likelihood (ARML) 34

2.6 The Bayesian approach 35

2.6.1 Full Bayesian 35

2.6.2 Approximate Bayes Estimator (AB) 36

2.6.3 Bayes modal (BM) 38

2.6.4 Rukhin's approach 38

2.7 Malzahn, Böhning and Holling (MBH) 40

2.8 Alternative within-study variance estimates 40

2.8.1 Bhaumik et al (2012) 41

2.8.2 Berkey et al (1995) and Knapp & Hartung (2003) 41

2.9 A summary of methods 42

2.10 Example meta-analysis 44

2.11 Conclusions 46

3 Methods for confidence intervals of the summary effect 49

3.1 Introduction 50

3.2 Wald-type confidence interval 51

3.3 t-distribution confidence interval 51

3.4 Hartung-Knapp confidence interval 52

3.5 Example meta-analysis 53

3.6 Concluding remarks 54

4 An empirical comparison of heterogeneity variance estimators 57

4.1 Introduction 58

4.2 Methods 59

4.2.1 Included methods for estimating the heterogeneity variance . . 59

4.2.2 Empirical study dataset 60
4.2.3 Summary statistics 60
4.2.4 Data analysis 61
4.3 Results 63
4.3.1 A summary of CDSR meta-analyses 64
4.3.2 Agreement between heterogeneity variance estimates 65
4.3.3 Agreement between summary effects 69
4.3.4 Agreement between precision of the summary effect 69
4.3.5 Agreement between p-values 71
4.3.6 Factors to explain the level of agreement between heterogeneity variance estimates 75
4.4 Examples 75
4.5 Discussion 80
4.6 Conclusion 82

5 A review of simulation studies to compare heterogeneity variance estimators 83

5.1 Introduction 84
5.2 Methods 85
5.2.1 Search Strategy 85
5.2.2 Eligibility criteria 85

5.2.3 Data extraction 86

5.3 Results 86

5.3.1 Search Results 86

5.3.2 Simulation methods and parameter values 87

5.3.3 Performance measures 92

5.3.4 Heterogeneity variance estimators 93

5.3.5 Performance of point estimators of the heterogeneity variance 94

5.3.5.1 DerSimonian-Laird (DL) 96

5.3.5.2 Cochran's ANOVA (CA) 97

5.3.5.3 Paule-Mandel (PM), and its variants (PM_{CA} and PM_{DL}) 97

5.3.5.4 Restricted maximum likelihood (REML) and its approximation (ARML) 98

5.3.5.5 Maximum likelihood (ML) 100

5.3.5.6 Hunter-Schmidt (HS) 100

5.3.5.7 Sidik- Jonkman estimators (SJ and SJ_{CA}) 100

5.3.5.8 Malzahn, Bohning and Holling (MBH) 101

5.3.5.9 Bayesian estimators: Bayesian modal (BM) and Rukhin's estimators (RU, B0, BP, SB) 101

5.3.6 Performance of estimators of the summary effect 102

5.3.7 Performance of confidence intervals for the summary effect . . 102

5.3.8 Performance of heterogeneity variance estimators using other methods to estimate the within-study variance 103

5.3.9 A summary of recommendations 103

5.4 Discussion 104

5.5 Conclusion 108

6 Methods for a new simulation study **111**

6.1 Introduction 112

6.2 Aims 113

6.3 Heterogeneity variance estimators 113

6.4 Performance measures 115

6.5 Simulation methods 116

6.5.1 Simulating true study effects 118

6.5.2 Standardised mean difference (SMD) meta-analyses 118

6.5.3 Odds ratio meta-analyses 119

6.6 Parameter values 120

6.6.1 Number of studies 120

6.6.2 Heterogeneity variance parameter values 122

6.6.2.1 Method to derive heterogeneity variance parameter values 122

6.6.2.2 A summary of heterogeneity variance parameter values 123

6.6.3 Summary effect 124
6.6.4 Distribution of true study effects 124
6.6.5 Study sample sizes 125
6.6.6 Average probability of event across study groups 125
6.7 An overview of the simulation study 126

7 Main simulation study results **129**

7.1 Introduction 130
7.2 Heterogeneity variance estimators excluded from the main analysis . . 131
7.3 Simulated scenarios not presented in full 132
7.4 Selected performance measures 133
7.5 Results 133
7.5.1 Properties of heterogeneity variance parameter estimates . . . 133
7.5.1.1 DerSimonian-Laird (DL) 134
7.5.1.2 Cochran's ANOVA (CA) 135
7.5.1.3 Paule-Mandel (PM) 142
7.5.1.4 Two-step Cochran's ANOVA (PM_{CA}) 143
7.5.1.5 Two-step DerSimonian-Laird (PM_{DL}) 143
7.5.1.6 Maximum likelihood (ML) and Hunter-Schmidt (HS) 144
7.5.1.7 REML 144
7.5.1.8 Hartung-Makambi (HM) 145

7.5.1.9 Sidik-Jonkman (SJ) 146

7.5.1.10 Sidik-Jonkman (CA initial estimate) (SJ_{CA}) 146

7.5.1.11 A summary of all simulated scenarios 147

7.5.2 Properties of heterogeneity variance estimates for varying effect sizes 149

7.5.3 Properties of estimates of the summary effect 151

7.5.4 Coverage of 95% confidence intervals for the summary effect . 153

7.5.4.1 Wald-type confidence interval 153

7.5.4.2 t-distribution confidence interval 154

7.5.4.3 Hartung-Knapp confidence interval 158

7.5.4.4 A summary of coverage in all simulated scenarios . . 159

7.5.5 Convergence of ML and REML estimates of heterogeneity . . 160

7.5.6 An overview of the results 160

7.5.6.1 Properties of estimates of the heterogeneity variance 161

7.5.6.2 Properties of estimates of the summary effect 162

7.5.6.3 Properties of 95% confidence intervals of the summary effect 162

7.6 Discussion 163

7.7 Conclusions 166

8 Properties of heterogeneity variance estimators in meta-analyses of Cochrane reviews 169

8.1 Introduction . . . 170

8.2 Methods . . . 170

8.2.1 Mapping empirical to simulated meta-analyses . . . 171

8.2.2 Performance measures . . . 173

8.2.3 Included estimators of the heterogeneity variance . . . 174

8.2.4 Analysis methods . . . 174

8.3 Results . . . 175

8.3.1 The proportion of CDSR meta-analyses matched to each simulated scenario . . . 175

8.3.2 Performance of heterogeneity variance estimators in CDSR meta-analyses . . . 177

8.3.2.1 Predicted bias of the heterogeneity variance . . . 178

8.3.2.2 Predicted mean squared error of the heterogeneity variance . . . 179

8.3.2.3 Predicted bias of summary effect estimates . . . 179

8.3.2.4 Predicted coverage of 95% confidence intervals of the summary effect . . . 182

8.4 Discussion . . . 184

8.5 Conclusions . . . 186

9 Discussion and conclusions **189**

9.1 Introduction 190

9.2 Thesis summary 190

9.3 Discussion 192

9.4 Further work 196

9.4.1 Logistic regression models for meta-analysis 196

9.4.2 Distributions of study size 198

9.4.3 Wider strategies for heterogeneity variance estimation in *problem* meta-analyses 198

9.4.4 Confidence intervals for the heterogeneity variance 199

9.4.5 Implementation in statistical software 199

9.5 Conclusion 199

Appendix A: Supplementary material from chapter 2 **203**

A.1 Search strategy 204

Appendix B: Supplementary material from chapter 4 **205**

Appendix C: Supplementary material from chapter 5 **209**

Appendix D: Supplementary material from chapter 6 **213**

Appendix E: Simulation study protocol **237**

Appendix F: Supplementary material from chapter 7 **259**

Abbreviations **267**

Bibliography **269**

Volume II: Supplementary graphs from the simulation study

1. Bias of heterogeneity estimates **287**

2. Mean squared error of heterogeneity estimates **312**

3. Proportion of zero heterogeneity estimates **337**

4. Mean bias of mean effect estimates **362**

5. Coverage of Z-type confidence intervals **387**

6. Coverage of t-distribution confidence intervals **412**

7. Coverage of Knapp-Hartung confidence intervals **437**

List of Figures

2.1 Forest plot with five studies in a meta-analysis evaluating hawthorn extract for chronic heart failure 45

3.1 The summary effect and 95% confidence interval for all combinations of heterogeneity variance and confidence interval methods 55

4.1 The numbers of studies included in all 2,894 meta-analyses 65

4.2 The distribution of τ^2 and I^2 estimates for OR and standardised mean difference meta-analyses calculated from the DL method. 66

4.3 Bland-Altman scatter plots comparing I^2 estimates from different heterogeneity variance methods. 67

4.4 Bland-Altman scatter plots comparing summary effect estimates using different heterogeneity variance estimation methods. 68

4.5 Bland-Altman scatter plots comparing summary effect estimates and standard errors using different heterogeneity variance estimation methods. 70

4.6 Bland-Altman scatter plots comparing differences in I^2 estimates against (upper-right panel) the number of studies and (lower-left panel) the total information 76

4.7 Forest plot of a meta-analysis of seven studies, with combined effects illustrated from various methods of heterogeneity variance estimation 77

4.8 Forest plot of a meta-analysis of six studies, with combined effects illustrated from various methods of heterogeneity variance estimation 79

5.1 Mean bias from selected simulation results in Novianti et al. [78] including simulated meta-analyses of type SMD and OR. 98

5.2 Mean squared error from selected simulation results in Novianti et al. [78] including simulated meta-analyses of type SMD and OR. 99

6.1 Probability density function of skew-normal distribution 125

7.1 Proportional mean bias (left-hand-side) and proportional mean squared error (right-hand-side) in selected scenarios with B0, BP and MBH heterogeneity variance estimators included. 132

7.2 Mean bias of heterogeneity variance estimates in standardised mean difference outcome meta-analyses 136

7.3 Mean bias of heterogeneity variance estimates in odds ratio meta-analyses with event probability 0.1 to 0.5 137

7.4 Mean squared error of heterogeneity variance estimates in standardised mean difference meta-analyses 138

7.5 Mean squared error of heterogeneity variance estimates in odds ratio meta-analyses with event probability 0.1 to 0.5 139

7.6 Proportion of zero heterogeneity variance estimates in standardised mean difference meta-analyses 140

7.7 Proportion of zero heterogeneity variance estimates in odds ratio meta-analyses with event probability 0.1 to 0.5 141

7.8 Mean bias of heterogeneity variance estimates in odds ratio meta-analyses containing small studies and with event probability 0.1 to 0.5 150

7.9 Mean bias of the summary effect estimates in odds ratio meta-analyses with rare events. 151

7.10 Coverage of 95% confidence intervals of the summary effect in standardised mean difference meta-analyses with small-to-medium studies . 155

7.11 Coverage of 95% confidence intervals of the summary effect in odds ratio meta-analyses with small-to-medium studies and an average event probability of 0.05. 156

7.12 Coverage of 95% confidence intervals of the summary effect in odds ratio meta-analyses with small and large studies and an average event probability of 0.1 to 0.5. 157

8.1 A heat map of CDSR meta-analyses falling into each simulated scenario176

8.2 Predicted distribution of proportional bias of the heterogeneity variance estimators 180

8.4 Predicted distribution of proportional MSE of heterogeneity variance estimates 181

8.3 Predicted distribution of bias of the summary effect (θ) 182

8.5 Predicted distribution of the coverage of summary effect confidence intervals; Wald-type, t-distribution and Hartung-Knapp. 183

B.1 Bland-Altman scatter plots comparing I^2 estimates from different heterogeneity variance methods excluded from the main results 206

B.2 Bland-Altman scatter plots comparing τ^2 estimates from different heterogeneity variance methods. 207

E.1 Histogram of the numbers of studies in meta-analyses in the Cochrane Database of Systematic Reviews (CDSR) 254

E.2 Probability density function of skew-normal distribution 255

F.1 Proportional median bias (left-hand-side) and proportional mean bias (right-hand-side) of the heterogeneity variance in selected scenarios to show why median bias is excluded from the main results. 260

F.2 Proportional median squared error (left-hand-side) and proportional mean squared error (right-hand-side) of the heterogeneity variance in selected scenarios to show why median squared bias is excluded from the main results. 260

F.3 Mean squared error of the summary effect in selected scenarios to show why this measure is excluded from the main results. 261

F.4 Power to detect a statistically significant summary effect in selected scenarios to show why this measure is excluded from the main results. 261

F.5 Mean error of the error-interval estimation of effect in selected scenarios to show why this measure is excluded from the main results. . . 262

F.6 Variance error of the error-interval estimation of effect in selected scenarios to show why this measure is excluded from the main results. 262

F.7 Mean squared error of heterogeneity variance estimates in odds ratio meta-analyses containing small studies and with event probability 0.1 to 0.5 263

List of Tables

1.1 Standard contingency table notations for a study i with a binary outcome 6

2.1 Summary of heterogeneity variance estimators 43

2.2 Heterogeneity variance estimates derived from different methods and associated I^2 estimates 46

4.1 The original outcome measures of included meta-analyses from the CDSR 64

4.2 The difference between heterogeneity variance estimation methods in terms of p-value categories derived from the Wald-statistic. 72

4.3 The difference between heterogeneity variance estimation methods in terms of p-value categories derived from the Hartung-Knapp method. 73

5.1 Simulation methods, parameter values and performance measures used in the 12 included simulation studies 90

5.2 Underlying ranges of I^2 in each publication 91

5.3 Summary of performance measures reported in the 12 included simulation studies 93

5.4 Summary of heterogeneity variance estimators compared in the 12 included simulation studies 95

5.5 A summary of recommendations from the 12 included publications . . 105

6.1 Set of parameter values and distributions to simulate meta-analyses . 121

7.1 A summary of the properties of heterogeneity variance estimators for all scenarios of standardised mean difference and odds ratio meta-analyses with effect size *0.5*. 148

7.2 A summary of coverage for all scenarios of standardised mean difference and odds ratio meta-analyses with effect size 0.5 159

8.1 Matching criteria for simulated and empirical CDSR meta-analysis data 172

8.2 The average proportional MSE of heterogeneity variance estimates in CDSR meta-analyses 181

9.1 Heterogeneity variance estimators available in popular statistics software 197

C.1 Performance measures in simulated data 212

D.1 τ^2 parameter values for each simulated scenario. 214

D.2 90% reference range of underlying I^2 values for each simulated scenario. 215

E.1 Set of parameter values and distributions to simulate meta-analyses . 253

F.2 The percentage of scenarios and meta-analyses in which ML failed to converge 264

F.3 The percentage of scenarios and meta-analyses in which REML failed to converge 264

F.1 Excluded performance measures and reasons for exclusion 265

F.4 Proportion of studies with zero events in either study arm in simulated odds ratio meta-analyses 266

Acknowledgements

I would first like to thank my co-supervisors Dr Mark Simmonds and Professor Julian Higgins for offering me this research opportunity and their substantial help and guidance throughout. Their help not only steered me in the right direction with my research, but also helped me to become a more effective researcher. Thank you to the other members of the Thesis Advisory Panel, Professor Lesley Stewart and Professor Martin Bland, for their invaluable guidance and continued interest outside of our meetings. Thank you to Dr Rebecca Turner, for granting permission to use the dataset of Cochrane meta-analyses and for responding to my queries on the dataset. I would also like to thank the other collaborators involved in the design of my simulation study; Dr Dan Jackson, Dr Jack Bowden, Dr Areti Angeliki Veroniki, Dr Evangelos Kontopantelis and Professor Wolfgang Viechtbauer. I am also thankful to everyone in CRD for such a friendly working environment.

Author's declaration

I declare that, except where explicit reference is made to the contribution of others, that this thesis is the result of my own work and has not been submitted for any other degree at the University of York or any other institution. All sources are acknowledged as references.

The work presented in chapter 4, has been presented at the Young Statistician's Meeting (YSM) in Bristol (UK, July 2014), the annual international meeting of the Society of Research Synthesis Methods (SRSM) in York (UK, July 2014) and at the International Society for Clinical Biostatistics (ISCB) in Vienna (Austria, August 2014). This work has also been published, with the reference as follows:

> Langan, D., Higgins, J. P. T., and Simmonds, M. 2015. An empirical comparison of heterogeneity variance estimators in 12894 meta-analyses. *Res Synth Methods* 6, 2, 195-205.

The work presented in chapter 5 will also be published in Research synthesis methods. An early view of this paper is available at the time of writing:

> Langan, D., Higgins, J. P. T., and Simmonds, M. 2016. Comparative performance of heterogeneity variance estimators in meta-analysis: a review of simulation studies. *Res Synth Methods*.

Chapter 1

Introduction

1.1 Principles of systematic reviews and meta-analysis

A systematic review collates and synthesises all relevant evidence for a given research question using transparent and comprehensive procedures. Health researchers may be inundated with information in the absence of a systematic review [51], so may be unconsciously selective and potentially come to biased conclusions. A systematic review can identify gaps in knowledge, so that research can be more focused on areas where little is known or evidence in inconclusive. They allow those making changes to health policy to make better judgements based on all available evidence and can provide conclusive answers that would not be possible using individual studies.

Guidance for conducting a systematic review has been published by the Cochrane Collaboration [51], Cooper and Hedges [19] and the Centre for Reviews and Dissemination (University of York, UK) [13]. These guidance documents recommend the following common steps. First, a research question should be identified and formulated which is focused and concise, giving the review a clear aim. The next step is to carry out a search of all literature that may provide relevant evidence in answering the research question. Evidence is then appraised and selected for inclusion based on explicit criteria decided at the planning stage of the systematic review. Exclusion criteria may include, for example, the exclusion of literature in a foreign language or studies that are not randomised controlled trials. The PRISMA statement (Preferred Reporting Items for Systematic reviews and Meta-Analyses) advises the inclusion of a flow diagram in the final report that documents all searching and screening steps [76]. Finally, evidence is synthesised in a manageable and digestible way that allows readers to consider conclusions made by the reviewer and also investigate how these conclusions were drawn [13, 71].

Glass [32] coined the term meta-analysis, referring to the statistical collation of results of related studies for the purpose of integrating the findings. Since the aims of a meta-analysis complement those of a systematic review, they are frequently

included as a component of a systematic review. However, these two approaches to evidence synthesis can also exist in isolation [4]. A meta-analysis is a statistical synthesis or summary of studies, most commonly producing an 'average' result across the studies. It is an increasingly popular type of analysis in medical research [69, 112], and has also been used in other areas including social [31] and education research [1].

Many of the reasons to conduct a meta-analysis stem from the reasons to conduct a systematic review; they help to understand what is often a large and diverse base of evidence. Furthermore, a meta-analysis usually has greater power to detect a statistically significant result than any one of the included studies [13]. A meta-analysis can also help to reduce bias in a systematic review because of the transparency of its method; this is particularly true if methods are defined up-front and justified in a review protocol.

1.2 The Cochrane Database of Systematic Reviews (CDSR)

One of the most notable contributing factors in the increase of published systematic reviews and meta-analyses is the formation of Cochrane (formerly The Cochrane Collaboration). The collaboration was founded in 1993 [36] and has grown such that there are now over 15,000 contributing members and have produced and continue to maintain over 3000 open-access systematic reviews [51]. The main aim of collaboration is to provide researchers, or anyone with personal health concerns with high-quality resources for making informed health decisions. Their systematic reviews are published in the *Cochrane Database of Systematic Reviews* (CDSR) along with review protocols and editorials. Meta-analyses are often are often included within the reviews to statistically combine study results. Data from meta-analyses published by the collaboration are used in thesis, as noted in the front matter.

1.3 Types of study-level data

In this section, I define the type of study-level data required for meta-analysis. Ideally, this can be extracted directly from published papers, but this isn't always possible. Methods for meta-analysis using these data are given from section 1.5 onwards.

Required study-level data usually consist of an estimate of some parameter and its variance, denoted by $\hat{\theta}_i$ and $\hat{\sigma}_i^2$ respectively for a given study i. This parameter is commonly referred to as an effect size. In a health research setting, this parameter often represents a measure of the difference between two groups, such as an active treatment groups and a control/placebo group. For example, a study may measure the risk of myocardial infarction in a group of patients receiving intravenous magnesium and in a control group [114]. In this case, $\hat{\theta}_i$ represents an estimated difference in the risk between these groups [46]. To increase generality, I refer to them as groups one and two in this thesis. A number of measures can be used to calculate $\hat{\theta}_i$ depending on the type of study outcome, such as an odds ratio for a binary outcome, or a standardised mean difference for a continuous outcome. I show how these measures, among others, are calculated in the following two sections.

1.3.1 Summarising continuous outcome data

In studies with a continuous outcome, data from each group can be summarised by its mean, standard deviation and sample size. $\hat{\mu}_{1i}$ and $\hat{\mu}_{2i}$ denote the observed means, $\hat{sd}_{1i}$ and $\hat{sd}_{2i}$ denote the observed standard deviations and n_{1i} and n_{2i} denote the sample sizes in groups 1 and 2 respectively. The mean difference and the standardised mean difference are two commons ways to measure the difference in $\hat{\mu}_{1i}$ and $\hat{\mu}_{2i}$, which are calculated as follows.

1.3.1.1 Mean difference (MD)

The mean difference (MD) can be calculated by [8]:

$$\hat{\theta}_i = \hat{\mu}_{1i} - \hat{\mu}_{2i}$$

If we assume the equal variances of $\hat{\mu}_{1i}$ and $\hat{\mu}_{2i}$, the variance of $\hat{\theta}_i$ is:

$$\hat{\sigma}_i^2 = \frac{n_{1i} + n_{2i}}{n_{1i} n_{2i}} \cdot S_i^2$$

where

$$S_i^2 = \frac{(n_{1i} - 1)\,\hat{sd}_{1i}^2 + (n_{2i} - 1)\,\hat{sd}_{2i}^2}{n_{1i} + n_{2i} - 2} \tag{1.1}$$

Alternatively, without making the equal-variances assumption:

$$\hat{\sigma}_i^2 = \frac{\hat{sd}_{1i}^2}{n_{1i}} + \frac{\hat{sd}_{2i}^2}{n_{2i}}$$

1.3.1.2 Standardised mean difference (SMD)

If MDs are comparable but on different scales, it is not advisable to combine them in a meta-analysis in their current form. For example, the continuous outcome of physical functioning could be measured in rehabilitation studies using measures based on different questionnaires with different scoring methods [130]. To address this problem and allow studies to be pooled more meaningfully, we calculate standardised mean differences (SMD) [8]:

$$\hat{\theta}_i = \frac{\hat{\mu}_{1i} - \hat{\mu}_{2i}}{S_i}$$

where S_i is the estimated standard deviation of the mean difference (see formula 1.1) an assumes equal variances between groups.

When $\hat{\theta}_i$ is on the SMD scale, a good approximation of its variance is:

$$\hat{\sigma}_i^2 = \frac{n_{1i} + n_{2i}}{n_{1i} n_{2i}} + \frac{\hat{\theta}_i^2}{n_{1i} + n_{2i}}$$

Hedges [43] proved that the SMD measure has positive bias in studies with small sample sizes and therefore suggested applying a correction factor; the bias corrected $\hat{\theta}_i$ becomes $J_i \cdot \hat{\theta}_i$ and $\hat{\sigma}_i^2$ becomes $J_i^2 \cdot \hat{\sigma}_i^2$ where $J_i = 1 - 3/\left(4\left(n_{1i} + n_{2i} - 2\right) - 1\right)$. The correction factor has since become widely used and suggested in many meta-analysis texts [8, 19, 42, 128].

Continuous study outcomes are usually on the same scale, so the *unstandardised* mean difference (MD) is more commonly used in practice [19, 88].

1.3.2 Summarising binary outcome data

In studies with a binary outcome, data can be presented in the form of a contingency table (e.g. table 1.1). From this data, we can derive measures that compare the event probability between groups such as the relative risk or odds ratio.

	Event	No event	Total
Group 1	a_i	b_i	$n_{1i} = a_i + b_i$
Group 2	c_i	d_i	$n_{2i} = c_i + d_i$
Total	$a_i + c_i$	$b_i + d_i$	$N_i = a_i + b_i + c_i + d_i$

Table 1.1: Standard contingency table notations for a study i with a binary outcome

1.3.2.1 Relative risk (RR)

The risk of an event in groups one and two are a_i/n_{1i} and c_i/n_{2i}. The relative risk (RR) is a comparison between the two groups and is thus $\left(a_i/n_{1i}\right) / \left(c_i/n_{2i}\right)$

[8]. This measure is transformed onto the log scale for meta-analysis to make the effect estimate within a given study conform approximately to a normal sampling distribution. Normality of study estimates is one of the assumptions in the standard meta-analysis models introduced in section 1.5.1. The log RR ($\hat{\theta}_i$) and its variance ($\hat{\sigma}_i^2$) in each study are therefore:

$$\hat{\theta}_i = log\left(\frac{a_i/n_{1i}}{c_i/n_{2i}}\right)$$

$$\hat{\sigma}_i^2 = \frac{1}{a_i} - \frac{1}{n_{1i}} + \frac{1}{c_i} - \frac{1}{n_{2i}}$$

$\hat{\theta}_i$ and $\hat{\sigma}_i^2$ cannot be calculated in the above formulae when zero events are observed in one or both groups. Throughout this thesis, a continuity correction is applied when required by adding 0.5 to a_i, b_i, c_i and d_i [10]. Other methods are available for dealing with zero events [113].

1.3.2.2 Odds ratio (OR)

The odds of an event in study groups one and two are a_i/b_i and c_i/d_i. From this, the odds ratio (OR) is $(a_i/b_i)\,/\,(c_i/d_i)$. As with RRs, ORs are transformed onto the log scale:

$$\hat{\theta}_i = log\left(\frac{a_i/b_i}{c_i/d_i}\right)$$

$$\hat{\sigma}_i^2 = \frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}$$

As for RR, a continuity correction is required when zero events are observed.

The OR measure has superior statistical properties over RR [5, 19]. First, estimated ORs don't change when the cause and effect is reversed (i.e. if the study groups become the event of interest and vice versa). Second, logistic regression methodology of binary data has developed using the odds ratio measure, so this gives added convenience when conducting a logistic regression analysis [5]. Finally, ORs follow the normal distribution more closely than RRs when transformed to the log scale. These are perhaps some of the reasons why it is the most commonly used measure in systematic reviews of health research (see chapter 4 for a summary of outcome measures used in meta-analyses from Cochrane reviews).

1.4 Heterogeneity of true study effects

Studies brought together in a meta-analysis usually differ to some extent in how they are designed and conducted. Therefore, observed study effects ($\hat{\theta}_i$) in a meta-analysis often have a higher dispersion than what is expected from their observed variances ($\hat{\sigma}_i^2$) alone. This extra variability is known as *heterogeneity* and is the central theme of this thesis. Meta-analyses containing heterogeneous studies should not only combine studies into an average result, they should also estimate and explore the plausible range of study effects [8, 48]. This is particularly necessary, for instance, in a meta-analysis where the intervention is proven effective on average but may be harmful in certain cases.

The Cochrane handbook suggests causes of heterogeneity can be split into two categories; methodological and clinical [51]. Methodological heterogeneity is the variability caused by differences in study design and risk of bias. Clinical heterogeneity is the variability caused by differences in the participants, interventions and outcomes studied. Clinical and methodological heterogeneity are both observed because of differences in the study design and its conduct. The key difference is that for clinical heterogeneity, differences in study design uncover heterogeneity that is ultimately attributable to variation of intervention effects in real settings. Thus, Glasziou and

Sanders [33] alternatively defines these two distinct categories of heterogeneity as *artificial* and *real*. The former is considered nuisance heterogeneity, exploring this can tells us little about how an intervention works outside of the study setting.

The causes of heterogeneity can be numerous and diverse, which often makes identifying them infeasible in practice. Alternatively, they may simply be unknown. To address the presence of heterogeneity in such cases, we may (1) refrain from pooling the studies together in a meta-analysis if extent of heterogeneity is too great, (2) choose to ignore it or (3) account for this extra variability in the analysis stage [51]. The latter two can be accomplished by implementing fixed and random-effects models respectively, which I introduce in the next section.

In some cases, researchers may hypothesise that quantified study characteristics explain all or a proportion of the observed heterogeneity. These characteristics can be used to carried out a sub-group analysis or added as a covariate in the meta-analysis model [112, 115]. This can be accomplished by implementing a meta-regression model, which I also introduce in the next section. Meta-regression typically only works with few study characteristics because meta-analyses in health research contain few studies [115].

1.5 Meta-analysis models

I first introduce the fixed-effect model in section 1.5.1 which makes the assumption that study effects are homogeneous. In section 1.5.2, the random-effects model includes an added variance parameter to taken into account any observed, but unexplained, heterogeneity. I show in section 1.5.3 how study covariates can be added random effects model when some heterogeneity can be explained. In section 1.5.4, I introduce the Bayesian approach to random-effects models.

1.5.1 The standard fixed-effect model

The term 'fixed-effect' refers to the fact that all studies in the meta-analysis are assumed to be estimating a common parameter value θ. Meta-analysis models usually assume observed study effects in a meta-analysis conform to the normal distribution:

$$\hat{\theta}_i \sim N\left(\theta, \sigma_i^2\right)$$

where θ is the true fixed summary effect size in the meta-analysis and σ_i^2 is the true variance in studies $i = 1, ..., k$.

1.5.2 The standard random-effects model

A random-effects model accounts for the possibility that $\hat{\theta}_i$ are estimates of different true study effect parameters θ_i:

$$\hat{\theta}_i \sim N\left(\theta_i, \sigma_i^2\right)$$

In this model, the distribution of θ_i has a mean θ and a heterogeneity variance denoted by τ^2. Some meta-analysis methods assume the normal distribution for θ_i:

$$\theta_i \sim N\left(\theta, \tau^2\right)$$

A crucial, but sometimes overlooked distinction [89], from a fixed-effect model is that $\hat{\theta}$ is an estimate of the *average* from a distribution of study effects [48].

1.5.3 Modelling with covariates

Study-level variables may be available that explain a proportion of the total heterogeneity. In this case, a meta-regression can be carried out by adding these as parameters in the random-effects model:

$$\hat{\theta}_i \sim N\left(\theta_i, \sigma_i^2\right)$$

$$\theta_i \sim N\left(\theta + \sum_{j=1}^{m} \beta_j, \tau^2\right)$$

where β_{ij} is the jth variable in a model containing m covariates.

In this model, τ^2 can be considered the residual variance of the true effects. The same methods are available to estimate τ^2 in both models, which are introduced in section 1.7.1 and detailed comprehensively in chapter 2.

1.5.4 Bayesian models

A Bayesian approach to meta-analysis may be taken using any of the above models [106, 111]. The distinguishing feature of a Bayesian model is that parameters are considered random quantities so we want to estimate its distribution rather than its value. The approach also involves combining the meta-analysis data with our prior beliefs about the parameters. For example, if we assume the random-effects model with no covariates, we can select prior distributions for the parameter values τ^2 and θ. These prior distributions can be specified based on expert opinion or on similar meta-analyses in the same research field [49, 120]. Prior distributions can also be vague if researchers can deduce little about the parameter value prior to conducting the meta-analysis.

Applying Bayes theorem results in the following joint posterior distribution, where φ is the vector of parameters of interest (e.g. τ^2 and θ):

$$P(\varphi \mid data) \approx P(\varphi)\, P(data \mid \varphi)$$

$P(\varphi \mid data)$ is the posterior distribution based on a combination of the prior distribution, $P(\varphi)$, and the observed data $P(data \mid \varphi)$.

Running a Bayesian analysis with many model parameters can be complicated and involves high-dimensional integration. However, it is becoming easier with recent technological advances. The posterior distribution is calculated using Markov Chain Monte Carlo (MCMC) methods. MCMC works by simulating from the high dimensional joint probability distribution, most commonly for the parameters θ and τ^2, to find a solution. Gibbs Sampling is used often in practice and is known to work well in this setting [111].

1.6 The inverse-variance method

One of the main aims of a meta-analysis is to combine studies and produce an average estimate for θ. Studies typically vary in terms of size and assuming studies are all of equal quality, larger studies tend to estimate the parameter with more precision. The inverse-variance method is commonly used to combine studies in a meta-analysis, which gives more precise studies a larger weighting and thus more influence on the average effect size. Using this method, θ and its variance can be estimated by:

$$\hat{\theta} = \frac{\sum_{i=1}^{k} \hat{w}_i \hat{\theta}_i}{\sum_{i=1}^{k} \hat{w}_i} \tag{1.2}$$

$$Var\left(\hat{\theta}\right) = \frac{1}{\sum_{i=1}^{k} \hat{w}_i} \tag{1.3}$$

Study weights can only be estimated from the data and are denoted by $\hat{w}_i$, which are calculated by the reciprocal of $Var\left(\hat{\theta}_i\right)$. If we assume a common effect, like the fixed-effect model in section 1.5.1, the within-study variance is assumed to account for all the variability of $\hat{\theta}_i$ and therefore $\hat{w}_i = 1/\hat{\sigma}_i^2$. If we allow for random effects, like the random-effects model in section 1.5.2, $\hat{w}_i = 1/\left(\hat{\sigma}_i^2 + \hat{\tau}^2\right)$.

Confidence intervals for θ are commonly calculated using the Wald-type method [25] based on the above variance (1.3):

$$\left[\hat{\theta} - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \cdot \sqrt{Var\left(\hat{\theta}\right)},\ \hat{\theta} + \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \cdot \sqrt{Var\left(\hat{\theta}\right)}\right] \tag{1.4}$$

where $\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$ is the $\left(1 - \frac{\alpha}{2}\right)$th percentile of the normal distribution and α is the significance level. α is usually 0.05, leading to a 95% confidence interval.

Other confidence interval methods are available, which I describe in chapter 3. Alternative methods exist for combining studies that are specific to meta-analysis of binary data, including Peto [81] and Mantel-Haenszel methods [75]. These methods can only estimate the summary effect if common effects are assumed, but heterogeneity can still be estimated and incorporated in the confidence interval for the summary effect [128].

1.7 Quantifying heterogeneity

I now outline a common method to estimate the heterogeneity variance parameter, τ^2, as defined in the standard random-effects model in section 1.5.2. Then, I introduce the I^2 statistic, which estimates the proportion of study effect inconsistency that is attributable to heterogeneity.

1.7.1 Estimating the heterogeneity variance

A common approach to estimating the heterogeneity variance parameter, τ^2, is that proposed by DerSimonian and Laird [25]. The method is based on the Q-statistic:

$$Q = \sum_{i=1}^{k} \hat{w}_i(\hat{\theta}_i - \hat{\theta})^2$$

Q is also used as a test statistic for the presence of heterogeneity, a p-value for this test is derived by referring Q to the χ^2-distribution with $k-1$ degrees of freedom. If $\hat{\sigma}_i^2$ adequately account for the total observed variance and $\hat{\theta}_i$ are normally distributed around $\underline{\theta_i}$, then $E[Q] = k-1$, i.e. the expected value of χ^2_{k-1}.

The DerSimonian-Laird method to estimate τ^2 is based on Q:

$$\hat{\tau}^2 = max\left\{0,\ \frac{Q-(k-1)}{\sum_{i=1}^{k}\hat{w}_i - \frac{\sum_{i=1}^{k}\hat{w}_i^2}{\sum_{i=1}^{k}\hat{w}_i}}\right\}$$

The numerator is a standardised measure of the extent that Q exceeds its expected value under the common effect assumption. This measure is converted to the same scale as $\hat{\theta}$ by the denominator. The method estimates τ^2 with no distribution assumption for θ_i.

When $Q < k-1$, heterogeneity variance estimates are truncated at zero as shown in the formula above. Other methods to estimate τ^2 are available; these are detailed in chapter 2.

1.7.2 The I^2 statistic

I^2 represents the proportion of the total variance that can be attributed to heterogeneity of true study effects and is more intuitive to interpret than τ^2. This is derived by transforming the Q-statistic [47]:

$$I^2 = \frac{Q - (k-1)}{Q} \cdot 100\% \tag{1.5}$$

Unlike τ^2, I^2 is independent of the scale of measurement (i.e. I^2 can be compared between meta-analyses with SMD and OR outcome measures). An alternative formula for I^2 is also given in the original paper [47], where its interpretation becomes more apparent:

$$I^2 = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{\sigma}^2} \cdot 100\% \tag{1.6}$$

where $\hat{\sigma}^2$ is an estimate of the *typical* within-study variance:

$$\hat{\sigma}^2 = \frac{(k-1)\sum_{i=1}^{k} \hat{w}_i}{\left(\sum_{i=1}^{k} \hat{w}_i\right)^2 - \sum_{i=1}^{k} \hat{w}_i^2}$$

and $\hat{w}_i = 1/\hat{\sigma}_i^2$

$\hat{\sigma}^2$ is a measure mainly proposed to calculate I^2 via formula 1.6 and represents the estimated within-study variance if studies have equal σ_i^2. In reality, σ_i^2 vary from study-to-study and therefore formula 1.6 is not considered a true definition of I^2 [8]. I^2 is more commonly calculated from Q-based formula 1.5 for this reason and because of its convenient relationship with the DerSimonian-Laird method for estimating τ^2.

1.8 Prediction intervals

A prediction interval is used to measure the spread of true effect sizes (θ_i) in a meta-analysis with heterogeneous studies [48]. The interval can also predict the effect size of a new study to be included in the meta-analysis, hence the name 'prediction interval'. The prediction interval, as proposed by Higgins et al. [48], is:

$$\left[\hat{\theta} - t_{k-2}^{1-(\alpha/2)} \cdot \sqrt{\hat{\tau}^2 + V\left(\hat{\theta}\right)},\ \hat{\theta} + t_{k-2}^{1-(\alpha/2)} \cdot \sqrt{\hat{\tau}^2 + V\left(\hat{\theta}\right)}\right]$$

where $\hat{\theta}$ is estimated as in equation 1.2 with random-effects weights and $t_{k-2}^{1-(\alpha/2)}$ represents the $\left(1 - \frac{\alpha}{2}\right)$th percentile of the t-distribution with $k-2$ degrees of freedom. α is usually 0.05, leading to a 95% prediction interval.

The interval should not be confused with the confidence interval for $\hat{\theta}$, which I introduced in section 1.5.1.

1.9 Fixed effect vs. random-effects models

Using a fixed-effect model is appropriate when there is a contextual reason to believe study results are homogeneous. However, there is usually doubt whether the common effect assumption holds due to differences in study design or sampling population. Until recently, a common approach in such cases was to test for the presence of heterogeneity using the method based on the Q-statistic (section 1.7.1). The decision to use a fixed effect or random-effects model is then based on whether evidence for heterogeneity is statistically significant. However, simulation studies show this test has low power in meta-analyses typically seen in practice [44, 57] and therefore its use has been discouraged [8, 51]. A stronger argument against this test for heterogeneity is that meta-analyses are likely to contain heterogeneous studies even if there is no evidence from their results [40, 112].

The random-effects model is advocated when studies are potentially heterogeneous. The model reduces to the fixed-effect model when the heterogeneity variance is estimated to be zero, which occurs frequently using the DerSimonian-Laird method in practice [64]. The random-effects model is commonly used in meta-analysis and leads to more conservative results than the fixed-effect model when the heterogeneity variance parameter is positive by producing wider confidence intervals of the summary effect [8].

The random-effects model is not without criticisms. True study effect sizes are assumed to be normally distributed, which is unlikely given that bias (e.g. publication bias) is often present but undetectable [109]. Second, small studies are sometimes given a disproportionately high weight when studies are pooled using the inverse variance method [46]; small and large studies can have almost equal weight if the heterogeneity variance is comparatively high. Finally, estimating the heterogeneity variance parameter in a random-effects model presents many additional problems, as I will detail in the next section and address in the rest of this thesis.

1.10 An overview of the thesis

1.10.1 An overview of problems associated with heterogeneity variance estimation

Estimates of the heterogeneity variance in a random effects meta-analysis are usually imprecise [48, 120]. This is mainly because there are rarely sufficient numbers of studies contained in meta-analyses of health interventions [21]. Additional problems are apparent in meta-analyses with binary outcome measures such as the log odds ratio [78]. Binary outcome effect measures are correlated with their variance but the random-effects model assumes they are independent. Also, studies with rare events have large within-study variances, may require a continuity correction and contribute less to the summary effect [3].

Aside from the general problems of estimating the heterogeneity variance, the DerSimonian and Laird method in particular has been criticised. Simulation studies show the method underestimates heterogeneity variance when the underlying level of heterogeneity is high [78, 79, 124]. The method's bias is thought to be attributed to a failure of the methods only assumption: within-study variances used to calculate study weights are assumed to be known. The method has been proven theoretically

unbiased when this assumption holds [7, 124], but within-study variances can usually only be estimated in practice.

An estimate of the heterogeneity variance is required to calculate many other commonly reported statistics in meta-analysis, such as the summary effect and its confidence interval. The Wald-type method for producing these confidence intervals has been shown in simulation studies to be artificially narrow [61, 96]. The two main reasons for this are: (1) As already mentioned, DerSimonian-Laird on average underestimates the heterogeneity variance and (2) the Wald-type method assumes the heterogeneity variance is known, but is usually estimated and imprecise [96].

Alternative methods have since been proposed to estimate the heterogeneity variance and confidence interval of the summary effect. Some of these methods show more promising results in simulated data [16, 37, 40, 74, 93, 100, 102]. However, most of these studies are not comprehensive and recommend conflicting alternative methods. Therefore, there is currently no overall consensus as to which methods should be used in frequentist random-effects meta-analysis.

1.10.2 Focus of the thesis

My thesis reviews and compares methods to estimate the heterogeneity variance as defined in the standard random-effects model in section 1.5.2. This model in particular contains no covariates and assumes all observed between-study variance is random and cannot be explained. However, heterogeneity estimation methods can readily be applied to a random-effects model with covariates, so I discuss in the conclusion chapter whether my findings can be applied in this context. I focus almost solely on heterogeneity estimation in the frequentist framework, but allow exceptions for Bayesian methods that do not require complex integration and inexplicit prior distributions.

I focus specifically on methods for two-stage meta-analysis. A two-stage meta-analysis refers to the methods outlined in this chapter. It is defined as 'two-stage' be-

cause study summary data must first be extracted/derived before the meta-analysis is conducted. An alternative to a two-stage meta-analysis is a one-stage meta-analysis of individual participant data (IPD) [90, 104]. Many IPD meta-analyses combine each study dataset and perform a meta-analysis in one step, which allows for a full exploration of the study data [8]. However, the approach is usually more time-consuming and therefore not as common as the two-stage approach [119].

1.10.3 Aims

Expanding on the last section, the aims are to:

1. Conduct a comprehensive review of heterogeneity variance estimation methods currently available in the literature.
2. Assess the level of agreement between different heterogeneity variance methods in practice.
3. Compare the relative performance of methods in simulated data to establish which method(s) have the best properties.
4. Investigate the absolute performance of methods in simulated data to establish if and when all methods perform poorly.
5. Investigate whether any characteristics of meta-analyses can explain the properties of methods.
6. Recommend methods for random-effects meta-analysis and propose alternative strategies when all methods perform poorly.

In all chapters of this thesis where heterogeneity variance estimators are compared, I compare not only their estimates but their impact on the summary effect estimate and its confidence interval. As a consequence, the properties of methods to calculate these statistics are investigated as a secondary aim.

1.10.4 Structure of the thesis

The introductory chapters continue with a review of methods in two-stage random-effects meta-analysis. Chapter 2 contains a comprehensive review of heterogeneity variance estimation methods. Confidence interval methods for the summary effect follow in chapter 3. The latter is not intended to be a complete review, but an introduction to the confidence interval methods I have selected to use throughout the rest of the thesis.

To address the second aim of the thesis, chapter 4 compares methods in 12,894 empirical meta-analyses extracted from the *Cochrane Database of Systematic Reviews* (CSDR). Chapter 5 then reports a systematic review of previous studies that compare the properties of heterogeneity variance estimators in simulated meta-analysis data. Findings from this systematic review show that aims 3-5 have not been sufficiently addressed in the literature and so a further simulation study was required. Chapter 6 details the protocol of a further simulation study, designed in light of the limitations of current evidence identified in the systematic review. Chapters 7 and 8 then present the results from this simulation study.

Finally, chapter 9 concludes the thesis by summarising the main findings and discusses their implications for methods in random-effects meta-analysis. I make recommendations informed by findings from this thesis and from a review of the wider evidence base. Limitations of the heterogeneity variance methods and any limitations caused by scope of the thesis are discussed along with opportunities for further research.

Chapter 2

Methods for estimating the heterogeneity variance

2.1 Introduction

An estimate of the heterogeneity variance is often used to gauge inconsistency between study effects and is required to conduct a random-effects meta-analysis. I described in chapter 1 the DerSimonian-Laird method, commonly used to estimate the heterogeneity variance. This method has been widely criticised in the literature for producing negatively biased estimates [78, 101, 124] and therefore, many other methods have since been proposed. I present a comprehensive review of heterogeneity estimation methods in this chapter, highlight any methodological similarities and finish with an example to show these similarities in the context or a real meta-analysis.

This chapter was written concurrently with a systematic review of methods for estimating the heterogeneity variance that I co-authored [122]. I provided amendments after it was initially drafted by the first author. Methods in this paper were identified in a formal search of the PubMed database by the first author (Areti Angeliki Veroniki); the details of this search are given in appendix 9.5. All articles identified in this search are referenced in this chapter, which otherwise represents my own work.

The heterogeneity variance estimators in this chapter fall within a number of distinct approaches. I first introduce the method of moments approach in section 2.2, which consists of the DerSimonian-Laird estimator and a number of others. The maximum likelihood approach is introduced in section 2.5 and the Bayesian approach in section 2.6.

A number of estimators are proposed that only allow for positive heterogeneity variance estimates. These are proposed with the ethos described by Hartung and Makambi [40]: "...it may sometimes be difficult, in many applications, to accept zero as an estimate of the between-study variance when it is well known that there is some variation (albeit small) between groups/studies under consideration. This makes it desirable to derive a positive estimator for between-study variance". These positive estimators include Hartung-Makambi (section 2.3.1), those belonging to the Sidik-Jonkman approach (section 2.3.2) and all Bayesian estimators (section 2.6).

I introduce acronyms for each estimator in this chapter, which are used in the rest of the thesis. Table 2.1 on page 43 details all acronyms for reference. Also note, I use the general notation $\hat{w}_i$ for estimated study weights but these vary from estimator-to-estimator; I specify their functional form in each section.

Methods in this chapter were programmed in R (3.2.3) for use in all analyses in the rest of this thesis. I coded all methods that were not already available in the R package *metafor* [126].

2.2 Method of moments approach

This unified approach to heterogeneity estimation was initially proposed by Kacker [59], then DerSimonian and Kacker [24] explained the approach within a meta-analysis context. DerSimonian-Laird, Cochran's ANOVA and Paule-Mandel are estimators that pre-date this approach but have since been recognised within this unified identity [24]. I explain the commonality between these estimators before introducing them individually in sections 2.2.1 - 2.2.5.

The approach is based on the generalised Q-statistic:

$$Q_{MM} = \sum_{i=1}^{k} w_i \left(\hat{\theta}_i - \hat{\theta}\right)^2$$

$$\hat{\theta} = \sum_{i=1}^{k} \left(w_i \hat{\theta}_i\right) / \sum_{i=1}^{k} w_i \tag{2.1}$$

where k denotes the number of studies in the meta-analysis and study weights are denoted by w_i

We assume w_i takes no specific form for the general method of moments approach and may be known or estimated from the study data. Therefore, $\hat{\theta}$ is a generic

weighted average of study effects $\hat{\theta}_i$. Q_{MM} becomes the Q-statistic introduced in chapter 1 when $\hat{w}_i = 1/\hat{\sigma}_i^2$ is substituted for w_i.

The aim of this section is to derive a general method of moments formula for the heterogeneity variance. First, recall that $Var\left(\hat{\theta}_i\right)$ is made of the two variance components $\sigma_i^2 + \tau^2$ in a random-effects model and $\hat{\theta}$ is calculated as in equation 2.1. If we take the expected value of the unweighted squared error for a given study i [59]:

$$\begin{aligned} E\left[\left(\hat{\theta}_i - \hat{\theta}\right)^2\right] &= Var\left(\hat{\theta}_i - \hat{\theta}\right) = Var\left(\hat{\theta}_i\right) + Var\left(\hat{\theta}\right) - 2Cov\left(\hat{\theta}_i, \hat{\theta}\right) \\ &= \left(\sigma_i^2 + \tau^2\right) + \frac{\sum_{i=1}^k w_i^2\left(\sigma_i^2 + \tau^2\right)}{\left(\sum_{i=1}^k w_i\right)^2} - \frac{2w_i\left(\sigma_i^2 + \tau^2\right)}{\sum_{i=1}^k w_i} \end{aligned}$$

Therefore, the expected value of Q_{MM} is:

$$\begin{aligned} E\left[\sum_{i=1}^k w_i\left(\hat{\theta}_i - \hat{\theta}\right)^2\right] &= \sum\nolimits_{i=1}^k E\left[w_i\left(\hat{\theta}_i - \hat{\theta}\right)^2\right] \\ &= \sum_{i=1}^k w_i^2\left(\sigma_i^2 + \tau^2\right) + \frac{\sum_{i=1}^k w_i^2\left(\sigma_i^2 + \tau^2\right)}{\sum w_i} - 2\frac{\sum_{i=1}^k w_i^2\left(\sigma_i^2 + \tau^2\right)}{\sum w_i} \\ &= \sum_{i=1}^k w_i^2\left(\sigma_i^2 + \tau^2\right) - \frac{\sum_{i=1}^k w_i^2\left(\sigma_i^2 + \tau^2\right)}{\sum w_i} \end{aligned}$$

To derive a formula for the estimated heterogeneity variance ($\hat{\tau}^2$), equate the expected value to its observed value:

$$\begin{aligned} \sum_{i=1}^k w_i\left(\hat{\theta}_i - \hat{\theta}\right)^2 &= \sum_{i=1}^k w_i\sigma_i^2 + \hat{\tau}^2\sum_{i=1}^k w_i - \frac{\sum_{i=1}^k w_i^2\sigma_i^2}{\sum_{i=1}^k w_i} - \hat{\tau}^2\frac{\sum_{i=1}^k w_i^2}{\sum_{i=1}^k w_i} \\ &= \hat{\tau}^2\left(\sum_{i=1}^k w_i - \frac{\sum_{i=1}^k w_i^2}{\sum_{i=1}^k w_i}\right) + \sum_{i=1}^k w_i\sigma_i^2 - \frac{\sum_{i=1}^k w_i^2\sigma_i^2}{\sum_{i=1}^k w_i} \Longleftrightarrow \end{aligned}$$

$$\hat{\tau}^2 = \frac{\sum_{i=1}^{k} w_i \left(\hat{\theta}_i - \hat{\theta}\right)^2 - \left(\sum_{i=1}^{k} w_i \sigma_i^2 - \frac{\sum_{i=1}^{k} w_i^2 \sigma_i^2}{\sum_{i=1}^{k} w_i}\right)}{\sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i}} \tag{2.2}$$

The approach assumes that σ_i^2 are known and these are usually replaced by estimates $\hat{\sigma}_i^2$ in the above formula [6]. All method of moments estimators can be derived from the formula 2.2, but with w_i taking different functional forms.

None of the proposed study weights ensure $\hat{\tau}^2 > 0$ in formula 2.2. Therefore, for all method of moments estimators, $\hat{\tau}^2$ is truncated to zero whenever it would otherwise be negative. Method of moments estimators make no assumption about the distribution of θ_i, unlike most other methods introduced in this chapter [62, 93, 94].

2.2.1 DerSimonian-Laird (DL)

The DerSimonian-Laird (DL) estimator [25] was introduced in the introduction chapter. I showed that DL is derived from the Q-statistic, which is the same as Q_{MM} with study weights $\hat{w}_i = 1/\hat{\sigma}_i^2$. With these weights, the method of moments formula 2.2 derived earlier for $\hat{\tau}^2$ becomes:

$$\hat{\tau}^2_{DL} = max \left\{ 0, \frac{\sum_{i=1}^{k} (1/\hat{\sigma}_i^2) \left(\hat{\theta}_i - \hat{\theta}_{DL}\right)^2 - (k-1)}{\sum_{i=1}^{k} (1/\hat{\sigma}_i^2) - \frac{\sum_{i=1}^{k} \left(1/\hat{\sigma}_i^2\right)^2}{\sum_{i=1}^{k} \left(1/\hat{\sigma}_i^2\right)}} \right\}$$

where $\hat{\theta}_{DL} = \sum_{i=1}^{k} \left(\hat{w}_i \hat{\theta}_i\right) / \sum_{i=1}^{k} \hat{w}_i$

Kontopantelis et al. [64] introduced an alternative DL estimator (DL_P), with a cut-off value of 0.01 to ensure all estimates are positive.

2.2.2 Cochran's ANOVA (CA)

Cochran's ANOVA (CA) estimator [18] was proposed in 1954 before being introduced into the random-effects meta-analysis context by Hedges and Olkin [42]. Kacker [59] then showed CA is part of the general method of moments approach. CA assigns equal weight to each study, most commonly $w_i = 1/k$ [24], but any positive constant would produce identical heterogeneity variance estimates.

Weights $w_i = 1/k$ are substituted into formula 2.2:

$$\hat{\tau}^2_{CA} = max\left\{0,\ \frac{1}{k-1}\sum_{i=1}^{k}\left(\hat{\theta}_i - \hat{\theta}_{CA}\right)^2 - \frac{1}{k}\sum_{i=1}^{k}\hat{\sigma}^2_i\right\} \tag{2.3}$$

where $\hat{\theta}_{CA} = \sum_{i=1}^{k}\hat{\theta}_i/k$

Cochran [18] originally derived the estimator by rearranging $\hat{\theta}$'s variance components:

$$\tau^2 \ = \ Var\left(\hat{\theta}\right) - \sigma^2$$

σ^2 denotes the typical within-study variance, estimated assuming equal variances: $\hat{\sigma}^2 = \sum_{i=1}^{k}\hat{\sigma}^2_i/k$. $Var\left(\hat{\theta}\right)$ is estimated using the unweighted formula for a sampling variance. This equates to formula 2.3 above.

2.2.3 Non-parametric Bootstrap DerSimonian-Laird ($\mathbf{DL_B}$)

Kontopantelis et al. [64] proposed a bootstrap version of the standard DL estimator ($\mathrm{DL_B}$) to reduce the number of zero τ^2 estimates. A bootstrap estimate of τ^2 is calculated in four steps:

1. k studies are sampled randomly with replacement (the same number as in the meta-analysis)

2. $\hat{\tau}^2_{DL}$ is calculated for the sample.

3. 1-2 is repeated 10,000 times in order to derive a distribution for $\hat{\tau}^2_{DL}$. Note, fewer samples may be required if k is small.

4. The mean of this distribution is the bootstrap estimate of τ^2.

The bootstrap method could trivially be applied to other τ^2 estimators [26].

2.2.4 Paule-Mandel (PM)

The Paule-Mandel estimator (PM) [80] was originally proposed in the more general context of combining measurements from different experiments. The method has since been introduced into the meta-analysis framework [9, 24].

Weights of each study take the form $\hat{w}_i = 1/\left(\hat{\sigma}^2_i + \hat{\tau}^2_{PM}\right)$. A non-closed form expression for $\hat{\tau}^2_{PM}$ can be derived by substituting these weights $\hat{w}_i$ into formula 2.2:

$$\hat{\tau}^2_{PM} = \frac{\sum_{i=1}^{k} \hat{w}_i(\hat{\theta}_i - \hat{\theta}_{PM})^2 - \left(\sum_{i=1}^{k} \hat{w}_i\sigma^2_i - \left(\sum_{i=1}^{k} \hat{w}^2_i\sigma^2_i\right) / \left(\sum_{i=1}^{k} \hat{w}_i\right)\right)}{\sum_{i=1}^{k} \hat{w}_i - \left(\sum_{i=1}^{k} \hat{w}^2_i\right) / \left(\sum_{i=1}^{k} \hat{w}_i\right)} \tag{2.4}$$

where $\hat{\theta}_{PM} = \sum_{i=1}^{k} \left(\hat{w}_i\hat{\theta}_i\right) / \sum_{i=1}^{k} \hat{w}_i$

As the heterogeneity variance parameter is included in the study weights, $\hat{\tau}^2_{PM}$ is found by a process of iteration until convergence. The initial estimate of $\hat{\tau}^2_0 = 0$ is commonly used to begin the process. If at any step the estimate is negative, then set $\hat{\tau}^2_{PM} = 0$. Rukhin et al. [94] demonstrated that there is always just one solution and the process always converges irrespective of the initial estimate.

Alternatively, a neater expression than formula 2.4 can be derived if we substitute $\hat{w}_i$ into Q_{MM}. Since $\hat{w}_i$ are random-effects weights that account for the total study variance, we constrain Q_{MM} to asymptotically follow the χ^2_{k-1}-distribution with expected value $k-1$ on the assumption that $\hat{\theta}_i$ are normally distributed around θ_i. Equating Q_{MM} in this case to this expected value:

$$\sum_{i=1}^{k} \frac{\left(\hat{\theta}_i - \hat{\theta}_{PM}\right)^2}{\hat{\sigma}_i^2 + \hat{\tau}^2_{PM}} = k - 1$$

However, finding $\hat{\tau}^2_{PM}$ using this expression involves an iterative process that is less intuitive [24, 80].

PM is also known as the empirical Bayes estimator because they were thought to be separate methods until Rukhin [93] noted their equivalence. Despite this alternative name, PM is not an empirical Bayes approach to heterogeneity variance estimation. It gets this name from Morris [77], who used the PM heterogeneity variance estimator in an empirical Bayes approach to estimate θ.

2.2.5 Two-step versions of Paule-Mandel ($\mathbf{PM_{CA}}$ & $\mathbf{PM_{DL}}$)

DerSimonian and Kacker [24] introduced two alternative versions of PM that do not require complete iteration to convergence. They use the same random-effects study weights as PM but iteration is restricted to two-steps with initial estimates $\hat{\tau}^2_0 = \hat{\tau}^2_{CA}$ and $\hat{\tau}^2_0 = \hat{\tau}^2_{DL}$. For the former, the general method of moments formula 2.2 becomes the following closed form expression:

$$\hat{\tau}^2_{PM_{CA}} = \frac{\sum_{i=1}^{k} \hat{w}_i(\hat{\theta}_i - \hat{\theta}_{CA})^2 - \left(\sum_{i=1}^{k} \hat{w}_i\sigma_i^2 - \left(\sum_{i=1}^{k} \hat{w}_i^2\sigma_i^2\right) / \left(\sum_{i=1}^{k} \hat{w}_i\right)\right)}{\sum_{i=1}^{k} \hat{w}_i - \left(\sum_{i=1}^{k} \hat{w}_i^2\right) / \left(\sum_{i=1}^{k} \hat{w}_i\right)} \quad (2.5)$$

where $\hat{w}_i = (\hat{\sigma}_i^2 + \hat{\tau}_{CA}^2)^{-1}$ and $\hat{\theta}_{CA}$ is defined in section 2.2.2

For the latter two-step estimator, weights are $\hat{w}_i = (\hat{\sigma}_i^2 + \hat{\tau}_{DL}^2)^{-1}$ and $\hat{\theta}_{DL}$ is defined in section 2.2.1.

The main criticism of $\mathrm{PM_{CA}}$ and $\mathrm{PM_{DL}}$ is that restricting PM to two iterative steps is unnecessary and not considered statistically optimal. PM heterogeneity variance estimates can easily be computed using reliable iterative techniques widely available in many statistical software packages.

2.3 Non-truncated moments-based approaches

2.3.1 Hartung-Makambi (HM)

As mentioned in section 2.2, method of moments estimators allow for negative estimates of the heterogeneity variance and should be truncated at zero. Hartung and Makambi [40] proposed a correction to $\hat{\tau}_{DL}^2$ so that $\hat{\tau}^2$ is always positive and truncation is not required. Recall that:

$$\hat{\tau}_{DL}^2 = max\left\{0,\ \frac{Q-(k-1)}{c}\right\}$$

where $c = \sum_{i=1}^{k} \hat{w}_i - \left(\sum_{i=1}^{k} \hat{w}_i^2 / \sum_{i=1}^{k} \hat{w}_i\right)$, $\hat{w}_i = 1/\hat{\sigma}_i^2$ and Q is the statistic defined in chapter 1.

$\hat{\tau}_{DL}^2$ is negative and must be truncated to zero when $Q < (k-1)$. The estimator is derived by taking the first term of $\hat{\tau}_{DL}^2$ (Q/c), which is always positive and applying a positive multiplicative correction factor denoted by ε. This correction factor accounts for the bias introduced as a result of excluding the term $(k-1)/c$. The estimator takes the form $\hat{\tau}_{HM}^2 = \varepsilon \cdot Q/c$, where:

$$\varepsilon = \frac{Q}{2\left(k-1\right)+Q}$$

and therefore:

$$\hat{\tau}^2_{HM} = \varepsilon \cdot \frac{Q}{c} = \frac{Q}{2\left(k-1\right)+Q}\left(\frac{Q}{c}\right) = \frac{Q^2}{c\left(2\left(k-1\right)+Q\right)}$$

It is not clear in the original paper [40] why ε takes this form or the extent that Q/c is biased before a correction factor is applied.

2.3.2 Sidik-Jonkman (SJ)

This method was first introduced by Sidik and Jonkman [101] and yields only positive estimates. The method is derived from the standard formula for $Var\left(\hat{\theta}\right)$ (formula 1.3 in the introduction chapter) with study weights defined as:

$$\hat{w}_i = \frac{1}{(\hat{\sigma}^2_i/\hat{\tau}^2)+1} = \frac{\hat{\tau}^2}{\hat{\sigma}^2_i+\hat{\tau}^2}$$

$Var\left(\hat{\theta}\right)$ can be re-expressed in terms of the new weights $\hat{w}_i$:

$$Var\left(\hat{\theta}\right) = \frac{\hat{\tau}^2}{\sum_{i=1}^{k}\hat{w}_i} \tag{2.6}$$

This method also uses an alternative weighted estimator of $Var\left(\hat{\theta}\right)$ proposed by Hartung [38]:

$$Var_{HK}\left(\hat{\theta}\right) = \frac{\sum_{i=1}^{k}\hat{w}_i\left(\hat{\theta}_i-\hat{\theta}\right)^2}{(k-1)\sum_{i=1}^{k}\hat{w}_i}$$

If we equate $Var_{HK}\left(\hat{\theta}\right)$ and $Var\left(\hat{\theta}\right)$ from equation 2.6, then:

$$\hat{\tau}^2 = \frac{1}{k-1}\sum_{i=1}^{k}\frac{1}{(\hat{\sigma}_i^2/\hat{\tau}^2)+1}\left(\hat{\theta}_i - \hat{\theta}\right)^2$$

The formula above is naturally iterative, so Sidik and Jonkman [101] proposed a two-step process with initial estimate $\hat{\tau}_0^2 = \frac{1}{k}\sum_{i=1}^{k}\left(\hat{\theta}_i - \hat{\theta}_{CA}\right)^2$. The SJ estimator is defined as:

$$\hat{\tau}_{SJ}^2 = \frac{1}{k-1}\sum_{i=1}^{k}\frac{1}{(\hat{\sigma}_i^2/\hat{\tau}_0^2)+1}\left(\hat{\theta}_i - \hat{\theta}_{SJ}\right)^2 \tag{2.7}$$

where $\hat{\theta}_{SJ}$ is the weighted least squares estimate of θ with weights $\hat{w}_i = 1/\left((\hat{\sigma}_i^2/\hat{\tau}_0^2)+1\right)$ and $\hat{\theta}_{CA}$ is the unweighted estimate of θ from section 2.2.2.

Study weights $\hat{w}_i$ are undefined when $\hat{\tau}_0^2 = 0$, which occurs in the unlikely case when all $\hat{\theta}_i$ are equal. In this case, set $\hat{\tau}_{SJ}^2 = 0$.

Sidik and Jonkman [101] noted that alternative $\hat{\tau}_0^2$ estimates may lead to an estimator with better properties. Therefore, Sidik and Jonkman [102] proposed $\hat{\tau}_0^2 = max\left(0.01, \hat{\tau}_{CA}^2\right)$ in a follow-up paper; I denote the resulting estimator as $\mathrm{SJ_{CA}}$. As with the original estimator, $\hat{\tau}_{SJ_{CA}}^2$ is a two-step estimator that is simple to compute and always results in a positive estimate of the heterogeneity variance.

These estimators have methodological similarities with PM, introduced in section 2.2.4. Their weights are equivalent to the PM random-effects study weights, multiplied by the constant $\hat{\tau}^2$; this transformation ensures SJ estimators are positive.

2.4 Hunter-Schmidt (HS)

The Hunter-Schmidt estimator (HS) [53] is derived by expressing the variance components for $\hat{\theta}$ as $Var\left(\hat{\theta}\right) = \tau^2 + \sigma^2$. A 'typical' within-study variance from all studies

i is denoted by σ^2. Weighted unbiased estimators of $Var\left(\hat{\theta}\right)$ and σ^2 are substituted into the variance components to derive the estimator of heterogeneity:

$$\hat{\tau}^2_{HS} = max\left\{0,\ \frac{\sum_{i=1}^{k}\hat{w}_i\left(\hat{\theta}_i - \hat{\theta}\right)^2}{\sum_{i=1}^{k}\hat{w}_i} - \frac{\sum_{i=1}^{k}\hat{w}_i\sigma_i^2}{\sum_{i=1}^{k}\hat{w}_i}\right\} \tag{2.8}$$

with fixed-effect weights $\hat{w}_i = 1/\hat{\sigma}_i^2$

Using these weights, the estimator can be re-expressed as [96]:

$$\hat{\tau}^2_{HS} = max\left\{0,\ \frac{\sum_{i=1}^{k}\hat{w}_i\left(\hat{\theta}_i - \hat{\theta}\right)^2 - k}{\sum_{i=1}^{k}\hat{w}_i}\right\} \tag{2.9}$$

Other weights that have been proposed include [124]; (1) $w_i = 1/\left(n_{1i} + n_{2i}\right)$ where n_{1i} and n_{2i} are the sample sizes in groups one and two of study i and (2) $w_i = 1/k$. A method for deriving the CA estimator (see section 2.2.2) also directly involves splitting the variance components, the only difference being that CA uses unweighted estimates of $Var\left(\hat{\theta}\right)$ and σ^2.

2.5 Maximum likelihood approach

Maximum likelihood and restricted maximum liklelihood (REML) estimators can both be derived from a log-likelihood function derived from the probability density function of $\hat{\theta}_i \sim N\left(\theta, \sigma_i^2 + \tau^2\right)$ [37]. Hence, it is assumed that θ_i and $\hat{\theta}_i$ are normally distributed around the central parameter θ, unlike the estimators in the method of moments approach.

2.5.1 Maximum likelihood (ML)

The log likelihood function for the maximum likelihood estimator is:

$$l_{ML}\left(\theta,\tau^2\right) = -\frac{1}{2}\sum_{i=1}^{k} ln\left(2\pi\left(\sigma_i^2+\tau^2\right)\right) - \frac{1}{2}\sum_{i=1}^{k}\frac{\left(\theta_i-\theta\right)^2}{\sigma_i^2+\tau^2} \tag{2.10}$$

The maximum likelihood estimate of τ^2 is the value that maximises $l_{ML}\left(\theta,\tau^2\right)$. The maximum can be found by partially differentiating l_{ML} with respect to τ^2 and equating this to zero. This leads to:

$$\hat{\tau}^2_{ML} = max\left\{0,\ \frac{\sum_{i=1}^{k}\hat{w}_i^2\left(\left(\hat{\theta}_i-\hat{\theta}_{ML}\right)^2-\hat{\sigma}_i^2\right)}{\sum_{i=1}^{k}\hat{w}_i^2}\right\} \tag{2.11}$$

where $\hat{w}_i = \left(\hat{\sigma}_i^2+\hat{\tau}^2_{ML}\right)^{-1}$ and $\hat{\theta}_{ML}$ is the maximum likelihood estimate of θ. The formula for $\hat{\theta}_{ML}$ is derived by partially differentiating l_{ML} with respect to θ:

$$\hat{\theta}_{ML} = \frac{\sum_{i=1}^{k}\hat{w}_i\hat{\theta}_i}{\sum_{i=1}^{k}\hat{w}_i} \tag{2.12}$$

Maximum likelihood estimates are calculated by solving (2.11) and (2.12) simultaneously and iteratively, starting with an initial estimate $\hat{\tau}_0^2$. There are many iterative methods used in maximum likelihood including the Newton-Raphson method, Fisher's scoring algorithm and the simplex algorithm [58]. If at any step $\hat{\tau}^2 < 0$, then the process of iteration stops and we evaluate $\hat{\tau}^2_{ML} = 0$. For any iteration method, convergence is not guaranteed [64]. In this thesis, I use $\hat{\tau}_0^2 = \hat{\tau}^2_{CA}$ and use Fisher's scoring algorithm.

The method assumes that within-study variances and θ are known, when in reality they must be estimated from study data [102]. Cheung [14] suggested this is likely to lead to an underestimate of τ^2.

2.5.2 Restricted maximum likelihood (REML)

To derive the REML estimator of τ^2, the log-likelihood function (2.10) is transformed so that it excludes the summary effect parameter θ [41]. In doing so, REML avoids making the assumption that θ is known and is therefore thought to be an improvement on the ML estimator [124]. This produces the following log-likelihood function:

$$l_{REML}\left(\tau^2\right) = -\frac{k}{2}ln\left(2\pi\right) - \frac{1}{2}\sum_{i=1}^{k} ln\left(\sigma_i^2 + \tau^2\right) - \frac{1}{2}\sum_{i=1}^{k}\frac{\left(\hat{\theta}_i - \hat{\theta}\right)^2}{\sigma_i^2 + \tau^2} - \frac{1}{2}ln\left(\sum_{i=1}^{k}\frac{1}{\sigma_i^2 + \tau^2}\right)$$

To derive a formula for $\hat{\tau}^2$, we partially differentiate l_{REML} with respect to τ^2 and setting this differential to zero. This results in the following equation:

$$\hat{\tau}^2_{REML} = max\left\{0,\ \frac{\sum_{i=1}^{k}\hat{w}_i^2\left(\left(\hat{\theta}_i - \hat{\theta}_{REML}\right)^2 - \hat{\sigma}_i^2\right)}{\sum_{i=1}^{k}\hat{w}_i^2} + \frac{1}{\sum_{i=1}^{k}\hat{w}_i}\right\}$$

where $\hat{\theta}_{REML} = \sum_{i=1}^{k}\left(\hat{w}_i\hat{\theta}_i\right) / \sum_{i=1}^{k}\hat{w}_i$ and $\hat{w}_i = \left(\hat{\sigma}_i^2 + \hat{\tau}^2_{REML}\right)^{-1}$

$\hat{\tau}^2_{REML}$ is found by the same iterative process as $\hat{\tau}^2_{ML}$ from section 2.5.1 and convergence is also not guaranteed.

2.5.3 Approximate restricted maximum likelihood (ARML)

The approximate restricted maximum likelihood (ARML) estimator, is thought to give similar estimates as REML [2, 77, 124]. Heterogeneity variance estimates are calculated by iteration and the following formula:

$$\hat{\tau}^2_{ARML} = max \left\{ 0, \frac{\sum_{i=1}^{k} \hat{w}_i^2 \frac{k}{k-1} \left(\left(\hat{\theta}_i - \hat{\theta}_{ARML} \right)^2 - \hat{\sigma}_i^2 \right)}{\sum_{i=1}^{k} \hat{w}_i^2} \right\}$$

where the weights are defined as $\hat{w}_i = \left(\hat{\sigma}_i^2 + \hat{\tau}^2_{ARML} \right)^{-1}$.

ARML is a simplified version of REML, with the additional assumption that sampling variances σ_i^2 for all studies are equal. Regardless, the process of finding $\hat{\tau}^2_{ARML}$ involves iteration and has no obvious benefit over REML. When all σ_i^2 are equal, both give identical estimates and only differ slightly otherwise although this is difficult to prove algebraically [124].

2.6 The Bayesian approach

I described the general framework for Bayesian meta-analysis in the introduction chapter (section 1.5.4). Bayesian heterogeneity variance estimators are based on this model and allow for prior beliefs of model parameters to be combined with meta-analysis data. This is the defining difference from the frequentist estimators outlined in the rest of this chapter. I introduce the full Bayesian approach in section 2.6.1. Following this, I introduce a series of semi-Bayesian τ^2 estimators that are more simple to compute including approximate Bayes (AB), empirical Bayes (EB), Bayes modal (BM) and estimators proposed by Rukhin [93].

2.6.1 Full Bayesian

The full Bayesian approach estimates the heterogeneity variance simultaneously with all other parameters of interest in the model. In doing so, it can account for uncertainty of these parameters [106]. In a Bayesian random-effect model with no covariates, we can define prior distributions for τ^2 and θ:

$$\tau^2 \sim p_1(\varphi_1)$$

$$\theta \sim p_2(\varphi_2)$$

where p_1 and p_2 are the chosen probability distributions with fixed parameter vectors φ_1 and φ_2.

Prior distributions and their fixed parameters vary between meta-analyses in practice and are chosen based on external evidence, expert opinion or they are vague to reflect a lack of prior knowledge [111]. Therefore, it is not possible to define a distinct full Bayesian method. Possible assumed distributions for τ^2 include the inverse gamma, uniform or normal [66, 88]. A normal distribution for θ_i is often assumed and is therefore the chosen prior distribution for θ [111].

The aim of this approach is to calculate a joint posterior distribution for τ^2 and θ by combining prior distributions with meta-analysis data. The posterior distribution is derived by Markov Chain Monte Carlo (MCMC) methods such as Gibb's sampler [106]. This requires specialist software such as WinBUGS [72]. From the joint posterior distribution, expected values and credibility intervals for τ^2 and θ can be extracted.

2.6.2 Approximate Bayes Estimator (AB)

The approximate Bayes estimator (AB) was originally proposed within the context of sequential meta-analysis [50]. It does not require any form of Gibbs sampling or process of iteration and is therefore more simple to compute than full Bayes.

The prior for τ^2 follows the inverse-gamma distribution $\Gamma^{-1}(\eta, \lambda)$, with parameters η and λ defining the shape and spread of the distribution respectively and zero probability of $\tau^2 < 0$. The inverse-gamma distribution has the following p.d.f:

$$p\left(\tau^2;\eta,\lambda\right)=\frac{\lambda^{\eta}}{\Gamma\left(\eta\right)}\left(\tau^2\right)^{-\eta-1}exp\left(-\frac{\lambda}{\tau^2}\right)$$

where Γ represents the gamma function.

The underlying effects of each study i are assumed known, i.e. $\sigma_i^2=0$. Higgins et al. [50] suggest this has minimal impact on $\hat{\tau}^2$. By making this assumption, it follows that the posterior distribution will also be an inverse-gamma distribution with parameters $\eta=\eta_0+(k/2)$ and $\lambda=\lambda_0+(k\tau^2/2)$ [35] . τ^2 in this case represents the heterogeneity variance from the data, for which Higgins et al. [50] suggests using the DL estimate (section 2.2.1). A posterior estimate of τ^2 can be derived by substituting the formulas for the posterior parameters η and λ into the formula for the mean of an inverse-gamma distribution:

$$\hat{\tau}_{AB}^2=\frac{\lambda}{\eta-1}=\frac{2\lambda_0+k\hat{\tau}_{DL}^2}{2\left(\eta_0-1\right)+k}$$

The prior distribution for τ^2 has mean $\lambda_0/\left(\eta_0-1\right)$, implying the expected value of λ_0 is $\hat{\tau}_0^2\left(\eta_0-1\right)$. This can be substituted into the formula for $\hat{\tau}_{AB}^2$ above:

$$\hat{\tau}_{AB}^2=\frac{2\tau_0^2\left(\eta_0-1\right)+k\hat{\tau}_{DL}^2}{2\left(\eta_0-1\right)+k}$$

This last step is carried out because it is often easier to define a prior value for τ_0^2 than the spread parameter λ_0. To calculate $\hat{\tau}_{AB}^2$, we must provide two of three prior values for τ_0^2, η_0 and λ_0. In the context of sequential meta-analysis in the original publication [50], τ_0^2 is the estimate of τ^2 from the previous update to the meta-analysis. Outside of this context, τ_0^2 can represent our best estimate from prior beliefs.

2.6.3 Bayes modal (BM)

Bayes Modal (BM) can estimate τ^2 numerically without the need for MCMC methods. It imposes a gamma prior distribution for τ [16, 17]:

$$p(\tau;\eta,\lambda) = \frac{1}{\Gamma(\alpha)}\tau^{\alpha-1}e^{-\lambda\tau}$$

α and λ are the shape and scale parameters defined from prior information. Chung et al. [16] suggest using $\alpha = 2$ and $\lambda \approx 0$ for a vague prior. The gamma distribution is chosen because it has the property $p(\tau = 0;\eta,\lambda) = 0$ for any α or λ, thus avoiding zero estimates of τ^2 from the posterior. The density function of the posterior distribution can be derived if we assume an improper uniform prior for θ:

$$p(\tau,\theta) = l_{ML}\left(\theta,\tau^2\right) + (\alpha-1)\,log\tau - \lambda\tau + c$$

where c is an undefined constant and $l_{ML}(\theta,\tau^2)$ is the log-likelihood function (2.10).

Software packages are available to finding estimates of (τ,θ) that maximise the above equation such as *lmer* in R and *gllamm* in Stata [17]. The BM estimator can alternatively be considered a maximum likelihood approach with a penalty imposed to avoid zero estimates; the above log-likelihood is that of $l_{ML}(\theta,\tau^2)$ with added terms [17].

2.6.4 Rukhin's approach

Rukhin [93] proposed two semi-Bayesian heterogeneity variance estimators. These differ from estimators derived from a more typical Bayesian approach, which involves specifying prior distributions for the unknown parameters and requires MCMC methods fit observed data to the model. Rukhin's estimators are more simple to compute and only require a fixed prior estimate of τ^2, denoted $\hat{\tau}_0^2$.

Rukhin's estimators are derived from a generalised version of method of moments from section 2.2. He first explicitly derives the formula for $Var\left(\hat{\tau}^2\right)$ under this unified approach. Then, his general formula finds $\hat{\tau}^2$ such that $Var\left(\hat{\tau}^2\right)$ is locally minimised around the prior estimate of τ^2. I refer the reader to the original paper for a detailed derivation of this approach [93]. The general formula for Rukhin's heterogeneity variance estimators is:

$$\hat{\tau}^2_{RB} = \frac{\sum_{i=1}^{k}\left(\hat{\theta}_i - \hat{\theta}\right)^2}{k+1} + \frac{\left(\sum_{i=1}^{k}\left(n_{1i}+n_{2i}\right)-k\right)\left(2k\hat{\tau}_0^2 - (k-1)\sum_{i=1}^{k}\hat{\sigma}_i^2\right)}{\left(\sum_{i=1}^{k}\left(n_{1i}+n_{2i}\right)-k+2\right)k\left(k+1\right)} \quad (2.13)$$

where n_{1i} and n_{2i} are the sample sizes in intervention groups one and two, $\hat{\theta} = \sum_{i=1}^{k}\left(\hat{\theta}_i\hat{w}_i\right)/\sum_{i=1}^{k}\hat{w}_i$ and $\hat{w}_i = \left(\hat{\sigma}_i^2 + \hat{\tau}_0^2\right)^{-1}$.

Rukhin [93] proposed two formulae for $\hat{\tau}_0^2$:

1. $\hat{\tau}_0^2 = 0$, which leads to following heterogeneity variance estimator (B0):

$$\hat{\tau}^2_{B0} = \frac{\sum_{i=1}^{k}\left(\hat{\theta}_i - \hat{\theta}\right)^2}{k+1} + \frac{\left(\sum_{i=1}^{k}\left(n_{1i}+n_{2i}\right)-k\right)(k-1)\left(\sum_{i=1}^{k}\hat{\sigma}_i^2\right)}{\left(\sum_{i=1}^{k}\left(n_{1i}+n_{2i}\right)-k+2\right)k\left(k+1\right)}$$

2. $\hat{\tau}_0^2 = 0.5\left(k-1\right)\sum_{i=1}^{k}\left(\hat{\sigma}_i^2\right)/k$, which leads to (BP):

$$\hat{\tau}^2_{BP} = \frac{\sum_{i=1}^{k}\left(\hat{\theta}_i - \hat{\theta}\right)^2}{k+1}$$

For a given k, $\hat{\tau}^2_{BP}$ is a fixed proportion of the total sample variance $Var\left(\hat{\theta}\right) = \left(k-1\right)^{-1}\sum_{i=1}^{k}\left(\hat{\theta}_i - \hat{\theta}\right)^2$. There is little logic for this relationship, so BP is unlikely to have good properties. Rukhin [93] suggested this prior for τ^2 only to simplify the general formula 2.13.

Another estimator, which forms a part of the same approach, assumes σ_i^2 to be known. This estimator is given by the formula:

$$\hat{\tau}^2_{SB} = \frac{\sum_{i=1}^{k}\left(\hat{\theta}_i - \hat{\theta}\right)^2}{k+1} + \frac{2\hat{\tau}_0^2 k - (k-1)\sum_{i=1}^{k}\hat{\sigma}_i^2}{k\,(k+1)}$$

Prior estimates $\hat{\tau}_0^2$ must be specified to calculate $\hat{\tau}^2_{SB}$, but Rukhin [93] made no specific suggestions.

2.7 Malzahn, Böhning and Holling (MBH)

Malzahn et al. [74] proposed a τ^2 estimator that makes no assumption on how θ_i are distributed. The estimator can only be applied to SMD meta-analyses, the outcome measure I introduced in section 1.3.1. First, formulae for the typical variance of θ_i are derived under both fixed-effect and random-effects assumptions. An estimate of τ^2 is then derived by taking the difference between the two variances, this results in:

$$\tau^2_{MBH} = \left(\frac{1}{k-1}\right)\sum_{i=1}^{k}(1-K_i)\left(\hat{\theta}_i - \hat{\theta}_{CA}\right)^2 - \frac{1}{k}\sum_{i=1}^{k}\left(\frac{n_{1i}+n_{2i}}{n_{1i}n_{2i}}\right) - \frac{1}{k}\sum_{i=1}^{k}\left(K_i\hat{\theta}_i^2\right)$$

where $K_i = 1 - ((N_i - 2)/N_i J_i^2)$, $N_i = n_{1i} + n_{2i} - 2$, $\hat{\theta}_{CA} = \sum_{i=1}^{k}\hat{\theta}_i/k$ and $J_i = 1 - 3/(4N_i - 1)$. J_i is the bias correction factor proposed by Hedges [43] and already introduced in section 1.3.1.

2.8 Alternative within-study variance estimates

All heterogeneity variance estimators in this chapter are dependent on within-study variances. So far in this chapter, I assumed within-study variances have been calculated by conventional means as introduced in chapter 1. Two other methods have

been proposed to improve within-study variance estimates, which then in theory lead to improved heterogeneity variance estimates. All alternative methods identified and presented in this section can only be applied to binary outcome meta-analyses.

2.8.1 Bhaumik et al (2012)

Bhaumik et al. [3] proposed an alternative method of calculating within-study variances in meta-analyses with an odds ratio outcome measure and rare events. The method works by allowing studies to *borrow strength* from the other studies in a meta-analysis and in doing so assumes within-study variances are equal. Within-study variances are estimated by the following:

$$\begin{aligned}\hat{\sigma}_i^2 &= \frac{1}{n_{1i}+1}\left[exp\left(-odds_2 - \hat{\theta}_{CA} + \frac{\hat{\tau}^2}{2}\right) + 2 + exp\left(odds_2 + \hat{\theta}_{CA} + \frac{\hat{\tau}^2}{2}\right)\right] \\ &\quad + \frac{1}{n_{2i}+1}\left[exp\left(-odds_2\right) + 2 + exp\left(odds_2\right)\right]\end{aligned}$$

where $odds_2$ are the observed odds of an event in the group 2 of study i, $\hat{\theta}_{CA}$ is the equally-weighted combined effect estimate as in section 2.2.2 with a continuity correction to deal with zero events 2.2.2.

In the original paper [3], $\hat{\tau}^2$ is calculated using the PM estimator from section 2.2.4. However, these alternate estimates of σ_i^2 could be applied to any other τ^2 estimator.

2.8.2 Berkey et al (1995) and Knapp & Hartung (2003)

Berkey et al. [2] proposed a smoothed estimator of the within-study variances in meta-analyses with a relative risk outcome measure. The idea was introduced to reduce correlation between the relative risk estimate and the within-study variances:

$$\hat{\sigma}_i^2 = \frac{1}{k \cdot n_{1i}} \sum_{i=1}^{k} \left(\frac{n_{1i} - a_i}{a_i} \right) + \frac{1}{k \cdot n_{2i}} \sum_{i=1}^{k} \left(\frac{n_{2i} - c_i}{c_i} \right)$$

where a_i and c_i are the number of events in treatment groups 1 and 2 respectively.

Knapp and Hartung [61] proposed an adapted version that includes a continuity correction:

$$\hat{\sigma}_i^2 = \frac{1}{k \cdot n_{1i}} \sum_{i=1}^{k} \left(\frac{n_{1i} - a_i + 0.5}{a_i + 0.5} \right) + \frac{1}{k \cdot n_{2i}} \sum_{i=1}^{k} \left(\frac{n_{2i} - c_i + 0.5}{c_i + 0.5} \right)$$

These estimators cannot be applied to other binary outcome meta-analyses, where the outcome is for example an odds ratio [86].

2.9 A summary of methods

In this chapter, I reviewed methods for estimating the heterogeneity variance in a random-effects meta-analysis. A complete list of methods is given in table 2.1.

The estimators share many methodological aspects in common and can be grouped into the following categories. Method of moments estimators are derived from a generalised Q-statistic with weights that depend on the specific estimator being used. Other moments based estimators include HM, SJ and SJ_{CA}, which were developed so that the heterogeneity variance is positive in all meta-analyses. Maximum likelihood estimators are derived from maximising the log-likelihood function of a random-effects model assuming normality; these estimators include ML, REML and its approximate version (ARML). Bayesian estimators are derived from Bayes rule of conditional probabilities and take into account prior beliefs about the model parameters. Alternative methods for estimating the within-study variances in section 2.8 are not strictly heterogeneity variance estimators, but have been proposed to improve its estimates indirectly.

Estimator	Acronym	Section
Method of moments approach (section 2.2)		
DerSimonian-Laird	DL	2.2.1
Positive DerSimonian-Laird	DL_P	2.2.1
Cochran's ANOVA	CA	2.2.2
Paule-Mandel	PM	2.2.4
Two-step Cochran's ANOVA	PM_{CA}	2.2.5
Two-step DerSimonian-Laird	PM_{DL}	2.2.5
Other non-truncated moments-based approaches (section 2.3 and 2.4)		
Hartung-Makambi	HM	2.3.1
Sidik-Jonkman	SJ	2.4
Sidik-Jonkman (CA initial estimate)	SJ_{CA}	2.3.2
Hunter-Schmidt	HS	2.4
Maximum likelihood approach (section 2.5)		
Maximum Likelihood	ML	2.5.1
Restricted Maximum Likelihood	REML	2.5.2
Approximate Restricted Maximum Likelihood	ARML	2.5.3
Bayesian approach (section 2.6)		
Full Bayes	FB	2.6.1
Approximate Bayes	AB	2.6.2
Bayes modal	BM	2.6.3
Rukhin (zero prior)	B0	2.6.4
Rukhin (simple)	BP	2.6.4
Rukhin (alternate)	SB	2.6.4
Bootstrap approach		
Bootstrap DerSimonian-Laird	DL_B	2.2.1
SMD outcome only		
Malzahn, Böhning and Holling	MBH	2.7

Table 2.1: Summary of heterogeneity variance estimators

Some heterogeneity variance estimates are easier to compute than others. Many can be expressed explicitly, such as DL, and are therefore simple to compute. PM and estimators derived from the maximum likelihood approach require a process of iteration to converge to a solution; this is because they define random-effects study weights that include the heterogeneity variance parameter. Alternatives to PM have been proposed that that restrict the process of iteration to two-steps; those proposed under the method of moments approach in section 2.2.5 and Sidik and Jonkman estimators in section 2.3.2. The full Bayesian method requires MCMC

simulation methods to converge to a solution. Researchers have often favoured the more simple methods [64, 95, 101]. However, I believe reliable iterative methods should not be considered inferior to simple methods on this basis alone, particularly when software packages exist that automate the process and always converge to the optimal estimate for the given method.

Heterogeneity variance estimators introduced in this chapter require statistical assumptions, of which many are unlikely to hold. A frequent assumption is that within-study variances are known, when in practice they can only be estimated from the study data. When estimated within-study variances account for more than the observed total variance (i.e. when they are overestimated), methods such as DL produce a heterogeneity variance estimate truncated to zero as it would otherwise be negative. Only the full Bayesian method in section 2.6.1 does not require this assumption.

The maximum likelihood estimators assume that true study effects are normally distributed. The validity of this assumption has been questioned in medical meta-analyses [11, 28, 48]. The assumption is particularly questionable in the presence of publication or reporting bias [92] or in binary outcome meta-analyses with small study sample sizes [116]. A typical check for normality, such as that proposed by Egger et al. [27], often lacks sufficient power [11]. We also typically make the assumption of normally distributed effects in the Bayesian approach, but other more flexible distributions can be assumed.

2.10 Example meta-analysis

I present a meta-analysis of studies comparing hawthorn extract with placebo for treatment of chronic heart failure [82] to put the methods outlined in this chapter into context. The primary outcome is 'maximum work load' measured in METS (metabolic equivalents), which captures the amount of oxygen consumed and groups are compared by a mean difference (the outcome measure introduced in section 1.3.1).

The results of each study are presented graphically in the form of a forest plot in figure 2.1. All studies had positive effect estimates in favour of hawthorn extract but only Zapfe [129] was statistically significant at the 5% level.

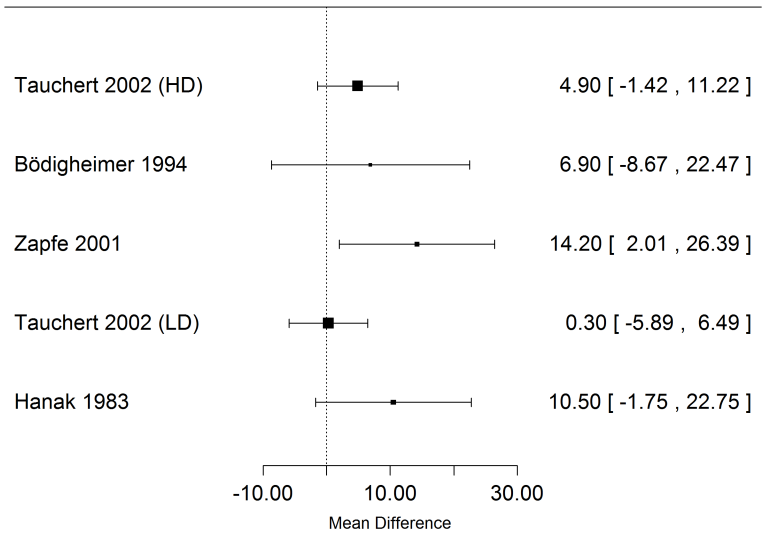


Figure 2.1: Forest plot with five studies in a meta-analysis evaluating hawthorn extract for chronic heart failure

Table 2.2 shows estimates of the heterogeneity variance according to most methods introduced in this chapter along with the associated estimate of I^2 (calculated using formula 1.6 in the last chapter). Estimates of the heterogeneity variance ranged from 0 (using Cochran's ANOVA method) up to 24.56 (using Rukhin's simple estimator, BP). Corresponding I^2 estimates range from 0% to 53.1%. All other estimators derived from the method of moments approach, except Cochran's ANOVA, have relatively similar τ^2 estimates (DL, DL_P, PM, PM_CA, PM_DL). Sidik and Jonkman estimators (SJ and SJ_CA) has wildly different estimates despite being methodologically similar; this is perhaps because SJ_CA uses the small initial estimate derived from Cochran's ANOVA, which results in a final estimate of CA's truncated value. HM is also a non-truncated estimator and has a relatively large heterogeneity variance estimate ($\tau^2 = 11.14$). There is a large difference between ML and REML heterogeneity variance estimates, despite their methodological similarity.

I use the same meta-analysis as an example in the next chapter, to show how dif-

ferences in heterogeneity estimates impact on the summary effect and its confidence interval.

Estimator	$\hat{\tau}^2$	$\hat{I}^2$	Estimator	$\hat{\tau}^2$	$\hat{I}^2$
DL	6.56	23.2	ML	1.27	5.5
DL_P	6.56	23.2	REML	9.31	30.0
CA	0	0.0	ARML	7.16	24.8
PM	5.88	21.3	AB **	6.56	23.2
PM_{CA}	6.56	23.2	BM	5.48	20.2
PM_{DL}	5.78	21.1	B0	2.47	10.2
HM	11.14	33.9	BP	24.56	53.1
SJ	13.93	39.1	DL_B	4.39	16.8
SJ_{CA}	0.01	0.1	MBH *	-	-
HS	0.80	3.6			

Table 2.2: Heterogeneity variance estimates derived from different methods and associated I^2 estimates
**Malzahn, Böhning and Holling's estimator (MBH) could not be calculated as study effects are not on the standardised scale*
***Approximate Bayes estimate based on $\tau^2 = 0$ and $\eta = 1$ priors.*

2.11 Conclusions

In this chapter, I reviewed methods for estimating the heterogeneity variance in a random-effects meta-analysis. I described how they are derived, their formulae, methodological similarities and weaknesses. I identified 20 distinct methods for heterogeneity variance estimation. Some of these methods have only recently been proposed, such as those by Rukhin [93]; these are unlikely to have been compared extensively with pre-existing methods in simulated or empirical data.

I have previously mentioned the weaknesses of the commonly used DerSimonian-Laird method that have been brought to light in the literature [78, 79, 124]. The numerous alternatives as outlined in this chapter suggests there may exist one or more methods with better properties. It is imperative that the properties of these

estimators are better understood so that informed recommendations can be made for frequentist random-effects meta-analysis.

Chapter 3

Methods for confidence intervals of the summary effect

3.1 Introduction

The impact of using a heterogeneity variance estimator in random-effects meta-analysis extends beyond its point estimate. In order to perform a comprehensive comparison of heterogeneity variance estimators in this thesis, I also compare these estimators in terms of their impact on the summary effect and its confidence interval. I focus only on the inverse variance method for calculating estimates of the summary effect, as described in section 1.6 of the introduction chapter.

Selected confidence interval methods are introduced in this chapter that will be used throughout the rest of this thesis; a comprehensive review is not required because these methods are not the main focus. Confidence interval methods introduced in this chapter include Wald-type (section 3.2), t-distribution (section 3.3) and Hartung-Knapp confidence intervals (section 3.4). The Wald-type confidence interval is currently reported as standard in meta-analyses in Cochrane reviews [51]. The other two methods are included because they have shown promising results in simulation studies [61, 93, 96] and are methodologically diverse. None of the chosen methods are theoretically related to a specific heterogeneity variance estimator, meaning any heterogeneity variance estimate can be used to derive a confidence interval. This is a key characteristic given the aims of this thesis. Other methods can be found in the literature [11, 12, 37, 45, 83, 96]. For example, the profile likelihood method [37] is based on the same log-likelihood function as the maximum likelihood heterogeneity variance estimator.

All confidence interval methods included in this chapter can be adapted for both random and fixed-effect meta-analysis. Given the focus of this thesis, I present these methods assuming random-effects. The methods can also be adapted for any confidence level; I present their generalised form in this chapter but I focus on the 95% confidence level in the rest of this thesis.

Many confidence interval methods are also available to express uncertainty around the heterogeneity variance estimate [60, 125]. These methods will not be considered

as the scope of this thesis is limited to the impact on point estimates of the heterogeneity variance, subsequent summary estimates and their confidence intervals.

3.2 Wald-type confidence interval

The Wald-type confidence interval is most commonly used in meta-analysis, and is previously described in the introduction chapter (section 1.6). Recall [25, 100]:

$$\left[\hat{\theta} - \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\cdot\sqrt{Var_W\left(\hat{\theta}\right)},\ \hat{\theta} + \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\cdot\sqrt{Var_W\left(\hat{\theta}\right)}\right]$$

$\Phi^{-1}\left(1-\frac{\alpha}{2}\right)$ is the $\left(1-\frac{\alpha}{2}\right)$th percentile of the normal distribution and, for a 95% confidence interval, set $\alpha = 0.05$. $Var_W\left(\hat{\theta}\right)$ is the variance of the summary effect and is calculated by the formula:

$$Var_W\left(\hat{\theta}\right) = \frac{1}{\sum_{i=1}^{k}\left(\hat{w}_i\right)} \tag{3.1}$$

where $\hat{w}_i = 1/\left(\hat{\sigma}_i^2 + \hat{\tau}^2\right)$ and k is the number of studies in the meta-analysis.

This confidence interval method assumes that study effect estimates $(\hat{\theta}_i)$ follow a normal distribution [12]. Also, the method assumes τ^2 and $\hat{\sigma}_i^2$ are known, but estimates of these parameters are used in practice [11, 96].

3.3 t-distribution confidence interval

The t-distribution confidence interval for the summary effect addresses the small sampling bias of the Wald-type confidence interval, and is therefore thought to improve coverage [28]:

$$\left[\hat{\theta} - t_{k-1}\left(1-\frac{\alpha}{2}\right)\cdot\sqrt{Var_W\left(\hat{\theta}\right)},\ \hat{\theta} + t_{k-1}\left(1-\frac{\alpha}{2}\right)\cdot\sqrt{Var_W\left(\hat{\theta}\right)}\right] \tag{3.2}$$

$t_{k-1}\left(1-\frac{\alpha}{2}\right)$ is the $\left(1-\frac{\alpha}{2}\right)$th percentile of the t-distribution with $k-1$ degrees of freedom. $Var_W\left(\hat{\theta}\right)$ is the same variance estimate as for the Wald-type confidence interval and makes all the same assumptions.

3.4 Hartung-Knapp confidence interval

The Hartung-Knapp method [38–40] also relies on a t-distribution with $k-1$ degrees of freedom and uses an alternative weighted variance of θ. I stated the formula for this weighted variance in the last chapter, as it was used to derived SJ estimators of the heterogeneity variance (section 2.3.2). Its formula can be derived as follows.

If we make the assumption that random-effects weights $\hat{w}_i = 1/\left(\hat{\sigma}_i^2 + \hat{\tau}^2\right)$ are known, then:

$$\frac{\hat{\theta} - \theta}{\sqrt{1/\sum_{i=1}^{k}\hat{w}_i}} \sim N(0,1)$$

and

$$\sum_{i=1}^{k}\hat{w}_i\left(\hat{\theta}_i - \hat{\theta}\right)^2 \sim \chi^2_{k-1}$$

Since these two variables are independent (as proven by Hartung [38] and Sidik and Jonkman [100]), and by definition of the t-distribution, we can derive:

$$\frac{N(0,1)}{\sqrt{\chi^2_{k-1}/(k-1)}} = \frac{\left(\hat{\theta} - \theta\right)\Big/\sqrt{1/\sum_{i=1}^{k}\hat{w}_i}}{\sqrt{\sum_{i=1}^{k}\hat{w}_i\left(\hat{\theta}_i - \hat{\theta}\right)^2\Big/(k-1)}} \sim t_{k-1} \tag{3.3}$$

Given that $\left(\hat{\theta} - \theta\right) \Big/ \sqrt{Var\left(\hat{\theta}\right)} \sim t_{k-1}$, we can equate this to equation 3.3 above and derive the following formula for the weighted variance:

$$Var_{HK}\left(\hat{\theta}\right) = \frac{\sum_{i=1}^{k} \hat{w}_i \left(\hat{\theta}_i - \hat{\theta}\right)^2}{(k-1)\sum_{i=1}^{k} \hat{w}_i}$$

and thus the Hartung-Knapp confidence interval is:

$$\left[\hat{\theta} - t_{k-1}\left(1 - \frac{\alpha}{2}\right) \cdot \sqrt{Var_{HK}\left(\hat{\theta}\right)},\ \hat{\theta} + t_{k-1}\left(1 - \frac{\alpha}{2}\right) \cdot \sqrt{Var_{HK}\left(\hat{\theta}\right)}\right] \tag{3.4}$$

This method is the equivalent to the t-distribution method in the last section, but the variance is multiplied by a *scaling factor* as explained by Sidik and Jonkman [100]. In many cases, the scaling factor can be < 1, which leads to a narrower confidence interval that the Wald-type method. A variation on this method have been proposed to deal with this problem by constraining the scaling factor to be ≥ 1 [61]. Throughout this thesis, the original Hartung-Knapp method without constraint will be used.

As with the other two confidence intervals in this chapter, the variance components are assumed to be known and study effects normally distributed.

3.5 Example meta-analysis

I now revisit the example meta-analysis from section 2.10 comparing hawthorn extract with placebo for increasing maximum workload in patients with chronic heart failure. I present its summary effect and 95% confidence interval according to all combinations of heterogeneity variance and summary effect confidence interval estimation

methods (16 heterogeneity variance methods and 3 confidence interval methods). All results of this meta-analysis are plotted in figure 3.1.

The heterogeneity variance estimate was zero according to Cochran's ANOVA method and more than zero for all other methods (see section 2.10). A narrower confidence interval is observed when Cochran ANOVA's zero estimate is used as shown in figure 3.1. Similarly, Rukhin's simple estimator (BP) estimated the highest heterogeneity variance ($\hat{\tau}^2 = 24.56$) and lead to the widest confidence intervals. Summary effects are fairly consistent between methods, but confidence interval widths differ more significantly when Wald-type and t-distribution methods are used. Hartung-Knapp confidence intervals appear more robust to changes in the heterogeneity variance estimate. Only Wald-type confidence intervals produced a statistically significant result at the 5% level.

3.6 Concluding remarks

In this chapter, I described three confidence intervals proposed for the summary effect in random-effects meta-analysis; namely Wald-type [25], t-distribution [28] and Hartung-Knapp confidence intervals [38].

The Wald-type confidence interval for the summary effect is most commonly used in meta-analysis. This method, coupled with the DerSimonian-Laird estimator of the heterogeneity variance, is often referred to as the DerSimonian-Laird approach to random-effects meta-analysis [25]. However, I associate *DerSimonian-Laird* solely with the heterogeneity variance estimator in this thesis. Wald-type confidence intervals, and the other two methods introduced in this chapter, can be calculated with any heterogeneity variance estimator.

In this chapter, I also presented these methods in the context of a real meta-analysis to show the choice of methods can lead to different confidence interval estimates. I

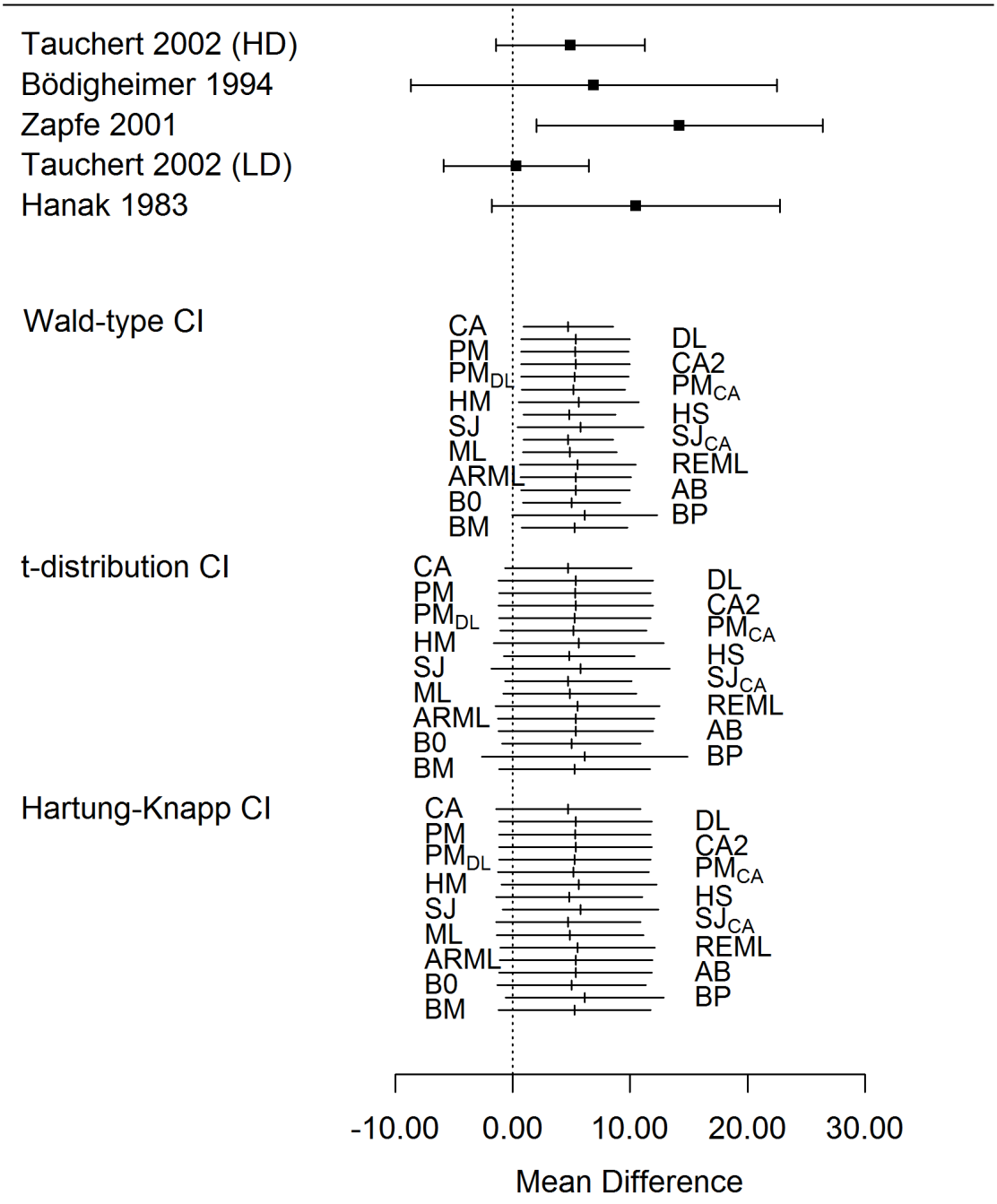


Figure 3.1: The summary effect and 95% confidence interval for all combinations of heterogeneity variance and confidence interval methods

compare heterogeneity variance estimators in more empirical meta-analysis data in the next chapter.

Chapter 4

An empirical comparison of heterogeneity variance estimators

4.1 Introduction

In this chapter, I assess the impact of using different heterogeneity variance estimators on published meta-analyses, using empirical data derived from a complete snapshot of the *Cochrane Database of Systematic Reviews* (CDSR) up to 2008 [107]. Because the Cochrane Collaboration's Review Manager software [87] is used to write all Cochrane reviews, the data are formatted in a highly consistent way. Results from each study of each meta-analysis could therefore be extracted, including type of outcome, sample size, 2×2 tables for dichotomous outcomes, and means with standard deviations for continuous outcomes. Extraction of the data is described in more detail elsewhere [21]. Permission to use the dataset for this analysis was granted by Rebecca Turner, who is acknowledged in the front matter of this thesis.

These data allowed me to look at (1) the magnitude of differences in heterogeneity variance estimates in practice, (2) the impact of the choice of heterogeneity variance estimator on conclusions and (3) the extent to which recommendations of the best heterogeneity variance estimator are required. I also examine two selected examples from the dataset, where differences between estimation methods are particularly prominent.

This chapter expands on a previous empirical study to compare statistical inference between heterogeneity variance estimators [117]. Limitations of the previously published study were as follows. First, estimates were compared from only five methods (DL, CA, HM, SJ and REML) so I compare many more (as detailed in section 4.2.1). Second, estimates were transformed to the scale of the D^2 statistic [127] and compared on this scale as a measure of agreement. The D^2 statistic is a measure of the degree of heterogeneity and, much like the I^2 statistic, takes values between 0% and 100%. The issue is the D^2 statistic is rarely used in practice, so I compare estimates on the scale of the I^2 statistic (more details and reasoning are given in section 4.2.4). Furthermore, the previous study is based on 920 meta-analyses; I conduct my analysis on 12,894 empirical meta-analyses to gain more precise estimates of

agreement.

4.2 Methods

4.2.1 Included methods for estimating the heterogeneity variance

I present a comparison of seven methods for estimating the heterogeneity variance in a meta-analysis: DerSimonian-Laird (DL) [25], Cochran's ANOVA (CA) [59], Paule-Mandel (PM) [80], Hartung-Makambi (HM) [40], Sidik-Jonkman (SJ) [101], maximum likelihood (ML) [37] and restricted maximum-likelihood (REML) [41]. These seven estimators were selected from the comprehensive list in chapter 2 because of their popularity and availability in statistical software. DL is derived from the method of moments approach to heterogeneity variance estimation and is the most frequently used heterogeneity variance estimator in practice. It is the default method in the Stata command *metan* and is currently the only method implemented in RevMan software [22]. CA assigns equal weightings to studies and represents a simple alternative to DL. PM assigns random-effects weights to studies, which are considered the statistically optimal weights in the method of moments approach [24]. REML is the default method in the R package *metafor* [126]. ML is a widely-used approach to statistical parameter estimation and is therefore also included in this analysis. In contrast to these estimators, HM and SJ were selected as non-truncated estimators that always estimate a positive heterogeneity variance.

I stated in chapter 2 that the PM estimator can theoretically be interpreted as a simple approximation of the REML approach in specific situations [94]. The extent of agreement between PM and REML estimates has not been investigated in other empirical studies [117] so both estimators are included and compared here.

Results in this chapter are representative of a comparison of all heterogeneity variance

estimators. To demonstrate this, many of the estimators not included in the main text are given in figure B.1 in the appendix. This figure shows a comparison between two included methods (DL and REML) and many that are excluded (PM_{CA}, PM_{DL}, HS, SJ_{CA} and ARML). I also exclude Bayesian methods that rely on a subjective choice of prior distribution because of difficulties defining these distributions out of context. Rukhin's estimators [93] are excluded because simulations results later in this thesis show they have poor properties. The estimator proposed by Malzahn, Böhning and Holling [74] is excluded because it can only be used in meta-analyses with a standardised mean difference outcome measure. I exclude bootstrapping because the approach could theoretically be applied to any heterogeneity variance estimator.

4.2.2 Empirical study dataset

A complete re-analysis of all meta-analyses in the CDSR dataset was possible from the study-level data available. I re-conducted all meta-analyses of dichotomous or continuous outcomes containing at least three studies. Those containing two studies were excluded from the results because it is arguably inappropriate to estimate heterogeneity in such cases. Effect estimates and standard errors were calculated for all studies from basic summary statistics. I calculated the log odds ratios for all dichotomous outcome meta-analyses and standardised mean differences for all continuous outcome meta-analyses. Hedges' g method was used to estimate standardised mean difference effects, which corrects for bias caused by small sample sizes [8] and is detailed in section 1.3.1.

4.2.3 Summary statistics

I used four summary statistics to compare the seven estimation methods: (i) the estimated heterogeneity variance, (ii) the estimated summary effect from a random-effects meta-analysis, (iii) the estimated standard error of the summary effect, and

(iv) the p-value for this result. These statistics were chosen because they are the key statistics used to draw inference from a meta-analysis and may be affected by the estimated heterogeneity variance. Furthermore, by comparing standard errors, I can also compare the widths of confidence intervals of the summary effect because confidence interval formulae for all included methods are otherwise independent of $\hat{\tau}^2$.

I calculated standard errors and hence p-values for the overall summary effect (i.e. summary statistics (iii) and (iv) above) using both Wald and Hartung-Knapp methods (i.e. two of the three methods outlined in chapter 3). The Wald method is the currently used as standard in Cochrane meta-analyses [51] and was introduced in section 1.6. The Hartung-Knapp method uses an alternative weighted standard error of the summary effect and derives a p-value from the t-distribution. This method was introduced in section 3.4 and derived from the same approach as the Hartung-Knapp confidence interval for the summary effect. I omitted p-values based on the t-distribution method outlined in section 3.3 because they are based on the same formula for the variance as the Wald-type method and therefore results would be be similar.

4.2.4 Data analysis

I illustrate pair-wise agreement between results from different estimation methods using Bland-Altman plots [5], thereby illustrating how the discrepancy between two methods depends on the underlying value of the parameter (estimated as the average result across the two methods). Pair-wise plots are arranged in a matrix to facilitate simultaneous comparison of each method with all others. Bland-Altman plots are used to examine the first three of our four summary statistics (heterogeneity variance, summary effect and precision of summary effect). I superimpose non-parametric 80% reference ranges on the same plots to illustrate the spread of agreement. To calculate the 80% reference ranges, I split meta-analyses into groups of 200 according to their

order on the x-axis and calculated the 10th and 90th percentiles of the discrepancies. The plotted reference range is a smoothed line between the calculated percentiles.

Bland-Altman plots traditionally present raw differences between parameter estimates on the y-axis, but precisions of the summary effect are compared as a ratio for two reasons: (1) They naturally conform to a log-normal distribution. (2) By including precision of the summary effect in this analysis, I can also compare the widths of summary effect confidence intervals and these comparisons are more meaningful on the ratio scale. For example, a confidence interval that is half the width of another is half as likely to include the null value with all else being equal. Heterogeneity variance estimates also have a skewed distribution in practice [21], but I apply a transformation as detailed below and present raw differences.

I sought to measure discrepancies between heterogeneity variance (τ^2) estimates on an appropriate scale that would maximise the generalisability of the results and be intuitively interpretable. The most obvious option is to present the raw differences of τ^2 estimates, but the scale of these differences is too dependent on the average τ^2 estimate (as shown in appendix A.1). Therefore, I transformed τ^2 estimates to the scale of the I^2 statistic and present their raw differences (see equation 1.6 for I^2 in the introduction chapter). I consider this a transformation because all parameter estimates other than the heterogeneity variance estimate τ^2 remain fixed between methods. Differences in I^2 statistics reflect only differences in values of τ^2.

The summary effect and its standard error depend on the scale of measurement. Therefore, I multiplied standardised mean differences and standard errors from each continuous meta-analysis by a value of 1.81 to obtain a result that is approximately comparable to a log odds ratio [15]. The I^2 statistic and p-values for the summary effect are independent of the scale of measurement and so do not require a transformation. I carried out separate analyses on continuous and binary outcome meta-analyses, but since I found no difference between the results I present results with all meta-analyses combined.

I compared p-values of the summary effect by tabulating categories of levels of statistical significance. First, p-values were dichotomised at the 5% level to explore agreement for the threshold most commonly applied in practice. Second, p-values were categorised to represent a wider range of levels of statistical significance: $p \leq 0.01$, $0.01 < p \leq 0.05$, $0.05 < p \leq 0.1$ and $p > 0.1$. I considered p-values that differ by at least 2 categories on this finer scale to be sufficiently different to change inference. I recognise the limitations of using statistical significance to draw inferences [110, 117], but also appreciate their widespread use.

In a secondary analysis, I explored whether the level of agreement between heterogeneity variance estimates can be explained by two meta-analysis characteristics; the number of studies (k) and the total information (V). Hardy and Thompson [37] defines the total information as $V = \sum_{i=1}^{k} \hat{w}_i$, which takes into account the number and sizes of studies. Hardy and Thompson [37] found using simulations that the power to detect heterogeneity (using the Q-statistic from section 1.7.1) depends on these characteristics, so I explored whether they also affect the level of agreement between heterogeneity variance estimators. I illustrate their effects using the same plots of pair-wise agreement as for the main analysis, but with the the number of studies and total information on the x-axes.

4.3 Results

A summary of the characteristics of meta-analyses in the CDSR is presented in section 4.3.1. Heterogeneity variance estimators are compared in sections 4.3.2 to 4.3.5 for each of the four summary statistics. In section 4.3.6, I show whether the level of agreement between heterogeneity variance estimates is affected by two meta-analysis characteristics.

4.3.1 A summary of CDSR meta-analyses

The 2008 version of the CDSR contains 22,453 meta-analyses, of which I excluded 9559 (42.6%). 8641 meta-analyses were excluded because they contain fewer than three studies and the remaining 918 meta-analyses were excluded because the type of outcome measure is missing or something other than binary or continuous. A total of 12,894 meta-analyses from 1817 systematic reviews are included in these analyses.

Type of outcome	Outcome measure	N (%)
binary	odds ratio	3295 (26%)
	relative risk	5568 (43%)
	risk difference	116 (1%)
continuous	standardised mean difference	948 (7%)
	mean difference	2967 (23%)
	Total	12,894

Table 4.1: The original outcome measures of included meta-analyses from the CDSR

8979 (70%) included meta-analyses have a dichotomous outcome and 3915 (30%) have a continuous outcome (as shown in table 4.1). I calculated odds ratio outcome measures for all binary outcome meta-analyses in this analysis, of which 3295 (37%) use this measure in the original publication. For continuous meta-analyses, I calculated a standardised mean difference outcome measure, of which 948 (24%) originally used this measure.

Figure 4.1 shows the numbers of studies contained in included meta-analyses from the CDSR dataset. Having excluded those that contain fewer than three studies, the median number of studies is 4 (inter-quartile range 3-7) and 11,009 (85.4%) meta-analyses contain fewer than 10 studies.

Figure 4.2 shows the distribution of DL estimates of the heterogeneity variance and I^2 statistics for both dichotomous outcomes (based on odds ratios) and continuous outcomes (based on standardised mean differences). The DL method estimated $\hat{\tau}^2 = 0$ for 4395 (49%) dichotomous outcome meta-analyses and 1315 (33.6%) continuous-outcome meta-analyses.

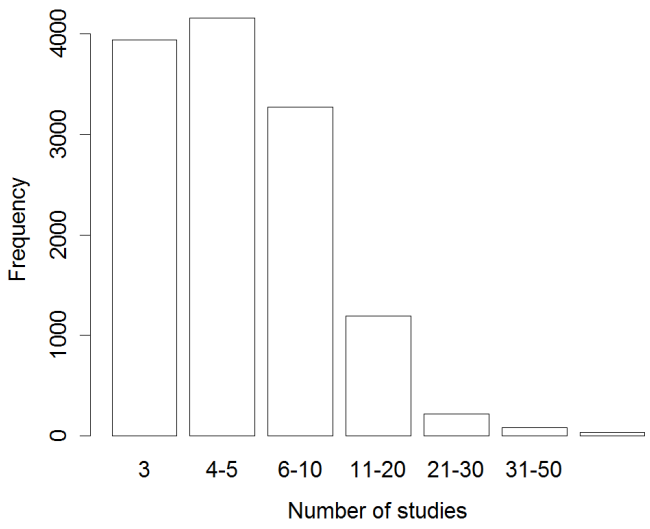


Figure 4.1: The numbers of studies included in all 2,894 meta-analyses

4.3.2 Agreement between heterogeneity variance estimates

I present in figure 4.3 Bland-Altman plots of pair-wise agreement between heterogeneity variance estimates, expressed on the I^2 scale, for the seven estimation methods: Cochran's ANOVA (CA), DerSimonian-Laird (DL), Paule-Mandel (PM), Hartung-Makambi (HM), Sidik-Jonkman (SJ), ML and REML. The plots show the difference between two I^2 statistics for a particular pair of methods on the y-axes as a measure of agreement. 80% reference ranges are shown by the thick red lines. Because I^2 values depend both on between-study variance and within-study variance, the horizontal positioning of the meta-analyses on this scale is affected by both of these: meta-analyses to the left have either low heterogeneity or high within-study variance (or both), and those to the right have high heterogeneity or low within-study variance.

There is a relatively high level of agreement between DL and PM estimates of I^2, with perfect agreement when estimates of the heterogeneity variance are zero. In few cases do DL and PM estimates of I^2 differ by more than 25% in absolute value. There is also relatively high agreement between SJ and HM estimates when $I^2 < 25\%$ because neither method produces zero heterogeneity variance estimates. SJ estimates of I^2

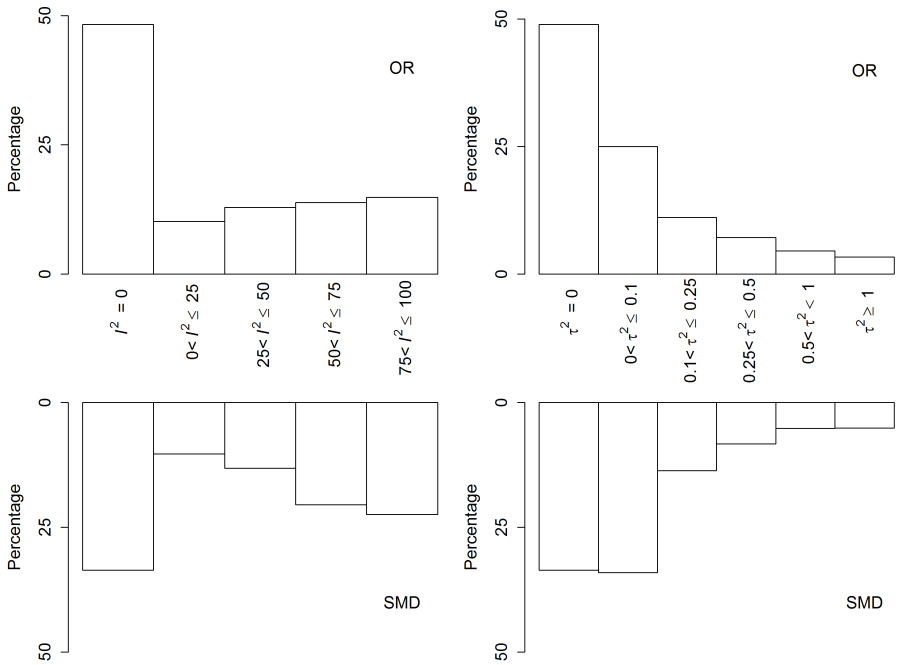


Figure 4.2: The distribution of τ^2 and I^2 estimates for OR and standardised mean difference meta-analyses calculated from the DL method.
Frequencies expressed as a percentage.

are generally larger than other estimates. REML has low agreement with both DL and PM estimators when $I^2 < 75\%$. ML generally produces lower I^2 estimates than other methods in all comparisons. Apart from the reasonable agreement observed between DL and PM, all other comparisons show a low level of agreement where in many cases one method estimates $I^2 = 0\%$ and the other estimates $I^2 > 50\%$. In particular, CA I^2 estimates have low agreement with all other I^2 estimates. Points which make up straight diagonal lines seen in most plots show where one estimate is $I^2 = 0\%$ and the other is positive; extreme differences in I^2 estimates occur more frequently in these cases. There appears to be less agreement between estimates when the average I^2 (x-axis) is around 50% and a high level of agreement close to 0% or 100%; this however is largely because absolute differences in I^2 estimates have a limited range when the average is close to the upper and lower limits of I^2.

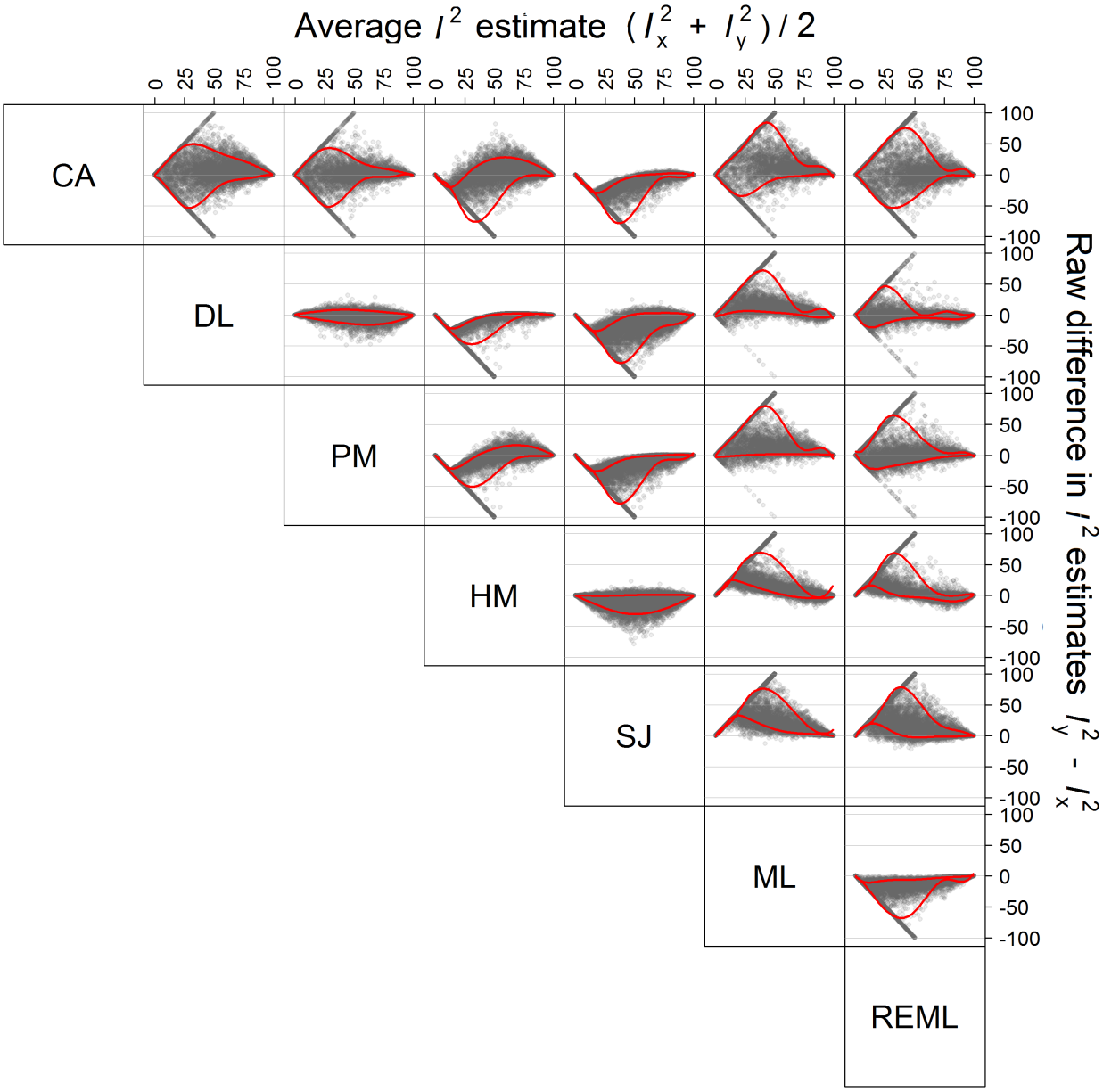


Figure 4.3: Bland-Altman scatter plots comparing I^2 estimates from different heterogeneity variance methods.

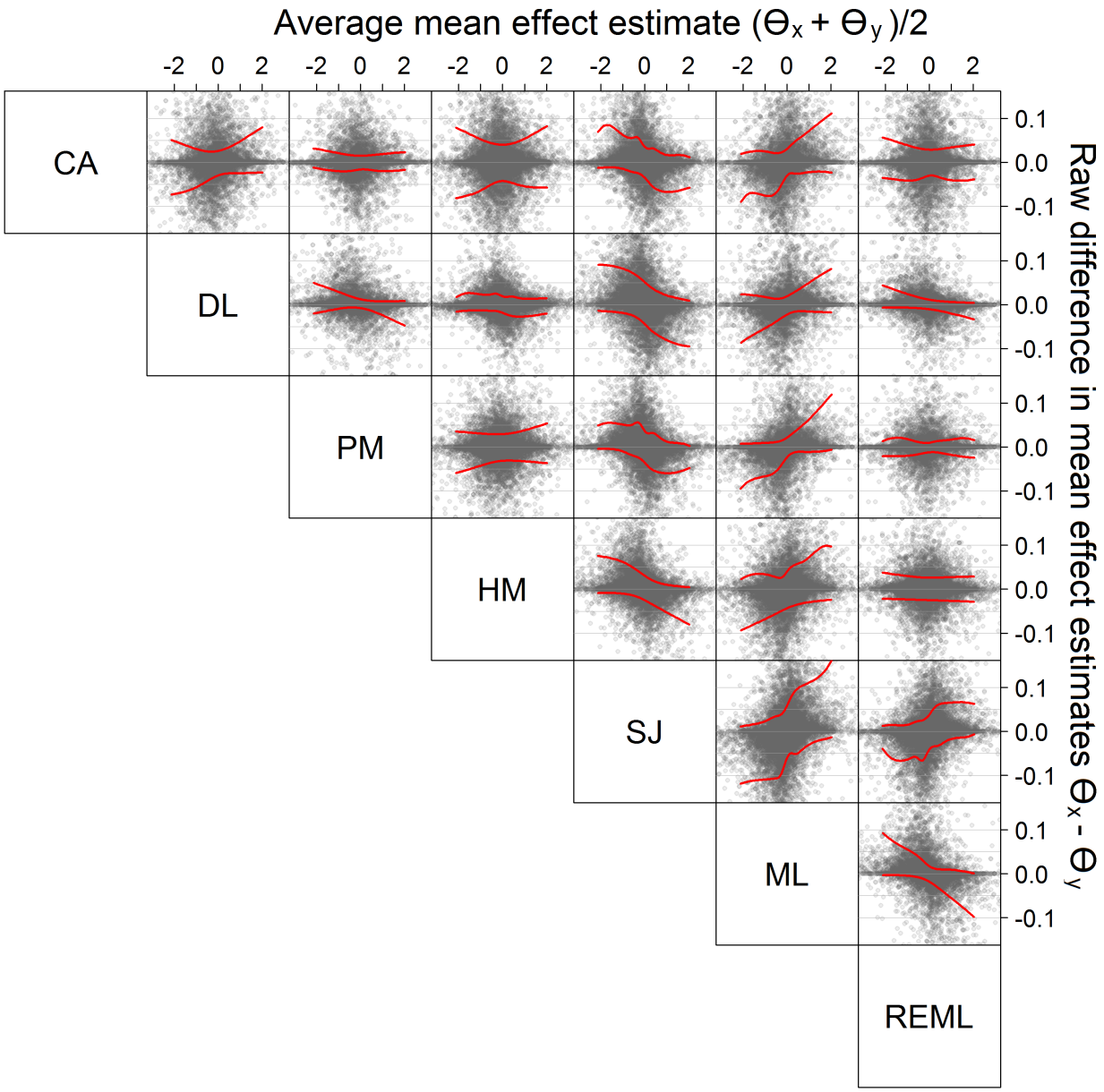


Figure 4.4: Bland-Altman scatter plots comparing summary effect estimates using different heterogeneity variance estimation methods.
θ estimates represent log odds ratios, including standardised mean differences in continuous outcome meta-analyses converted to the same scale

4.3.3 Agreement between summary effects

Figure 4.4 shows Bland-Altman plots of the pair-wise agreement between summary effects. These are expressed as log odds ratios, calculated either from a dichotomous outcome meta-analysis or from a continuous outcome meta-analysis after being transformed to the log odds ratio scale as detailed in the methods section. Summary effects agree most between DL, PM and REML heterogeneity variance estimators. However, the level of agreement is high for all pair-wise comparisons, in most cases differing by a negligible amount. Some 80% reference ranges appear to show poor agreement far from the null value; this is most likely because extreme summary effects are few in number and as such have considerable impact on the reference range. The agreement of summary effects between REML and all other heterogeneity variance estimators appear to depend on whether the summary effect is positive or negative. None of the methods depend on the direction of effect, so such effects are due to chance alone or due to differences in characteristics of meta-analyses with positive and negative effects.

4.3.4 Agreement between precision of the summary effect

Figure 4.5 shows Bland-Altman plots of the level of agreement between standard errors of the summary effect. The upper-right panel displays agreement of Wald standard errors and the lower-left displays agreement of Hartung-Knapp weighted standard errors. Agreement is measured as a ratio of standard errors (equivalently a ratio of confidence interval widths), and plotted on the log-scale.

Results suggest that changing the heterogeneity variance estimator can possibly halve or double the size of Wald standard errors. These standard errors agree most for the pair-wise comparisons DL v PM and DL v HM. The SJ estimator in most meta-analyses produced higher standard error than other heterogeneity variance estimators; this is expected given that the SJ estimate of I^2 was higher than other I^2

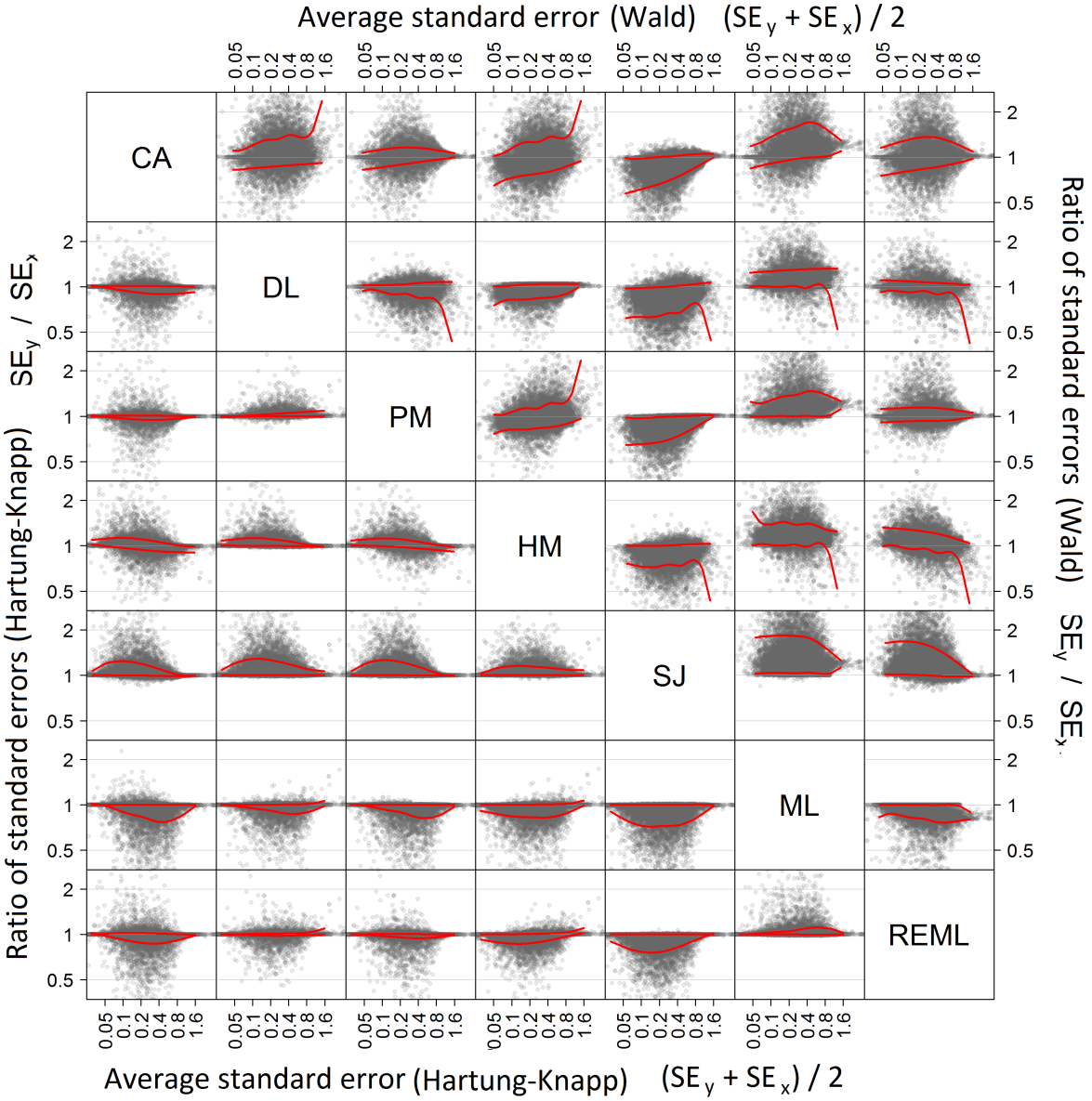


Figure 4.5: Bland-Altman scatter plots comparing summary effect estimates and standard errors using different heterogeneity variance estimation methods. *(1) Lower-left panel: Comparing Hartung-Knapp weighted standard errors (2) Upper-right panel: Comparing the Wald standard errors (differences presented on the log scale in all plots)*

estimates in almost all meta-analyses. The ML estimator in most meta-analyses produces a lower standard error than other heterogeneity variance estimators; this is expected given than results in section 4.3.2 show that ML tends to produce lower I^2 estimates than other estimators. All comparisons show poor agreement.

Hartung-Knapp standard errors have much higher agreement than Wald standard errors, a change in the heterogeneity variance estimator can lead to a 25% reduction or 50% increase in the standard error. Agreement of these standard errors is highest between DL, PM and REML estimators. Also, SJ and HM estimators have high agreement, even though agreement between their Wald standard errors is relatively low. ML and SJ estimators typically produce low and high Hartung-Knapp standard errors respectively (as they do for Wald standard errors).

4.3.5 Agreement between p-values

Tables 4.2 and 4.3 show pair-wise agreement between p-values of the summary effect when different heterogeneity variance estimators are used. The p-values in table 4.2 are derived from the Wald-statistic and p-values in table 4.3 are derived from the t-statistic; methods for deriving these p-values are given in section 4.2.2. The lower-left panels of both tables present agreement between p-values split into two categories: $p \leq 0.05$ and $p > 0.05$. The upper-right panels present agreement between p-values split into finer categories: $p \leq 0.01$, $0.01 < p \leq 0.05$, $0.05 < p \leq 0.1$ and $p > 0.1$.

Results in table 4.2 suggest that choice of heterogeneity variance estimation method can have an effect on inference when p-values are derived from the Wald-statistic. Statistical significance of these p-values at the 5% level is discordant between at least two heterogeneity variance estimators in 10.3% of meta-analyses. The lowest agreement in statistical significance at the 5% level is observed between SJ and ML with 8.6% of meta-analyses having discordant p-values. The highest agreement is observed between DL and REML and between PM and REML methods; 2% of meta-analyses were discordant.

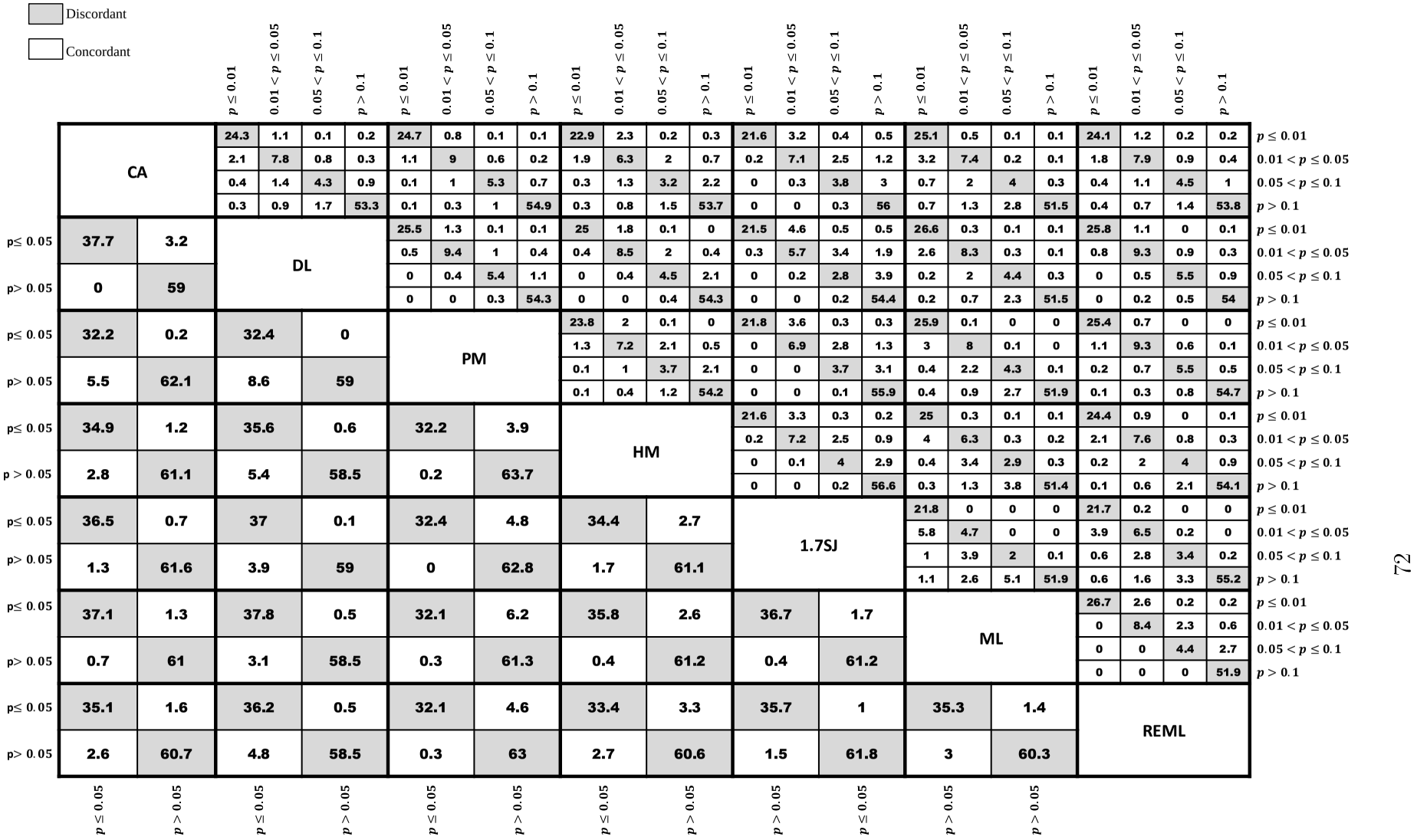


Table 4.2: The difference between heterogeneity variance estimation methods in terms of p-value categories derived from the Wald-statistic.

Upper-right panel: P-value categories split into 4: $p > 0.1$, $0.05 < p \leq 0.1$, $0.01 < p \leq 0.05$ *and* $p \leq 0.01$. *Lower-left panel: P-value categories split into 2:* $p > 0.05$ *and* $p \leq 0.05$.

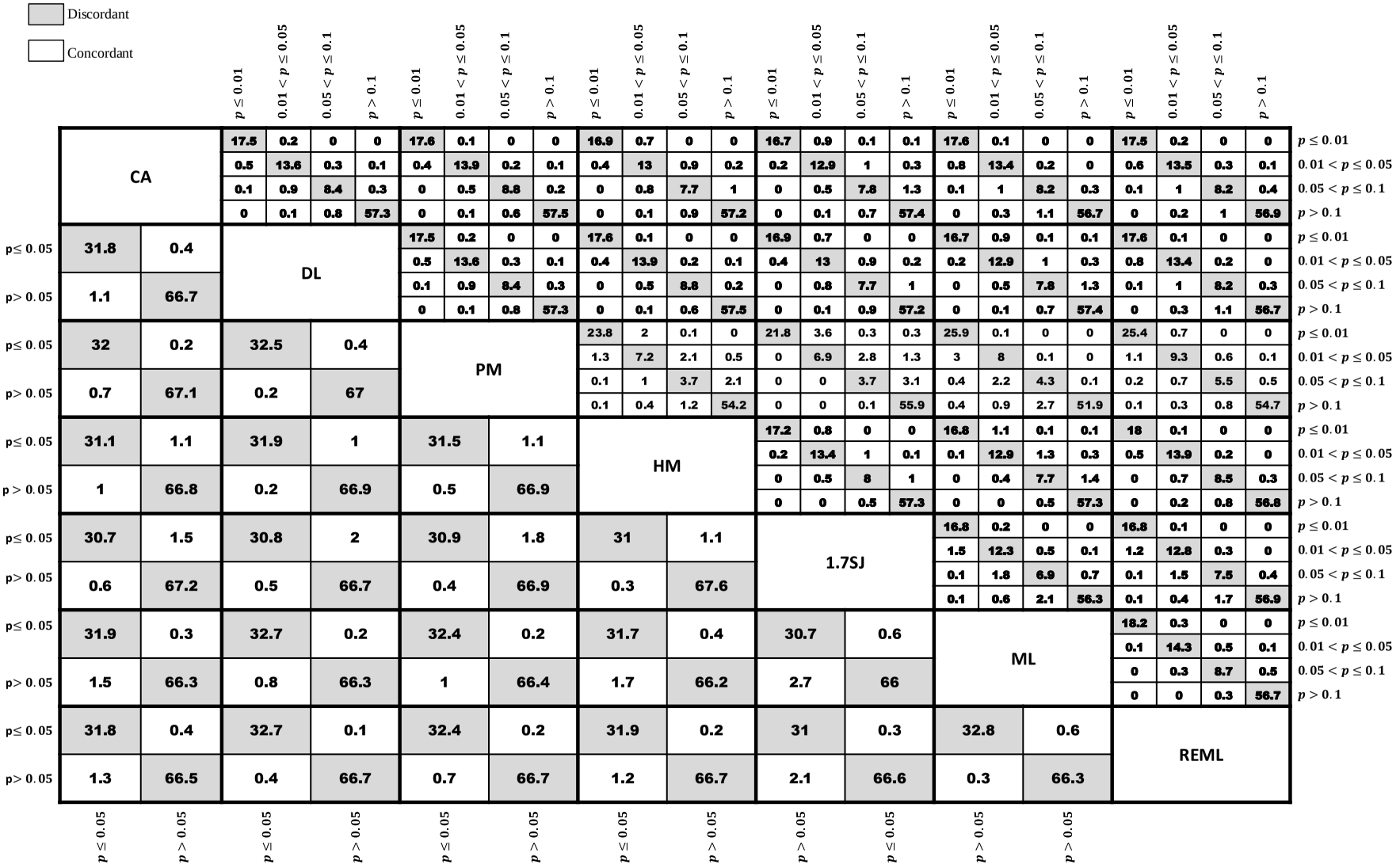


Table 4.3: The difference between heterogeneity variance estimation methods in terms of p-value categories derived from the Hartung-Knapp method.

Upper-right panel: P-value categories split into 4: $p > 0.1$, $0.05 < p \leq 0.1$, $0.01 < p \leq 0.05$ *and* $p \leq 0.01$. *Lower-left panel: P-value categories split into 2:* $p > 0.05$ *and* $p \leq 0.05$.

When p-values from the Wald-statistic are split into finer categories (also in table 4.2), there were differences of two or more categories between at least two heterogeneity variance estimators in 6.2% of meta-analyses. 1.4% of meta-analyses had p-values differing by 3 categories. P-values derived from SJ and ML heterogeneity variance estimators had the lowest agreement; 4.7% of meta-analyses differed by at least 2 p-value categories for this comparison and all where ML gave the lowest p-value. The highest agreement was observed between DL and HM methods; 0.6% of meta-analyses differed by at least 2 p-value categories and 0.05% of meta-analyses differed by 3 categories. There is also a high level of agreement between DL and PM methods; this is consistent with the level of agreement observed in terms of the other measures, including I^2, the summary effect and standard error.

Results from table 4.3 show p-values based on the t-statistic have much higher agreement between heterogeneity variance estimators. Statistical significance at the 5% level is discordant between at least two heterogeneity variance estimators in 3.7% of meta-analyses - roughly two-thirds less than the proportion of Wald-statistic p-values. Comparisons with low and high agreement are consistent between Wald-statistic and t-statistic p-values. The lowest agreement is observed between SJ and ML methods; 3.3% of meta-analyses have discordant statistical significance at the 5% level between these methods. The highest agreement is observed between DL and REML with discordance at the 5% level in 0.5% of meta-analyses.

When p-values derived from the t-statistic are split into finer categories (as shown in the upper-left panel of table 4.3), only 1.1% of meta-analyses had p-values that are discordant by two or more categories between any two estimators; almost six times fewer meta-analyses than p-values derived from the Wald-statistic. Again, the comparison between SJ and ML shows the lowest agreement, with p-values of 0.9% of meta-analyses discordant by two or more categories. In many other pair-wise comparisons, 0.1% or fewer meta-analyses are discordant by two or more categories.

4.3.6 Factors to explain the level of agreement between heterogeneity variance estimates

Figure 4.6 shows pairwise agreement between I^2 estimates plotted against the number of studies (in the upper-right panel) and total information (in the lower-left panel).

Results show that the general level of agreement between I^2 estimates is not correlated with how many studies there are in the meta-analysis. However, rare and extreme I^2 differences of close to 100% only occur in meta-analyses with fewer than 20 studies. The only exception is the comparison between ML and REML; differences between ML and REML estimates of I^2 are close to zero when there are 40 or more studies. For all other comparisons it appears the scatter of meta-analyses is showing a trend, but this is because most meta-analyses contain few studies and are situated to left hand side of the graph giving the appearance of less agreement between I^2 estimates.

When differences in I^2 are plotted against the total information, results show there is no trend in any of the comparisons. Extreme differences in I^2 close to 100% occur much less when the total information is high, but this can only be attributed to a high power to detect heterogeneity and therefore fewer zero I^2 estimates derived from the truncated estimators CA, DL, PM, ML and REML.

4.4 Examples

I selected two examples from the CDSR dataset, specifically chosen with widely different estimates of the heterogeneity variance to show how such differences may lead to different conclusions.

Example 1. Omega-3 fatty acids for prevention and treatment of cardiovascular disease

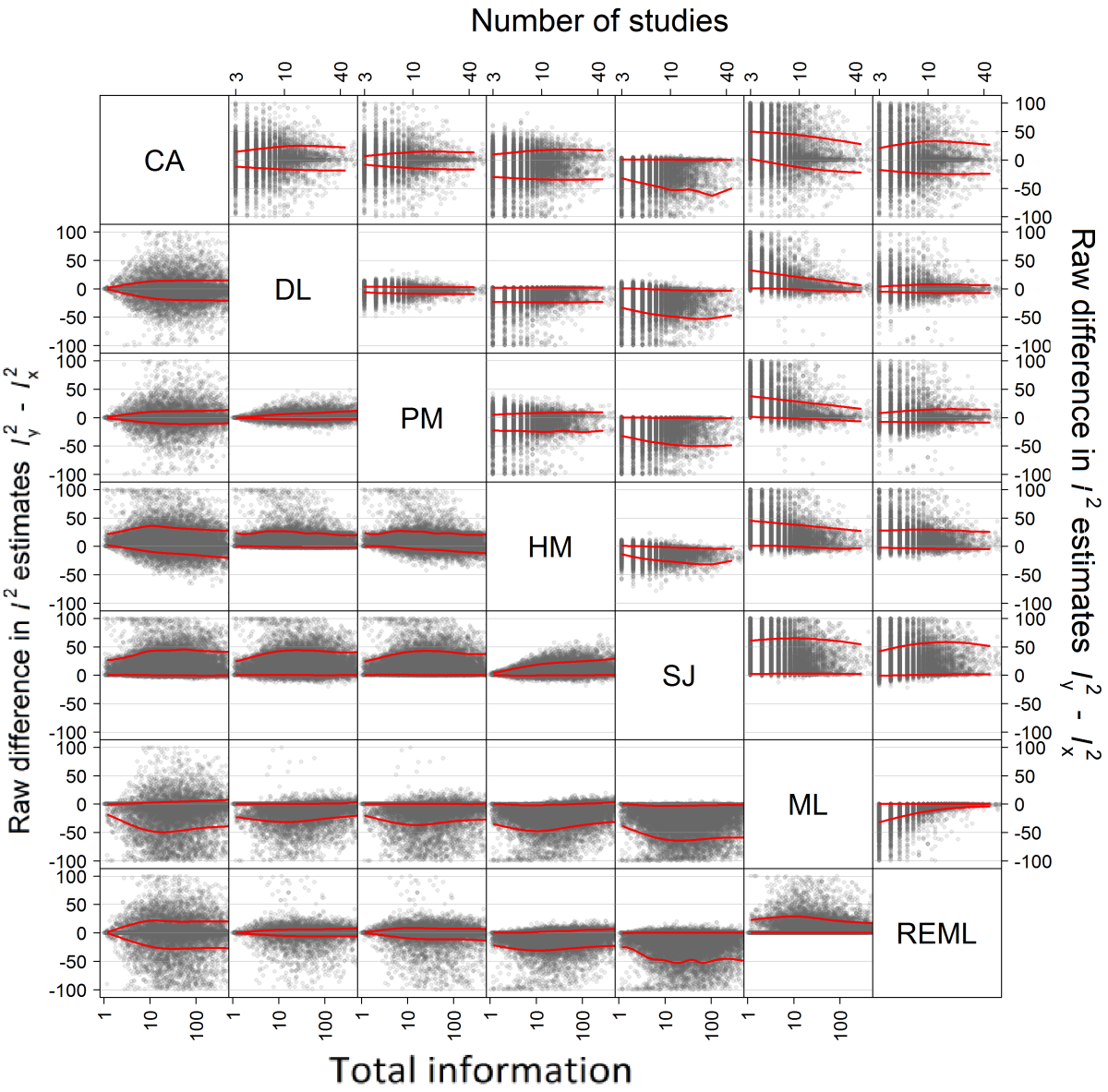


Figure 4.6: Bland-Altman scatter plots comparing differences in I^2 estimates against (upper-right panel) the number of studies and (lower-left panel) the total information

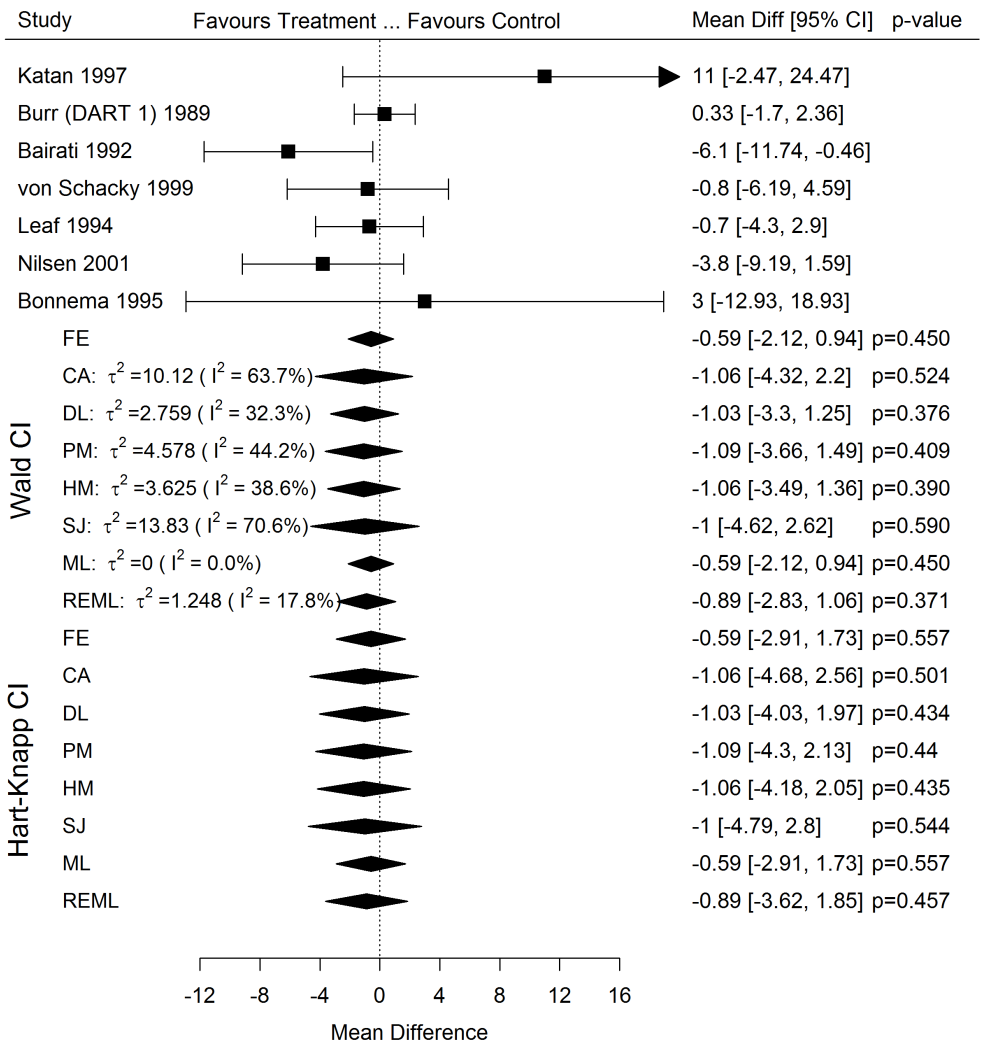


Figure 4.7: Forest plot of a meta-analysis of seven studies, with combined effects illustrated from various methods of heterogeneity variance estimation *Wald-type confidence intervals for the summary effect presented.*

The forest plot in figure 4.7 shows a meta-analysis of trials comparing the effect of high and low doses of omega-3 fatty acids in relation to systolic blood pressure (SBP) at the end of the study [52]. Seven studies were combined, originally by the random-effects inverse variance method with a DL estimate of the heterogeneity variance and summary effect confidence interval calculated using the Wald-type and Hartung-Knapp methods. The systematic review was used to inform guidance from The US Food and Drug Administration [29], although SBP was a secondary rather than a primary outcome. DL estimated I^2 as 32.3% with I^2 estimates ranging from the ML estimate of 0% to the SJ estimate of 70.6%. Estimates of the overall mean difference were affected by differences in heterogeneity variance estimates, ranging from -1.09 to -0.59. The Wald-type 95% confidence interval around the pooled effect when the SJ estimate of the heterogeneity variance is used is 2.37 times the size of the equivalent confidence interval derived using the ML estimate. By contrast, the Hartung-Knapp method produced confidence intervals with a up to 1.6 times difference in width (SJ vs ML also). All summary effects were not statistically significant with p-values ranging from 0.371 to 0.590. Therefore, choice of heterogeneity variance estimator did not affect inferences despite notably different estimates.

Example 2. Interventions used to improve control of blood pressure in patients with hypertension

The second example is from a systematic review of interventions to improve control of blood pressure in patients with hypertension [34]. Figure 4.8 shows the forest plot of a meta-analysis comparing educational interventions directed at the physician versus a control group. The outcome of the meta-analysis is whether the patient was able to control their blood pressure, and intervention groups are compared in the form of an odds ratio. The Cochrane systematic review containing this meta-analysis has informed health practice: it has been used in a NICE guideline for treating hypertension [91] and is referenced in a related Cochrane review [84]. The original fixed-effect analysis suggests a statistically significant result in favour of active intervention ($p < 0.001$). The I^2 statistic varied considerably between the CA

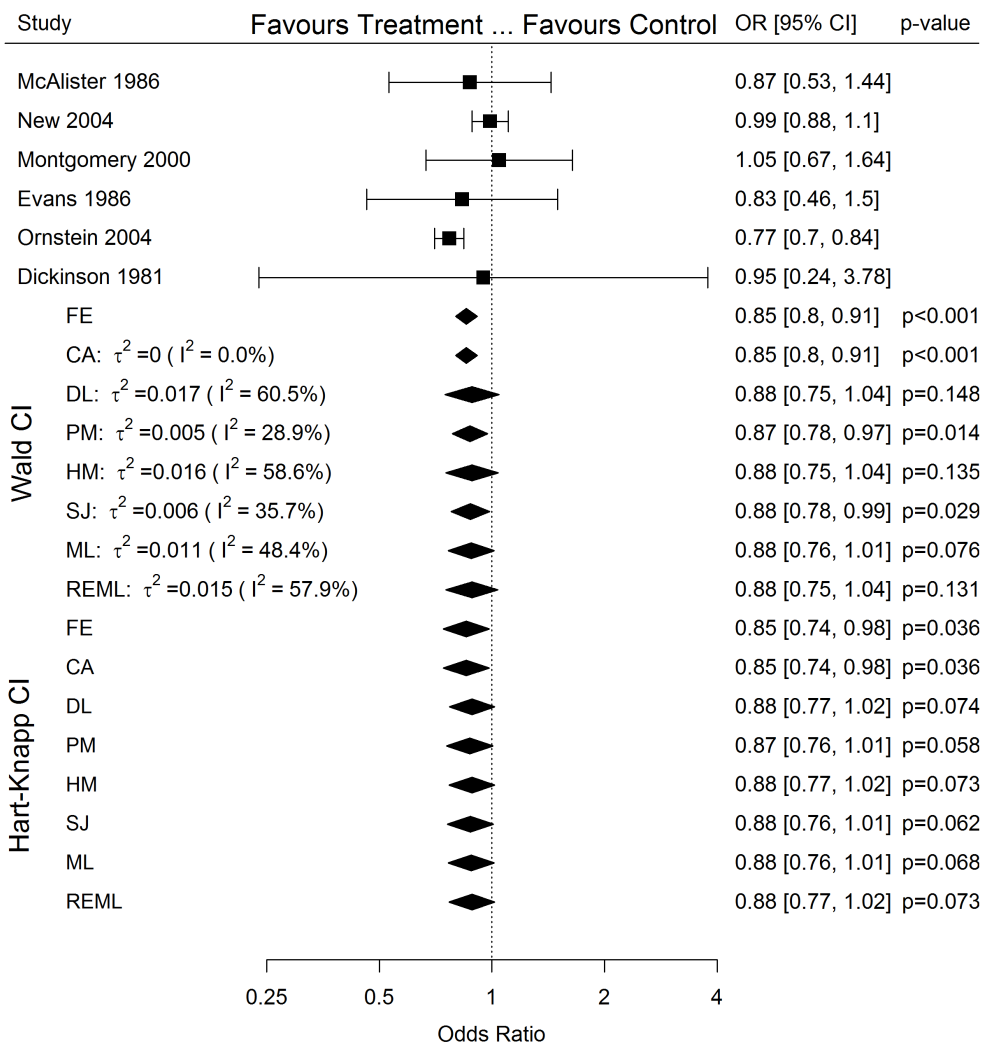


Figure 4.8: Forest plot of a meta-analysis of six studies, with combined effects illustrated from various methods of heterogeneity variance estimation
Wald-type confidence intervals for the summary effect presented.

estimate of 0% and the DL estimate of 60.5%. When an inverse-variance random-effects approach is used to combine the studies, results vary considerably with p-values ranging between < 0.001 and 0.148, even though there are minimal differences between pooled effect estimates. Therefore, using the Wald-type method, conclusions of this meta-analysis might change considerably depending on whether fixed- or random-effects are assumed as well as on which method of heterogeneity variance estimation is chosen. Conclusions are more robust using the Hartung-Knapp method.

4.5 Discussion

I have shown that there is considerable inconsistency in findings of meta-analyses when different methods are used to estimate heterogeneity variance in a random-effects model. In some cases, I^2 estimates differed by more than 50% in absolute value. In extreme cases, one method can produce an I^2 statistic of 0% while a different method can produce an I^2 close to 100%. Extreme inconsistencies mostly occurred when one method estimated $I^2 = 0\%$. Some methods are more consistent with each other, such as the DerSimonian-Laird (DL) and Paule-Mandel (PM) estimators. These methods give perfectly consistent results where $I^2 = 0\%$, but absolute differences in I^2 may still be up to 25%. Sidik-Jonkman (SJ) and Hartung-Makambi (HM) estimates had high agreement for low levels of heterogeneity and in few cases differences in I^2 exceeded 50%; this was likely due to both methods producing no I^2 estimates of 0%.

PM has been described as a simplified version of REML. Rukhin et al. [94] therefore suggested these methods should produce similar estimates of the heterogeneity variance, yet our results show low agreement between them. Rukhin et al addressed the simpler situation of a single sample of normally distributed observations arising in each study and estimated the within-study variance using REML methods rather than the maximum likelihood approach usually used for two-group studies such as

clinical trials [8, 124]. These differences may explain the low agreement between the methods.

While estimates of heterogeneity may be very different, this does not translate into substantial differences in the summary effect estimates. It can, however, lead to very different confidence intervals for the summary effect estimates, and hence to different p-values. Differences in p-values showed that conclusions of a meta-analysis may differ in a small number of cases, but frequently enough to cause concern. P-values derived from the commonly-used Wald-statistic were discordant at the 5% significance level in 10.3% of meta-analyses between at least two of the seven heterogeneity variance estimators. If the DL random-effects approach was used in all meta-analyses, 3.1% of meta-analyses would have had different conclusions (by at least two p-value categories) when at least one of the other heterogeneity variance estimates is used. If p-values were derived from the Hartung-Knapp method [38], conclusions would have changed much less frequently; p-values were discordant at the 5% level in 3.7% of meta-analyses between any two heterogeneity variance estimators. These results are consistent with an empirical study comparing DL with other heterogeneity variance estimators [117].

In a secondary analysis, I found the number of studies and total information in a meta-analysis has little impact on the level of agreement between heterogeneity variance estimates. Therefore, differences between estimates are not caused by lack of data, but perhaps by inherent differences between the heterogeneity estimation methods. The meta-analysis examples given in section 4.4 have a high level of disagreement between many summary statistics, but have no noticeable characteristics that would explain such disagreement. These findings suggest that large differences between heterogeneity variance estimates are possible in meta-analyses of all sizes.

My use of the CDSR dataset means that this analysis has some limitations. The dataset contained on average seven meta-analyses per systematic review and results within each systematic review are likely to be correlated, meaning heterogeneity variance estimates are not independent. Also, the dataset included meta-analyses

that were not really meta-analyses, in that the studies were presented together in a forest plot but not synthesised in the original publication. These 'meta-analyses' may have different characteristics from those that did combine the results and are likely to be more heterogeneous. Therefore, there may be a difference between the range of heterogeneity variance estimates in our analysis and the range reported in actual Cochrane reviews. Finally, the limitation of using empirical data is that the true heterogeneity variance is unknown and I cannot infer which method produces the closest estimate to the true value. Simulated data is required for this purpose, which I use predominantly in the rest of this thesis.

4.6 Conclusion

Differences across methods suggests the need for further research into the properties of the heterogeneity variance estimators, such as whether they give biased results. This can only be achieved through simulation rather than empirical meta-analysis data. Therefore, in the next chapter, I conducted a systematic review of simulation studies that compare heterogeneity variance estimators.

In summary, the choice of heterogeneity variance estimator can affect the results of a meta-analysis, including estimates of the degree of heterogeneity, the standard error of the summary effect and less frequently the statistical significance of results. The use of a single estimate of the heterogeneity variance may therefore lead to inappropriate conclusions in some meta-analyses. When conducting a random-effects meta-analysis, researchers should be aware that the choice of heterogeneity variance estimator may alter the conclusions drawn from the analysis. Sensitivity analysis using a wide range of plausible heterogeneity variances may be advised.

Chapter 5

A review of simulation studies to compare heterogeneity variance estimators

5.1 Introduction

In the last chapter, I presented an empirical comparison of heterogeneity variance estimators using meta-analysis data derived from the Cochrane Database of Systematic Reviews. Results showed considerable disagreement between heterogeneity variance estimates derived from many different methods, including DerSimonian-Laird [25]. Differences between these estimates may have led to discordant conclusions in a small, but noteworthy proportion of meta-analyses. These findings provide motivation for further investigation of the properties of heterogeneity variance estimators in this chapter and for the rest of my thesis.

I present in this chapter a systematic review of studies that compare heterogeneity variance estimators in simulated meta-analysis data. In studies based on simulated data, the properties of methods can be investigated. In this review, I aim to (1) identify whether there is consistent evidence across simulation studies, (2) understand how different heterogeneity methods impact on estimation of the heterogeneity variance itself, the meta-analytic summary effect and its confidence interval, (3) potentially recommend method(s) for estimating heterogeneity in practice without the need for further simulation studies; and (4) to identify areas where further simulation research may be required.

In section 5.2, I detail the methods I used to search and select simulation studies for inclusion in my review. I present a summary of the identified studies in section 5.3, including summaries of their designs in sections 5.3.1 - 5.3.4 and results from their simulations in section 5.3.5. The discussion and conclusion are in sections 5.4 and 5.5.

5.2 Methods

5.2.1 Search Strategy

I searched the databases MEDLINE, Web of Science Core Collection and JSTOR on 8th Nov 2014. Details of the search strategy are given in appendix C.1. For each of the included papers I examined reference lists and performed a citation search using Google Scholar. Search results were restricted to those written in English.

5.2.2 Eligibility criteria

I included papers if:

1. results were presented from simulated meta-analysis data;
2. simulated data were generated from a random-effects model with at least one scenario with $\tau^2 > 0$; and
3. results compared the performance of more than one heterogeneity variance estimator.

Papers were excluded if they contained only the following types of simulated data:

1. network meta-analyses;
2. one-stage individual participant data (IPD) meta-analyses;
3. meta-analyses of diagnostic accuracy studies; or
4. meta-regression, if covariates were responsible for all heterogeneity present.

5.2.3 Data extraction

I extracted details of the methods used to simulate data from each study. This included parameter values for the heterogeneity variance, summary effect and covariates for simulations of meta-regression data. Other method details include the type of outcome measure (e.g. odds ratio), the number of studies in each meta-analysis and the distributions for generating study effects and study sample sizes. Findings from simulation results relating to the performance of each heterogeneity variance estimator were collated and summarised. I also extracted raw study results (where available) to attempt a formal synthesis, but this was not feasible due to differences in the design of the included studies. Instead, I graphically present selected results from studies in which raw results were provided. Recommendations from the concluding sections of each paper were extracted, including which heterogeneity variance estimator(s) the authors thought performed best and any other general recommendations for heterogeneity variance estimation.

5.3 Results

5.3.1 Search Results

The database search returned 1,472 matches in MEDLINE, 1,918 matches in Web of Science Core Collection and 530 matches in JSTOR with a total of 3,225 non-duplicate matches. Ten publications met the eligibility criteria. I identified a further two from searching reference lists and citations [79, 102], leading to a total of 12 included simulation studies, which are listed in table 5.1. Among the 12 simulation studies, eight proposed new methods for heterogeneity estimation and then conducted a simulation study to compare the methods with existing methods. Sidik and Jonkman conducted two simulation studies [101, 102] and proposed new methods in each; their 2007 study was intended to supersede the earlier 2005 study. The remain-

ing four publications were simulation studies that only compared existing methods [78, 79, 96, 124].

5.3.2 Simulation methods and parameter values

Table 5.1 details methods that were used to simulate meta-analysis data in each study. Six (50%) studies simulated meta-analyses containing studies with a binary outcome, four of which used an odds ratio effect measure and two used a relative risk effect measure. Four (33%) studies simulated continuous outcome meta-analyses, all of which used a standardised mean difference effect measure, and one study also used the 'unstandardised' mean difference [124]. In both binary and continuous meta-analyses, study sample sizes were most commonly generated from a uniform distribution [78, 79, 101, 102]. The within-study variance of each study was then derived from these sample sizes. Only three (25%) studies simulated sample sizes or within-study variances using more than one set of parameter values [79, 93, 124].

Three (25%) studies simulated meta-analysis data with a generic effect measure [17, 64, 93]. These studies allow investigation of the properties of the estimators without conflation with estimation of specific outcome measures. Generic study effects in these three studies were simulated directly from the random-effects model. Chung et al. [16] and Kontopantelis et al. [64] used known within-study variances to calculate heterogeneity variance estimates, while Rukhin [93] used within-study variances estimated from simulated participant-level data. These within-study variances are rarely known in practice and therefore results from Chung et al. [16] and Kontopantelis et al. [64] represent the performance of heterogeneity variance estimators under ideal conditions.

All simulation studies presented results for a range of heterogeneity variance parameters, including zero in all but one case [64]. All studies except Kontopantelis et al. [64] gave no reasoning for their choice of parameter values or simply stated they chose values to reflect real meta-analyses in practice. In these studies, there was

Reference	**Effect measure**[1]	**Number of studies (k)**	**Heterogeneity variance (τ^2)**	**Summary effect (θ)** [2]	**Sample sizes / standard error of effect** [3]	**Event rate in control group (p_i) (binary outcomes only)**
Knapp and Hartung [61]	RR	5 - 15	0, 0.05, 0.1, 0.2, 0.3	-0.5	Sampled from example meta-analysis with replacement (sample size range 139 to 88,391 and split equally between arms)	0.05, 0.1
Panityakul et al. [79]	RR	10, 30	0 to 0.5 (by 0.05)	0	$n_{1i} \sim U(10, 30)$ or $n_{1i} \sim U(100, 300)$ (small or large studies), $n_{1i} = n_{2i}$	0.15
Bhaumik et al. [3]	OR	20	0, 0.2, 0.4, 0.6, 0.8, 1, 1.2	0	$n_{1i}, n_{2i} \sim U(50, 1000)$, n_{1i} and n_{2i} drawn independently and therefore unequal	0.004, 0.006, 0.01, ..., 0.99
Sidik and Jonkman [101]	OR	10 - 80	0, 0.10, 0.25,0.5, 0.75, 1, 1.25, 1.50, 2	0.5	$n_{1i} \sim U(20, 200)$, $n_{2i} \sim U(30, 300)$ (unequal n_{1i} and n_{2i} sampled independently)	$\sim U(0.05, 0.65)$
Sidik and Jonkman [102]	OR	10 - 50	0 - 0.5 (by 0.1) and 0.5 - 1.75 (by 0.25)	-0.5, 0, 0.5	$n_{1i}, n_{2i} \sim U(20, 200)$, $n_{1i} = n_{2i}$	$\sim U(0.05, 0.65)$

Reference	**Effect measure**[1]	**Number of studies (k)**	**Heterogeneity variance (τ^2)**	**Summary effect (θ)** [2]	**Sample sizes / standard error of effect**[3]	**Event rate in control group (p_i) (binary outcomes only)**
Novianti et al. [78]	OR	10 - 50	0, 0.5, 1, 1.5	0, 0.5	$n_{1i}, n_{2i} \sim U(20, 200)$, $n_{1i} = n_{2i}$	$\sim U(0.05, 0.65)$
	SMD	10 - 50	0, 0.0122, 0.0244, 0.0366	0, 0.5	$n_{1i}, n_{2i} \sim U(20, 200)$, $n_{1i} = n_{2i}$	NA
Malzahn et al. [74]	SMD	15	0, 0.09, 0.25, 1, 4	0.5	$N_i \sim U(11, 64)$ (unequal n_{1i} and n_{2i} with a ratio between 11/29 and 35/13)	NA
Sanchez-Meca and Marín-Martínez [96]	SMD	5 - 100	0, 0.04, 0.08, 0.16, 0.32	0.5, 0.8	n_{1i} and n_{2i} constant between meta-analyses in the same scenario. $n_{1i} = n_{2i}$ and sampled from skewed distribution with means 30, 50, 80 and 100	NA
Viechtbauer [124]	SMD	5 - 80	0, 0.001, 0.025, 0.05, 0.1	0, 0.2, 0.5, 0.8	$n_{1i} \sim N\left(m_j, \frac{1}{3}m_i\right)$ where $m_j = \{20, 40, 80, 160, 320\}$ and $n_{1i} = n_{2i}$	NA
	MD	5 - 80	0, 0.125, 0.25, 0.5, 1	0, 1, 2, 4	$sd_i = \{1, 0.5, 0.25, 0.125, 0.0625\}$, n_{1i} and n_{2i} as above	NA

Reference	**Effect measure**[1]	**Number of studies (k)**	**Heterogeneity variance (τ^2)**	**Summary effect (θ)** [2]	**Sample sizes / standard error of effect**[3]	**Event rate in control group (p_i) (binary outcomes only)**
Chung et al. [16]	G	5, 10, 30	0, 0.01, 0.05, 0.1, 0.2	0.5	σ_i^2 simulated from $0.25 \cdot \chi^2$ distribution and restricted to the range 0:009 - 0:6 (σ_i^2 assumed to be known,$\sigma_i^2 = \hat{\sigma}_i^2$)	NA
Kontopantelis et al. [64]	G	2 - 20	0.01, 0.03, 0.1	0.5 (sampled from normal, skew-normal, uniform, bi-model and double-spiked distributions)	σ_i^2 simulated from $0.25 \cdot \chi^2$ distribution and restricted to the range 0:009 - 0:6 (σ_i^2 assumed to be known,$\sigma_i^2 = \hat{\sigma}_i^2$)	NA
Rukhin [93]	G	3, 10	0 - 2 (by 0.1)	not clear	$N_i \sim U(4, 12)$, $\sigma_i^2 \sim \frac{(N_i - 3)}{\chi^2_{N_i-2}}$ (inverted χ^2-distribution such that $E[\sigma_i^2] = 1$)	NA

Table 5.1: Simulation methods, parameter values and performance measures used in the 12 included simulation studies

[1] *RR=relative risk; OR=odds ratio; SMD=standardised mean difference; MD=mean difference; G=generic.*

[2] *Underlying study effects sampled from the normal distribution unless otherwise stated.*

[3] *n_{1i} and n_{2i} denotes the sample size in the control and treatment groups respectively, the total sample size is denoted by N_i ($N_i = n_{1i} + n_{2i}$ in two-grouped studies). sd_i denote the standard deviation in both treatment and control groups respectively (equal between treatment groups in all publications and $sd_i = 1$ in SMD meta-analyses).*

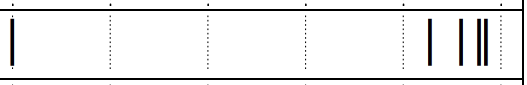
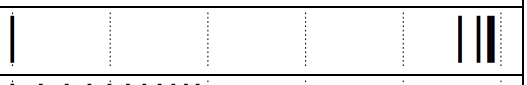
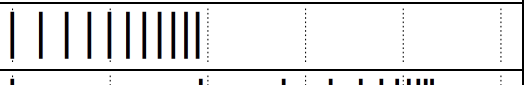
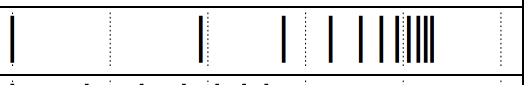
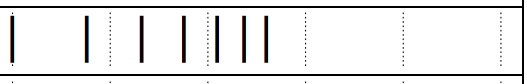
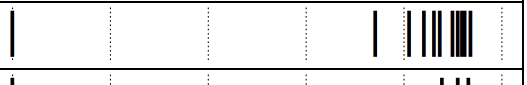
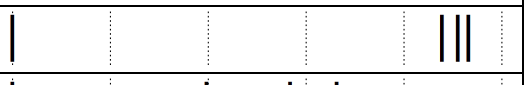
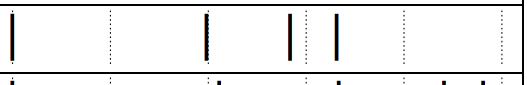
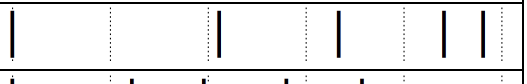
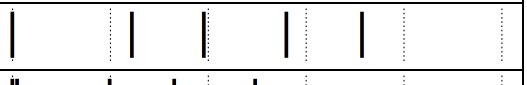
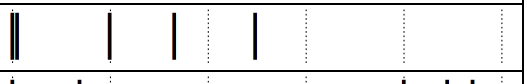
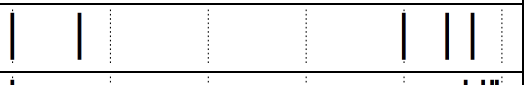
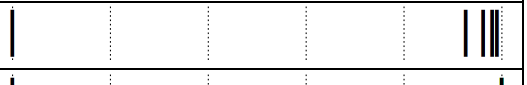
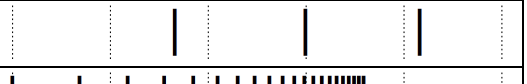
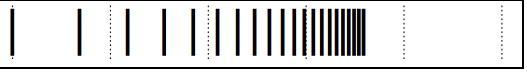
Reference	τ^2 parameter values	Scen-ario	Corresponding mean I^2 (%)
Knapp and Hartung (2003)	0, 0.05, 0.1, 0.2, 0.3	$p_{1i} = 0.05$	
		$p_{1i} = 0.1$	
Panityakul et al (2013)	0 to 0.5 (by 0.05)	Small studies	
		Large studies	
Bhaumik et al (2012)	0, 0.2, 0.4, 0.6, 0.8, 1, 1.2	$p_{1i} = 0.004$	
		$p_{1i} = 0.996$	
Sidik and Jonkman (2005)	0, 0.10, 0.25, 0.5, 0.75, 1, 1.25, 1.50, 2	-	
Sidik and Jonkman (2007)	0 - 0.5 (by 0.1) and 0.5 - 1.75 (by 0.25)	-	
Novianti et al (2014) *	0, 0.5, 1, 1.5 (OR)	-	
	0, 0.0122, 0.0244, 0.0366 (SMD)	-	
Malzahn et al (2000)	0, 0.09, 0.25, 1, 4	-	
Sanchez-Meca and Marin-Martinez (2008)	0, 0.04, 0.08, 0.16, 0.32	-	
Viechtbauer (2005)	0, 0.001, 0.025, 0.05, 0.1 (SMD)	Small studies	
		Large studies	
	0, 0.125, 0.25, 0.5, 1 (MD)	Small studies	
		Large studies	
Chung et al (2013) *	0, 0.01, 0.05, 0.1, 0.2	-	
Kontopantelis et al (2013) *	0.01, 0.03, 0.1	-	
Rukhin (2013)	0 - 2 (by 0.1)	-	

Table 5.2: Underlying ranges of I^2 in each publication

** Similar I^2 values are also reported in the original articles. I^2 values in* Kontopantelis et al. [64] *differ from those in the table because they used estimates of I^2 based on the formula dependent on the Q statistic (see formula 1.5). I^2 values in* Chung et al. [16] *differ because they were calculated individually for all k.*

inconsistency between the ranges of heterogeneity variances; Viechtbauer [124] used parameter values up to 0.1 and Malzahn et al. [74] used parameter values up to 4 to simulate SMD meta-analyses. Kontopantelis et al. [64] chose parameter values that correspond to a range of low, moderate and high I^2 (as defined in the study). I derived the range of underlying I^2 values for all simulation studies and present them in table 5.2[1]. There was low consistency between the range of I^2: seven studies (58%) contained only meta-analyses with non-zero underlying I^2 values greater than 40%. Nevertheless, I use the terms 'low', 'moderate' and 'high' heterogeneity in the rest this review as they were used in the original publications.

Over all publications identified in this review, meta-analyses were simulated that contain 2 - 100 studies. Six (50%) publications include simulated meta-analyses with fewer than 10 studies and five (42%) publications include simulated meta-analyses with 50 or more studies.

5.3.3 Performance measures

Performance measures reported from each simulation study are listed in Table 5.3. Ten studies compared bias of heterogeneity variance estimators. Nine compared the variance, efficiency or MSE; I define these three performance measures as measures of variability of heterogeneity variance estimates, and refer to them as such in the rest of this paper. For details on how these measures are calculated, see appendix C.2.

Many of the studies reported performance measures to quantify the impact of heterogeneity variance estimators on other commonly reported statistics in meta-analysis. Three reported the performance of estimates of the summary effect [93, 96, 101]. Six compared the coverage of 95% confidence intervals for the summary effect. Kontopantelis et al. [64] also reported performance of confidence interval for the summary

[1] I^2calculated from the formula $I^2 = 100 \cdot \tau^2 / \left(\tau^2 + \sigma^2\right)$, where σ^2 is the typical variance and derived from 1000 replications of each simulated scenario.

Simulation study	Heterogeneity variance					Summary effect		Confidence interval for mean effect	
	Bias	Mean Squared Error (MSE)	Variance	Efficiency	Probability of zero estimate	Bias	Mean Squared Error (MSE)	Coverage	Error interval
Knapp and Hartung (2003)	✓		✓					✓	
Panityakul et al (2013)	✓	✓							
Bhaumik et al (2012)	✓								
Sidik and Jonkman (2005)	✓	✓				✓	✓	✓	
Sidik and Jonkman (2007)	✓	✓							
Novianti et al (2014)	✓		✓						
Malzahn et al (2000)	✓		✓						
Sanchez-Meca and Marin-Martinez (2008)						✓		✓	
Viechtbauer (2005)	✓	✓		✓					
Chung et al (2013)	✓	✓			✓			✓	
Kontopantelis et al (2013)	✓				✓			✓	✓
Rukhin (2013)		✓					✓	✓	
Total	10	6	3	1	2	2	2	6	1

Table 5.3: Summary of performance measures reported in the 12 included simulation studies

effect in terms of mean error interval estimates, that is, the average ratio between observed and actual 95% confidence interval widths.

Given the range of reported performance measures, I present the performance of heterogeneity variance estimators in three sections: properties of the point estimate of heterogeneity in section 5.3.5, properties of the point estimate of the summary effect in section 5.3.6 and properties of confidence intervals for the summary effect in section 5.3.7.

5.3.4 Heterogeneity variance estimators

Table 5.4 shows which heterogeneity variance estimators were compared in each simulation study. DerSimonian-Laird was included in all 12 studies. Other estimators frequently included were Cochran's ANOVA (CA), restricted maximum likelihood (REML), maximum likelihood (ML), Paule-Mandel (PM) and Sidik-Jonkman (SJ).

Rukhin's Bayesian estimators (RU, B0, BP, SB) were only included in two simulation studies only with a generic outcome measure [64, 93]. Methods including the bootstrap DerSimonian-Laird (DL_B), positive DerSimonian-Laird (DL_P), two of Rukhin's estimators (RU, SB) and Bayes Modal (BM) have only been compared in the one study in which they were initially proposed. For details on each heterogeneity variance estimator, see chapter 2. I frequently refer to heterogeneity variance estimators by their acronyms in the rest of this chapter; the acronym definitions are given in table 5.4.

5.3.5 Performance of point estimators of the heterogeneity variance

In this section, I summarise the properties of heterogeneity variance estimators identified in this review in terms of bias, variability (e.g. mean squared error), and the proportion of zero estimates they produce. Estimators are summarised together when they are slight variations of the same method or have similar properties. This section excludes the Hartung and Makambi estimator, which has only been compared in terms of performance measures relating to the summary effect and its confidence interval [96] (see sections 5.3.6 and 5.3.7 for these results).

Selected results are presented in figures 5.1 and 5.2; these were recreated from the raw study results in the supplementary material of Novianti et al. [78] to back up some of the findings in this section. These raw study results included the performance measures bias and variance, but I present bias and mean squared error. The mean squared error is a more meaningful measure of performance than the variance and could be derived from the raw results given.

I present those compared in many simulation studies first, beginning with the DerSimonian-Laird estimator.

Estimators *	Acronym	Knapp and Hartung (2003) [1]	Panityakul et al (2013)	Bhaumik et al (2012)	Sidik and Jonkman (2005)	Sidik and Jonkman (2007)	Novianti et al (2014)	Malzahn et al (2000)	Sanchez-Meca and Marin-Martinez (2008)	Viechtbauer (2005) [2]	Chung et al (2013)	Kontopantelis et al (2013)	Rukhin (2013)	Total
DerSimonian-Laird (1986)	**DL**	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	12
Positive DerSimonian-Laird (2013)	**DL_P**											✔✔		1
Bootstrap DerSimonian-Laird (2013)	**DL_B**											✔✔		1
Cochran's ANOVA (1985)	**CA**		✔			✔	✔	✔	✔	✔	✔	✔	✔	9
Paule-Mandel (1982)	**PM**	✔	✔	✔		✔	✔						✔	6
Two-step Paule-Mandel (CA *a priori*) (2007)	**PM_{CA}**											✔		1
Two-step Paule-Mandel (DL *a priori*) (2007)	**PM_{DL}**			✔			✔					✔		3
Hartung-Makambi (2003)	**HM**								✔					1
Sidik-Jonkman (2005)	**SJ**		✔		✔✔	✔	✔		✔			✔		6
Sidik-Jonkman (CA *a priori*) (2007)	**SJ_{CA}**					✔✔	✔					✔		3
Hunter-Schmidt (2004)	**HS**								✔	✔				2
Maximum likelihood (1996)	**ML**		✔			✔			✔	✔	✔	✔		6
Restricted maximum likelihood (1977)	**REML**		✔			✔	✔		✔	✔	✔	✔		8
Approximate restricted maximum likelihood (1983)	**ARML**	✔												1
Rukhin (unbiased) (2013)	**RU**												✔✔	1
Rukhin (zero prior) (2013)	**B0**											✔	✔✔	2
Rukhin (simple) (2013)	**BP**											✔	✔✔	2
Rukhin (alternate) (2013)	**SB**												✔✔	1
Bayesian modal (2013)	**BM**										✔✔			1
Malzahn, Bohning and Holling (SMD only) (2000)	**MBH**	-	-	-	-	-	-	✔✔	✔		-	-	-	2

Table 5.4: Summary of heterogeneity variance estimators compared in the 12 included simulation studies

Key: ■ Binary outcome, □ continuous outcome and ■ generic outcome meta-analyses simulated.

✔✔ Signifies where estimators are compared and also proposed in the same publication.

** For references and details of each heterogeneity variance estimator, see chapter 2.*

[1] Heterogeneity variance estimates calculated using both smoothed and usual with-study variance estimates.

[2] Performance of estimators reported using non-truncated estimates.

5.3.5.1 DerSimonian-Laird (DL)

Performance of the DL estimator is documented in all 12 publications, which generally suggest that DL is negatively biased when the level of heterogeneity is moderate to high. The negative bias is more prominent when within-study variance estimates are imprecise, such as in SMD meta-analyses with small study sample sizes [74] and in binary outcome meta-analyses [79, 102], particularly when there are few events occurring in each study [3]. This negative bias can be observed in figure 5.1 (top right), where heterogeneity is high and meta-analyses have an OR outcome measure. Minimal negative bias was observed in continuous outcome meta-analyses with moderate study sample sizes [124] and binary outcome meta-analyses with large study samples sizes [61]. When within-study variances are known, Kontopantelis et al. [64] showed that DL becomes asymptotically unbiased as the number of studies in a meta-analysis increase; Viechtbauer [124] previously noted this in theory. Novianti et al. [78] showed that DL remains biased in binary outcome meta-analyses as the number of studies increases. In terms of mean squared error (MSE), DL performs relatively poorly in scenarios where negative bias is also observed [102]. In continuous and generic outcome meta-analyses, DL has a relatively low MSE and comparable performance to REML [16, 124].

Kontopantelis et al. [64] proposed a bootstrap version of DL (DL_B), with the aim of reducing the proportion of zero heterogeneity variance estimates. DL_B had the least number of zero estimates out of all methods that allowed zero estimates. In small meta-analyses (2-3 studies), DL_B has the highest positive bias of all estimators compared and comparable bias in meta-analyses with 5-10 studies [64].

Kontopantelis et al. [64] also proposed a positive version of DL (DL_P), which truncates heterogeneity variance estimates below 0.01. DL_P was one of the least biased estimators when the level of heterogeneity was low-to-moderate. However, this result may be misleading because 0.01 is also the lowest heterogeneity variance parameter value chosen in this study.

5.3.5.2 Cochran's ANOVA (CA)

CA has a small positive bias under most simulated conditions and remains positively biased even when the level of heterogeneity is high, unlike many other estimators such as DL, PM and REML [78, 79]. This can be observed in figure 5.1. Chung et al. [16] and Sidik and Jonkman [102] showed CA had the highest MSE out of all estimators compared, because of its large bias. Chung et al. [16] also found that CA had the highest percentage of zero estimates despite its positive bias in the same simulated conditions. It was, however, the least biased estimator considered by Panityakul et al. [79] when study sample sizes are small.

5.3.5.3 Paule-Mandel (PM), and its variants ($\mathbf{PM_{CA}}$ and $\mathbf{PM_{DL}}$)

PM was compared in six simulation studies, with five reporting bias and five reporting some measure of variability (see table 5.3). Novianti et al. [78] showed that PM performs well in terms of bias in SMD outcome meta-analyses, although it is comparable with many other estimators including CA, DL, PM_{DL}, SJ, SJ_{CA} and REML. PM was compared with self-proposed Bayesian estimators by Rukhin [93], in simulated generic outcome meta-analyses (see section 5.3.5.9 for performance of Bayesian estimators); results showed that PM has a lower MSE when the level of heterogeneity is low, but has a larger MSE than all Rukhin's Bayesian estimators for moderate to large levels of heterogeneity. In binary outcome meta-analyses [78, 79, 102], PM is negatively biased for high levels of heterogeneity but to a lesser extent than DL and REML, and approximately unbiased when study sample sizes are large (between 100 and 300 per group) [79]. Sidik and Jonkman [102] showed that PM is comparable with SJ_{CA} and both perform well in terms of bias and MSE. PM performs well overall but some of the most comprehensive simulation studies that included many estimators did not include PM [64, 124]. In particular, there is relatively little evidence for PM in continuous outcome meta-analyses.

PM_{CA} and PM_{DL} estimators are two-step versions of PM that use CA and DL as

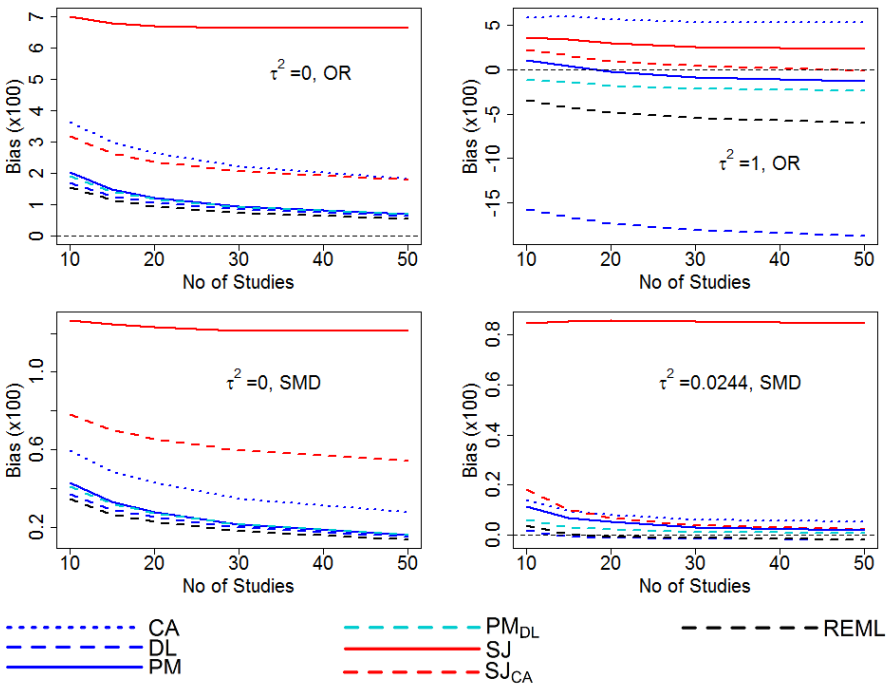


Figure 5.1: Mean bias from selected simulation results in Novianti et al. [78] including simulated meta-analyses of type SMD and OR.
The following parameters remain constant throughout these results: $n_{1i}, n_{2i} \sim U(20, 200)$, $n_{1i} = n_{2i}$, $p_i \sim U(0.05, 0.65)$ *and* $\theta = 0.5$

initial estimates respectively. Kontopantelis et al. [64] included both $\mathrm{PM_{CA}}$ and $\mathrm{PM_{DL}}$ in their comparisons, Novianti et al. [78] and Bhaumik et al. [3] included $\mathrm{PM_{DL}}$ only. Results showed that $\mathrm{PM_{CA}}$ and $\mathrm{PM_{DL}}$ have a level of bias comparable with PM (see figure 5.1). No publication reported the variability of $\mathrm{PM_{CA}}$ or $\mathrm{PM_{DL}}$ estimates (although Novianti et al. [78] included the variance of estimates as raw supplementary data).

5.3.5.4 Restricted maximum likelihood (REML) and its approximation (ARML)

REML was included in seven simulation studies: six reported bias in heterogeneity variance estimates and five reported some measure of variability (see Table 5.3 for details). In meta-analyses with moderate study sample sizes, REML becomes negatively biased as the level of heterogeneity increases, but to a lesser extent than DL

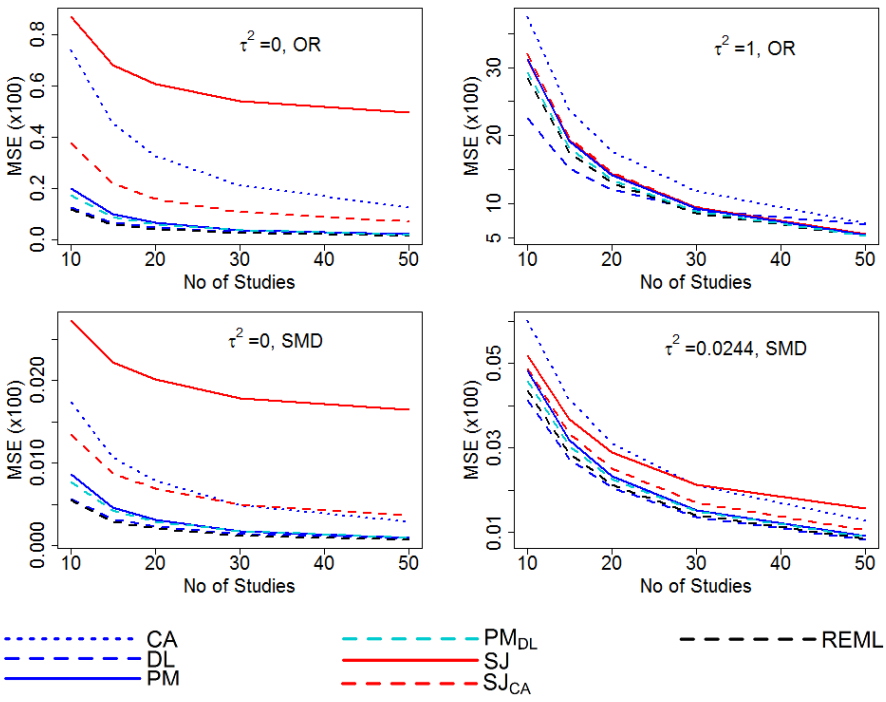


Figure 5.2: Mean squared error from selected simulation results in Novianti et al. [78] including simulated meta-analyses of type SMD and OR.
The following parameters remain constant throughout these results: $n_{1i}, n_{2i} \sim U(20, 200)$, $n_{1i} = n_{2i}$, $p_i \sim U(0.05, 0.65)$ *and* $\theta = 0.5$

or ML [78, 79, 102] as shown in figure 5.1 (top right). For low levels of heterogeneity and when studies are homogeneous, REML and DL have similar levels of bias [102]. For large study sample sizes (between 100 and 300 per group), Panityakul et al. [79] showed REML to be approximately unbiased, performing better in terms of bias than the other estimators compared, including CA, DL, PM, SJ and ML. Viechtbauer [124] also showed that, when studies typically have a large sample size, REML has a lower MSE than the other estimators compared including CA, DL, HS and ML.

Knapp and Hartung [61] compared the approximate restricted maximum likelihood (ARML) in terms of bias and variance in simulated OR meta-analyses. For all levels of heterogeneity above zero, ARML had a greater negative bias than DL and PM. ARML had the lowest variance of the three estimators compared, but this may be as a consequence of its negative bias.

5.3.5.5 Maximum likelihood (ML)

ML was included in six studies, of which five reported bias of heterogeneity variance estimates and three reported some measure of variability (see Table 5.3 for details). ML tends to underestimate the heterogeneity variance and is one of the least positively biased estimators when there is no underlying heterogeneity. ML has a negative bias particularly when there are fewer than 10 studies in the meta-analysis [16] and produces more zero estimates when there are fewer than five studies [16]. As a consequence of the negative bias, ML performs well in terms of MSE when the level of heterogeneity is low.

5.3.5.6 Hunter-Schmidt (HS)

HS was included in two studies [96, 124], of which only one reported bias and some measure of variability (variance) [124]. Viechtbauer [124] stated that the performance of HS is comparable with ML in terms of bias and MSE and grouped their results together.

5.3.5.7 Sidik- Jonkman estimators (SJ and SJ_{CA})

SJ is a two-step non-truncated heterogeneity variance estimator that only produces positive estimates of heterogeneity and was included in six simulation studies. Five publications reported bias and four reported some measure of variability. Results from all simulation studies showed that SJ is positively biased for small to moderate levels of heterogeneity and when study sample sizes are below 200 [78, 102]; this is illustrated in all scenarios in figure 5.1. For meta-analyses containing studies with larger sample sizes (between 100 and 300 per group), SJ has relatively a small amount of positive bias [79]. SJ's positive bias has been demonstrated in OR [78, 102], SMD [78] and generic [64] outcome meta-analyses and can be attributed to SJ being a non-truncated estimator that only produces positive heterogeneity variance estimates.

SJ_{CA} is derived from the same two-step approach as SJ but uses CA as an initial heterogeneity variance estimate for the first iterative step. Three studies compared SJ_{CA} with SJ and found that SJ_{CA} has less positive bias when the level of heterogeneity is low [64, 78, 102]. SJ and SJ_{CA} have a comparable MSE in all the reported results by Sidik and Jonkman [102].

5.3.5.8 Malzahn, Bohning and Holling (MBH)

MBH was included in two studies and can only be used in meta-analyses with a SMD outcome measure. Malzahn et al. [74] compared MBH with DL and CA in terms of bias and MSE. However, the results of this simulation were not given: the paper states only that MBH has a smaller MSE than CA. Sanchez-Meca and Marín-Martínez [96] only compared MBH in terms of performance measures relating to the overall summary effect and its confidence intervals.

5.3.5.9 Bayesian estimators: Bayesian modal (BM) and Rukhin's estimators (RU, B0, BP, SB)

Chung et al. [16] was the only study to include the Bayesian modal estimator (BM), and did so in simulated generic outcome meta-analyses. BM is a non-truncated estimator that only produces positive estimates. Therefore results showed that BM is more positively biased than CA, DL, ML and REML for low levels of heterogeneity and a larger MSE. For moderate to high levels of heterogeneity, BM has low bias comparable with DL and REML and performs better in terms of MSE.

Rukhin proposed a series of Bayesian heterogeneity variance estimators that only produce positive estimates, and compared them with CA, DL and PM in terms of MSE. All Rukhin's estimators have comparable MSE when the level of heterogeneity is moderate to high, which is lower than CA, DL and PM. When the level of heterogeneity is low, PM performs better than Rukhin's estimators in terms of MSE.

Kontopantelis et al. [64] also compared the bias of B0 and BP with many other estimators and found that BP had the highest positive bias of all estimators compared. B0 and BP remained positively biased in large meta-analyses containing 20 studies.

5.3.6 Performance of estimators of the summary effect

Three studies included performance measures relating to point estimates of the summary effect [93, 96, 101]. All of these calculated summary effects by the standard inverse variance method, where study weights are dependent on heterogeneity variance estimates calculated by a number of methods. Sidik and Jonkman [101] and Rukhin [93] found that estimates of the summary effect were unbiased, had small MSE and had a high level of agreement across all heterogeneity estimation methods. Sanchez-Meca and Marín-Martínez [96] showed $\hat{\theta}$ has a negligible amount of bias for all heterogeneity variance estimators compared.

5.3.7 Performance of confidence intervals for the summary effect

Six studies reported performance measures relating to 95% confidence intervals for the summary effect [16, 61, 64, 93, 96]. All of these used coverage as a performance measure. Three of these studies calculated Wald-type confidence intervals and reported that these are sensitive to which heterogeneity variance estimators are used and to the level of heterogeneity. For low levels of heterogeneity, coverages are above the nominal level of 95% and fall to 85-90% for moderate to high levels of heterogeneity. Coverage probabilities for t-distribution confidence intervals were reported in two studies and are also sensitive to level of heterogeneity but to a lesser extent than Wald-type confidence intervals [96]. Coverage of t-distribution confidence intervals with Hartung-Knapp variance estimates was reported in three studies; results

showed this confidence interval method is not sensitive to which method of heterogeneity estimation is used and maintains coverages close to the nominal 95% for all simulated scenarios [61, 93, 96].

5.3.8 Performance of heterogeneity variance estimators using other methods to estimate the within-study variance

Two studies investigated whether using alternatives to the usual study variance estimation methods help improve estimation of heterogeneity [3, 61]. I define the usual study variances as those calculated by methods described in chapter 1.

Knapp and Hartung [61] proposed a method for calculating within-study variances that reduces the correlation with study effects and found that using this method reduces the negative bias of DL and REML estimators. Results showed that using this method makes little difference to the coverage of confidence intervals for the summary effect.

Bhaumik et al. [3] proposed an alternative Paule-Mandel estimator, which calculates estimates of the heterogeneity variance using PM with other within-study variance estimates. Precision of these variance estimates in improved by borrowing strength from other studies in the meta-analysis. Results showed that PM heterogeneity variance estimates using this alternative method have less negative bias than DL, PM_{DL} and PM that use usual within-study variance estimates. The method can easily be applied with any heterogeneity variance estimator, but only PM was considered in this paper. The method can be applied only to odds ratio effects.

5.3.9 A summary of recommendations

Table 5.5 summarises recommendations made in the 12 publications. Ten make clear recommendations about which heterogeneity variance estimator(s) should be used in

practice. DL was included in all 12 simulation studies and was recommended twice: by Sidik and Jonkman [102] and Malzahn et al. [74]. However, in these studies, DL is compared only with SJ [74, 102] and CA [74], and in both publications is recommended only when the level of heterogeneity is low.

Three independent studies made recommendations from a comparison of only pre-existing estimators; two of these studies recommended PM [78, 79] over all other estimators compared. The other independent study, Viechtbauer [124], recommended REML in continuous outcome meta-analyses but did not include PM in the study. Novianti et al. [78] also stated that REML is a good alternative to PM in simulated SMD meta-analyses, but is not recommended in OR meta-analyses due to negative bias comparable with DL. Other estimators were recommended in the same publication where the estimator was initially proposed, including SJ_{CA} [102], BM [16], MBH [74], DL_B [64], B0 and BP [93]. Sidik and Jonkman [101] recommended their own SJ estimator, but a later (2007) recommendation of SJ_{CA} supersedes this. SJ is included in three other simulation studies [64, 78, 79] and is not recommended in any. There may have been a conflict of interest in these non-independent studies. Furthermore, many of these compared a small subset of methods, in few simulated scenarios with a limited range of heterogeneity levels (see table 5.2).

5.4 Discussion

Many papers have reported that the DerSimonian-Laird estimator of heterogeneity is negatively biased for moderate to large levels of heterogeneity, and suggest that better-performing heterogeneity variance estimators are available. In this review of comparative simulation studies, I found that the Paule-Mandel estimator generally performs well, is easy to compute and was specifically recommended in three publications from results based on both continuous and binary outcome meta-analyses [3, 78, 79]. REML was also recommended on the basis of two simulation studies of continuous outcome meta-analyses [78, 124]. However, computing REML estim-

Reference	Effect measure [1]	Heterogeneity variance estimators [2]	Recommendations / conclusions [3]
Knapp and Hartung (2003)	RR	DL, ARML, PM* (with and without using a smoothed within-study variance estimates)	No heterogeneity estimator recommended. Neither the smoothed or usual within-study variances are recommended. Smoothed are no better than 'usual' for calculating CIs of the summary effect.
Panityakul et al (2013)	RR	CA, DL, PM, SJ, ML, REML	PM recommended. Avoid CA and SJ because of positive bias.
Bhaumik et al (2012)	OR	DL, PM_{DL}, PM (PM with an without using alternative within-study variance estimates)	PM with alternative within-study variance estimates recommended for meta-analyses with OR effect measure.
Sidik and Jonkman (2005)	OR	DL, SJ	SJ recommended. DL preferred over SJ for low levels of heterogeneity. SJ recommended for Wald-type confidence intervals of the summary effect
Sidik and Jonkman (2007)	OR	CA, DL, SJ, SJ_{CA}, ML, REML, PM*	SJ where 'high levels of heterogeneity, SJ_{CA} or PM when low or moderate. SJ_{CA} preferred over PM because SJ_{CA} is easier to compute. Avoid DL, ML and REML (to a lesser extent) due to negative bias.
Novianti et al (2014)	OR & SMD	CA, DL, PM_{DL}, PM, SJ, SJ_{CA}, REML	PM and PM_{DL} recommended in meta-analyses with OR and SMD outcome measures. REML recommended as a valid alternative only in SMD meta-analyses.
Malzahn et al (2000)	SMD	CA, DL, MBH	MBH recommended. DL recommended only for low levels of heterogeneity and sample sizes are 'large'.
Sanchez-Meca and Marin-Martinez (2008)	SMD	CA, DL, HM, HS, SJ, ML, REML, MBH.	No heterogeneity estimator recommended. Simulation study focuses on confidence intervals for the summary effect
Viechtbauer (2005)	SMD & MD	CA, DL, HS, ML, REML (no estimator was truncated)	REML recommended. Avoid HS and ML due to negative bias and avoid FE meta-analysis.
Chung et al (2013)	G	CA, DL, ML, REML, BM	BM recommended. Avoid FE meta-analysis.
Kontopantelis et al (2013)	G	CA, DL, DL_P, DL_B, PM_{CA}, PM_{DL}, SJ, SJ_{CA}, ML, REML, B0, BP	DL_B recommended (to decrease the number of zero heterogeneity estimates). Avoid FE meta-analysis. Sensitivity analysis advised, particularly when a meta-analysis contains few studies.
Rukhin (2013)	G	CA, DL, PM, RU, B0, BP, SB	B0 or SB recommended for t-distribution CIs for the summary effect. BP recommended as a point estimate of the heterogeneity variance.

Table 5.5: A summary of recommendations from the 12 included publications

*[1] RR=relative risk; OR=odds ratio; SMD=standardised mean difference; MD=mean difference; G=generic. [2] Full names of estimators are given in table 5.4. [3] CI=confidence interval. *Publication refers to the empirical Bayes estimator, but this is equivalent to the Paule and Mandel estimator (PM) [93].*

ates involves a process of iteration that does not converge in a small proportion of meta-analyses [64]. Other recommended estimators included SJ_{CA}, DL_b, MBH, BM, B0, SB and BP, but recommendations came from the same publication where the method was originally proposed and therefore may be unduly influenced by conflicting interests; these estimators have been compared in few other simulation studies.

Studies show summary effect estimates are unbiased with low MSE, irrespective of which heterogeneity variance estimate is used. Wald-type confidence intervals of the summary effect are currently reported as standard in meta-analyses in Cochrane reviews, yet studies indicate that coverage depends highly on the heterogeneity variance estimate, and coverage can be as low as 85-90%. Sanchez-Meca and Marín-Martínez [96] recommended t-distribution confidence intervals with Hartung-Knapp variance estimates; this method has coverage closer to the nominal 95% and is not sensitive to the heterogeneity variance estimate used. A simulation study not included in this review (because only confidence interval methods were compared) has also called for wide-spread use of the Hartung-Knapp method [55].

There is still no overall consensus on which heterogeneity variance estimator to use in meta-analysis, in part because recommendations are based on subjective interpretation of the results and a trade-off between many performance measures. For example, ML and HS generally have low MSE, but only as a consequence of their negative bias. Viechtbauer [124] recommended REML as a compromise between bias and MSE and Novianti et al. [78] recommended PM based on bias alone. I summarised author's recommendations in section 5.3.9 in an attempt to make practical and collaborative conclusions, but my findings suggest that further research is still required.

This review has identified a number of limitations of the design of the simulation studies. All studies compared only a subset of all heterogeneity variance estimators available. This limits the conclusions of this review because, for example, PM and BM are not directly compared in any study and both have been recommended by different authors. Results described in section 5.3.8 suggest that using alternative

estimates of within-study variances improves estimation. However, evidence is based on meta-analyses with study effects generated from equal event probabilities, which represents optimal conditions for performance of these methods. Finally, most simulation studies generated study sample sizes from only one distribution. Those that simulated from a range of distributions [79, 93, 124] suggest the range of sample sizes in a meta-analysis affects the performance of these methods. More research is needed to investigate this effect.

Although the main limitations of this review stem from limitations in the evidence base, methods for this review could be improved in a number of ways. Relevant articles could have been missed from the online search by only including articles containing the word 'meta-analysis'. For example, simulation studies could have been carried out in the context of 'multiple laboratory experiments', given this is the context in which some of the estimators in chapter 2 are derived (CA and PM). Relevant articles may have been missed by restricted the search to English language only. Furthermore, the process of selecting of articles for inclusion could have been double checked by an independent reviewer to minimise the chance of human error.

In general, I found that simulations did not reflect the observed characteristics of meta-analyses in practice. 86% of meta-analyses from Cochrane reviews contain fewer than 10 studies [21], yet half of the reviewed simulation studies contained only meta-analyses with at least 10 studies [16, 61, 93, 96, 124]. Also, heterogeneity variance parameter values did not reflect the full range of levels of heterogeneity. For instance, seven studies contained only meta-analyses with non-zero underlying I^2 values greater than 40% [16, 61, 64, 74, 78, 101, 102]. Findings from this review suggest that properties of methods depend more strongly on I^2 than the heterogeneity variance parameter. I^2 depends on both the heterogeneity variance and the within-study variances.

Given that Cochrane meta-analyses typically contain few studies [21], heterogeneity variance estimates are imprecise regardless of the estimation method. As such, Kontopantelis et al. [64] suggested that a sensitivity analysis is required to test how

robust meta-analysis findings are to changes in this estimate. Other included simulation studies focused only on comparing the relative performance of estimators. In meta-analyses with few studies, there is little power to detect heterogeneity [16, 64] so truncated heterogeneity variance estimators produce a high number of zero estimates. Therefore, Chung et al. [16] recommended the non-truncated estimator, BM, and Kontopantelis et al. [64] recommended DL_B, which produces a lower number of zero estimates in comparison with DL. Meta-analyses containing a truly homogeneous group of studies and therefore zero heterogeneity is thought to be untenable in practice [48].

5.5 Conclusion

This review suggests there are better-performing heterogeneity variance estimators than the commonly used DerSimonian and Laird method. On the basis of the current evidence, the Paule-Mandel estimator may be the best alternative to calculate point estimates of heterogeneity and for calculating confidence intervals for the summary effect. Many recently proposed estimators including BM, DL_b, B0, SB and BP show promise, however, more research is required to compare them with a wider range of heterogeneity variance estimators before they can be recommended.

There are four main reasons why my recommendations based on this review are not conclusive: (1) many recommendations have been based on simulation studies proposing a new estimator, and so may have conflicts of interest, (2) they are based only on comparisons of a small subset of all heterogeneity estimation methods available, (3) they are based on simulated meta-analyses that do not reflect those found in typical systematic reviews and (4) they do not address sufficiently the practical situation that in many meta-analyses all heterogeneity variance estimates are very imprecise. Further independent simulation studies are needed to address these limitations. In the following chapters, I detail the design and present results of such

as simulation study. The study is designed in light of the identified limitations of existing simulation studies from this review.

Chapter 6

Methods for a new simulation study

6.1 Introduction

In the last chapter, I conducted a systematic review of previous studies that compared heterogeneity variance estimators in simulated meta-analysis data. Studies were in agreement that the DerSimonian-Laird estimator [25] of the heterogeneity variance is negatively biased in certain scenarios. Most studies advocated an alternative estimator, most commonly the Paule-Mandel estimator [80]. However, studies gave many other conflicting recommendations and therefore my systematic review was inconclusive overall. I suggested many reasons why studies came to conflicting conclusions, two of these reasons are: (1) most studies compared a small number of existing methods with those newly proposed and (2) recommendations were based on subjective trade-off between many performance measures. To address these issues, I propose to conduct a new simulation study that is collaborative and compares a comprehensive list of pre-existing heterogeneity variance estimators.

A study protocol was produced prior to simulating any meta-analysis data. This protocol was sent to a number of collaborators, who commented edited and approved the final protocol. These collaborators are: Mark Simmonds[1], Julian Higgins[2], Dan Jackson[3], Jack Bowden[3], Areti Angeliki Veroniki[4], Evangelos Kontopantelis[5] and Wolfgang Viechtbauer[6]. The protocol includes simulation methods for both binary and continuous outcome meta-analyses, the heterogeneity variance estimators compared and performance measures used to compare them. A summary of the design

[1]Centre for Reviews and Dissemination, University of York, York, YO10 5DD, UK

[2]School of Social and Community Medicine, University of Bristol, Bristol, UK

[3]School of Social and Community Medicine, University of Bristol, Bristol, UK

[4]Li Ka Shing Knowledge Institute, St. Michael's Hospital, 209 Victoria Street, East Building. Toronto, Ontario, M5B 1T8, Canada

[5]Centre for Health Informatics, Institute of Population Health, University of Manchester, Manchester, UK

[6]Department of Psychiatry and Psychology, Maastricht University, The Netherlands

of this new simulation study is presented here in this chapter, the original agreed document is in appendix D.3.

6.2 Aims

The aims of is simulation study are a subset of the main aims of this thesis (declared on page 19), namely:

1. Compare the relative performance of heterogeneity variance estimators in simulated data to establish which method(s) have the best properties.
2. Investigate the absolute performance of estimators in simulated data to establish if and when all methods perform poorly.
3. Investigate whether any characteristics of meta-analyses can explain the properties of estimators.
4. Recommend methods for random-effects meta-analysis and propose alternative strategies when all estimators perform poorly.

6.3 Heterogeneity variance estimators

Methods for estimating the heterogeneity variance in a random-effects model have been identified in chapter 2. For each simulated meta-analysis, heterogeneity variance estimates are calculated from the following 14 methods:

1. Cochran's ANOVA (CA) [18]
2. DerSimonian-Laird (DL) [25]
3. Paule-Mandel (PM) [80]

4. Two-step DerSimonian-Laird (PM_{DL}) [24]
5. Two-step Cochran's ANOVA (PM_{CA}) [24]
6. Hartung-Makambi (HM) [40]
7. Hunter-Schmidt (HS) [53]
8. Sidik-Jonkman (SJ) [101]
9. Sidik-Jonkman with Cochran's ANOVA initial estimate (SJ_{CA}) [102]
10. Maximum likelihood (ML) [37]
11. Restricted maximum likelihood (REML) [124]
12. Rukhin's estimator with zero prior (B0) [93]
13. Rukhin's simple estimator (BP) [93]
14. Malzahn, Böhning and Holling (MBH) [74]

Some of the estimators available (as detailed in chapter 2) are excluded from this study. Rukhin's optimal unbiased estimator (RU) [93], Rukhin's estimator with alternative prior (SB) [93] and positive-DerSimonian-Laird (DL_p) [64] have been shown to be inferior to other estimators in the systematic review of simulation studies in the previous chapter. Bayesian methods that rely on a subjective choice of prior distribution are excluded because of difficulty in objectively comparing them to frequentist methods. Methods that use alternative methods to estimate the within-study variances are excluded; these methods are beyond the scope of the thesis. Bootstrapping could be theoretically applied to any heterogeneity variance estimator so any variation of this approach is excluded.

6.4 Performance measures

Heterogeneity variance estimators are compared in terms of the following three primary performance measures:

1. Mean and median bias of the heterogeneity variance estimate
2. Mean and median squared error of the heterogeneity variance estimate
3. Proportion of zero estimates of the heterogeneity variance estimate

Furthermore, heterogeneity variance estimators are compared in terms of the following secondary performance measures relating to estimation of the mean of the random-effects distribution and its confidence interval. These are required to perform a comprehensive comparison of heterogeneity variance estimators, evaluating them not only as a point estimate of heterogeneity, but also on other meta-analysis statistics.

1. Mean absolute bias in estimate of the summary effect *
2. Mean squared error of estimate of the summary effect *
3. Coverage of 95% confidence intervals for the summary effect for the three confidence interval methods (i.e. the proportion of times the underlying summary effect falls inside the 95% confidence interval) *
4. Power to detect a significant summary effect using the three confidence interval methods **
5. Mean of the error-interval estimation of effect using the three confidence interval methods **
6. Variance of the error-interval estimation of effect using the three confidence interval methods **

* The mean effect is calculated by the weighted inverse variance method (on page 1.6 in the introduction).

** Confidence intervals are estimated using the three methods detailed in chapter 3. Namely, Wald-type [25], t-distribution [28] and Hartun-Knapp [38] confidence intervals. These methods are only a subset of all the confidence interval methods available, but since they relate only to the secondary performance measures, these three methods are sufficient for this simulation study. All these methods are independent of the choice of heterogeneity variance estimator so any combination of methods can be applied. All combinations are considered in this simulation study.

Bias and error are summarised using both the mean and median in performance measures 1 and 2, all previous simulation studies only used the mean. The median may be more appropriate because errors of heterogeneity variance estimates do not conform to the normal distribution. Sidik and Jonkman [102] noted that mean negative bias causes an artificially low mean squared error so we investigate whether the median has this same issue. Error-interval estimation (performance measures 8 and 9) is a ratio between the width of the estimated confidence interval and the true confidence interval, as defined by a previous simulation study [64]. The range of performance measures is comprehensive and includes measures of bias and variability of estimates of τ^2, the summary effect and confidence intervals of the summary effect. Measures of performance relating to confidence intervals for τ^2 are excluded as it is beyond the scope of the thesis. Details of how to calculate each included performance measure are in appendix C.2.

6.5 Simulation methods

We simulate meta-analysis data by the following main steps:

- A meta-analysis dataset is generated for specified parameter values using the methods outlined in this section.

- Heterogeneity variance estimates are calculated for the given meta-analysis using methods listed in section 6.3.
- Steps 1 and 2 are repeated 5,000 times and performance measures are calculated (see section 6.4)
- Steps 1-3 are repeated for all combinations of parameter values. The parameter values are given in table 6.1 on page 121.

All steps are carried out in R [85]. Bespoke code is used to calculate estimates of the heterogeneity variance and given in appendix D.2. Estimates from this code were compared with estimates produced by the *metafor* package in R for all methods included in this package (CA, DL, HS, SJ, ML and REML). For all other methods, checks were made against estimates from example meta-analyses in published methodology papers [3, 24, 73, 96, 101, 102]. Bespoke code was also written for Wald-type and Hartung-Knapp confidence interval methods for the summary effect and checked against the *metafor* package (see appendix D.3). Heterogeneity variance estimators are compared using the same simulated datasets to eliminate some of the sampling error.

ML and REML are iterative and fail to converge to a solution in a small number of cases [64], but this is primarily due the chosen iteration algorithm rather than the estimator [126]. In this study the default iteration algorithm in *metafor* is used, namely, Fishers scoring method with Cochran's ANOVA the initial estimate [126]. Simulated meta-analyses that cause such failures are not replaced and instances recorded and presented in the results. Heterogeneity variance estimates for each meta-analysis and performance measures for each combination of parameter values are stored for the analysis.

6.5.1 Simulating true study effects

For studies $i = 1, ..., k$ in each meta-analysis, first simulate true study effects (θ_i) such as a log odds ratio or standardised mean difference from some distribution D:

$$\theta_i \sim D\left(\theta, \tau^2\right)$$

where θ is the true summary effect parameter and τ^2 the heterogeneity variance parameter of D.

The standard random-effects model (on page 12 in chapter 1) assumes a normal distribution for D, but θ_i are also simulated from skew-normal distributions with moderate and high skew to test if the methods are robust when this assumption is violated. Distributions for D and parameter values for θ and τ^2 are listed in section 6.6. For each study i, study effect estimates $\hat{\theta}_i$ are then generated to simulate within-study sampling error. The process for doing so depends on the type of outcome of studies in each meta-analysis. In this study, two types of meta-analyses are simulated: (1) continuous outcome meta-analyses with a standardised mean difference effect measure; and (2) dichotomous outcome meta-analyses with an odds ratio effect measure, as detailed in sections 6.5.2 to 6.5.3.

6.5.2 Standardised mean difference (SMD) meta-analyses

To simulate observed standardised mean difference study effects, we use the following steps for each study i:

- Generate sample sizes for each group, denoted by n_{1i} and n_{2i} from one of a number of distributions as detailed in section 6.6.
- Generate n_{1i} observations from $N\left(0, \sigma_1^2\right)$ and n_{2i} observations from $N\left(\theta_i, \sigma_2^2\right)$, to represent participant-level data. Without loss of generality, variances are assumed equal by setting $\sigma_1^2 = \sigma_2^2 = 1$.

- Calculate the sample means ($\overline{Z_{1i}}$ and $\overline{Z_{2i}}$) and standard deviations ($\hat{sd}_{1i}$ and $\hat{sd}_{2i}$) of these observations.
- Calculate the standardised mean difference $\hat{\theta}_i$ and variance $\hat{\sigma}_i^2$ using Hedge's g method (described on page 4 in chapter 1).

I chose to simulate meta-analyses with this outcome measure because study effects are standardised and can be compared between meta-analyses.

6.5.3 Odds ratio meta-analyses

To simulate odds ratio study effects for each study i, I used the following steps:

- Generate the true average probability of an event across the two study groups, denoted by $\overline{p_i}$. p_{1i} and p_{2i} are found from solutions to the simultaneous equations:

$$\overline{p_i} = \frac{p_{2i} + p_{1i}}{2}$$

$$\theta_i = log\left(\frac{p_{2i}\left(1 - p_{1i}\right)}{p_{1i}\left(1 - p_{2i}\right)}\right)$$

  $\overline{p_i}$ are generated from one of a number of distributions as detailed in section 6.6, which represent where events are common and rare.

- Generate sample sizes for each intervention group, denoted by n_{1i} and n_{2i}, from one of a number of distributions in section 6.6.
- The numbers of events in the study groups are generated from the binomial distributions $B\left(n_{1i}, p_{1i}\right)$ and $B\left(n_{2i}, p_{2i}\right)$. Cell counts in a 2x2 contingency table can then be derived.

- Add 0.5 to all cell counts if there is any zero in the table. If there are zero events in both arms then exclude this study from the synthesis. When fewer than 2 studies remain after exclusions, the meta-analysis is withdrawn from the simulations without replacement. This is the current standard method for dealing with zero cell counts, but we recognise there may be other, better performing methods [10, 30].

- Calculate the sample log odds ratio, $\hat{\theta}_i$ and its variance $\hat{\sigma}_i^2$ using formulae on page 6 in chapter 1.

Meta-analyses with this outcome measure wer simulated for two reasons: (1) The systematic review of previous simulation studies (chapter 5) suggested heterogeneity variance estimators perform worse in this setting compared with standardised mean difference meta-analyses and (2) the odds ratio is one of the more common outcome measures in binary outcome meta-analyses [120].

6.6 Parameter values

Performance of the heterogeneity variance estimators are assessed for all combinations of parameter values and distributions given in table 6.1, for standardised mean difference and for odds ratio meta-analyses. There are a total of 960 standardised mean difference meta-analyses and 15,360 simulated scenarios for odds ratio meta-analyses. Parameter values were chosen to represent the range of values observed in published meta-analyses. Further details and a justification for these values are in sections 6.6.1 to 6.6.6 that follow.

6.6.1 Number of studies

Figure 4.1 on page 57 shows the number of studies typically included in CDSR meta-analyses and is used to inform parameter values for k. The number of studies

Parameter		Value/distribution
k	Number of studies in the meta-analysis	2, 3, 5, 10, 20, 30, 50, 100
I^2	Mean I^2 for each scenario	0%, 15%, 30%, 45%, 60%, 75%, 90%, 95%
θ	Summary effect	0.5 for SMD meta-analyses; 0, 0.5, 1.1 and 2.3 for log odds ratio meta-analyses (corresponding to ORs of 1, 1.65, 3 and 10)
θ_i	Distribution of true study effects	*(a)* $\theta_i \sim N(\theta, \tau^2)$(standard random-effects model) *(b)* Normal distribution with moderate skew: $\theta_i \sim SN(\theta, \tau^2, \gamma = 0.7)$ *(c)* Normal distribution with high skew: $\theta_i \sim SN(\theta, \tau^2, \gamma = 0.95)$ τ^2 takes parameter values that satisfy the I^2 values above
n_{1i}, n_{2i}	Study sample sizes	*(a)* Small studies: $n_{1i} = 20$ *(b)* Small to medium sized studies: $n_{1i} \sim U(20, 200)$ *(c)* Medium sized studies: $n_{1i} = 200$ *(d)* Small and large studies: $n_{11}, ..., n_{1m} = 20$ and $n_{1m}, ..., n_{1k} \sim U(1000, 2000)$ where m is the integer half way between 1 and k (when k is odd, one study is be generated from one of the two distributions at random) *(e)* Large studies: $n_{1i} \sim U(1000, 2000)$ In all scenarios, sample sizes are equal between groups ($n_{1i} = n_{2i}$)
Parameters only applying to **odds ratio meta-analyses**		
$\overline{p_i}$	Average probability of event across study groups	*(a)* $\overline{p_i} = 0.5$ *(b)* $\overline{p_i} \sim U(0.1, 0.5)$ *(c)* $\overline{p_i} = 0.05$ *(d)* $\overline{p_i} = 0.01$

Table 6.1: Set of parameter values and distributions to simulate meta-analyses

per meta-analysis in this simulation study range between 2 and 100. Meta-analyses with two studies are excluded from the analysis of CDSR in chapter 57, but these are included in the simulations. For completeness, meta-analyses with up to 100 studies are included, although Cochrane meta-analyses typically containing much fewer studies [21].

6.6.2 Heterogeneity variance parameter values

6.6.2.1 Method to derive heterogeneity variance parameter values

To ensure that heterogeneity variance (τ^2) parameter values represent the full range of inconsistency, τ^2 are defined so that they correspond to true I^2 between 0% and 95% (see table 6.1 for the full range of I^2). It is necessary for τ^2 values to vary between scenarios so that I^2 remain roughly constant. Also, there is difficulty in defining τ^2 using this method because I^2 are likely to vary to some extent due to sampling error in a given scenario. Therefore, we define τ^2 such that it produces meta-analyses with the desired true I^2 *on average* over 5,000 repetitions. We use trial and error to find the τ^2 that satisfy this definition.

The following formula is used to calculate the underlying I^2 of each repetition (similar to that introduced in chapter 1):

$$I^2 = \frac{\tau^2}{\tau^2 + \sigma^2} \cdot 100\%$$

where σ^2 is the 'true' typical study variance:

$$\sigma^2 = \frac{(k-1)\sum_{i=1}^{k} 1/\sigma_i^2}{\left(\sum_{i=1}^{k} 1/\sigma_i^2\right)^2 - \sum_{i=1}^{k} \left(1/\sigma_i^2\right)^2}$$

Preliminary analysis found that I^2 is insensitive to changes in k and the distribution of study effects, so τ^2 values are consistent between these scenarios. τ^2 values differ between all other parameters.

Threshold values are defined for I^2 to help interpret the results: 15% and 30% represent low inconsistency; 45% and 60% represent moderate inconsistency; and 75%, 90% and 95% represent considerable inconsistency. These threshold values roughly correspond to the guidelines in the Cochrane handbook [51], but modified merely so they correspond to simulated I^2 values. Recall that I^2 is a measure of heterogeneity relative to typical within study variances and so loosely represents the degree of overlap between study effect confidence intervals.

6.6.2.2 A summary of heterogeneity variance parameter values

Having applied this method, the derived τ^2 parameter values are given in table D.1 in the appendix. As expected, heterogeneity parameters are smallest in scenarios where within-study variances are also small. That is, in the scenarios containing standardised mean difference meta-analyses with large studies, where a heterogeneity parameter of 0.0256 produces meta-analyses with a mean I^2 of 95%. Conversely, heterogeneity parameters are largest in the scenarios with odds ratio meta-analyses containing small studies, in which a heterogeneity parameter of 15.6 produces a mean I^2 of 95%.

Underlying I^2 vary in any given scenario because they depend on the sampled true study effects and within-study variances. In meta-analyses with small and large study sizes, there are large differences in underlying I^2 within the same scenario; the 5th and 95th percentiles of I^2 have absolute differences of up to 50%. The differences are smaller in all other distributions of study sizes, in which 5th and 95th percentiles of I^2 have differences of up to 20%. Variation in I^2 within a given scenario is not considered a major issue for this analysis, since this method is used only to define

heterogeneity parameter values. The ranges of I^2 values in each scenario is shown in table D.2 in the appendix.

6.6.3 Summary effect

To reduce the number of scenarios, standardised mean difference meta-analyses are only simulated where $\theta = 0.5$. Three previous simulation studies simulated meta-analyses with multiple true SMD effects and all suggested that the value of θ has little bearing on any performance measure [78, 96, 124]. Odds ratio meta-analyses are simulated with a range of summary effects; 0, 0.5, 1.1 and 2.3 (corresponding to odds ratios of 1, 1.65, 3 and 10). Results may be affected by the underlying odds ratio, particularly when the odds ratio is extremely large or small, as shown previously in the simulation study by Bhaumik et al. [3]. An extreme underlying odds ratio causes imbalance in the event probabilities between groups, which can lead to one or both groups having rare events and this could affect results.

6.6.4 Distribution of true study effects

True study effects (θ_i) are generated from three distributions. In all scenarios, θ_i are sampled from distributions with mean θ and variance τ^2. First, θ_i are generated from the normal distribution which is assumed in the standard random-effects model and represents optimal conditions where estimators may perform best (scenario *a*). Some heterogeneity estimation methods such as Paule-Mandel do not assume normality of true effects and therefore it is hypothesised such estimators are more robust under non-normal conditions [24]. Second, θ_i are sampled from two skew-normal distributions (scenarios *b* and *c*) with 0.7 and 0.95 skew parameter values; this represents moderate and high negative skew as illustrated in figure 6.1. Kontopantelis et al. [64] previously looked at performance of heterogeneity variance estimators under skew-normal conditions, and defined this similar level of skew as 'moderate' and 'high'. This distribution is defined elsewhere [68].

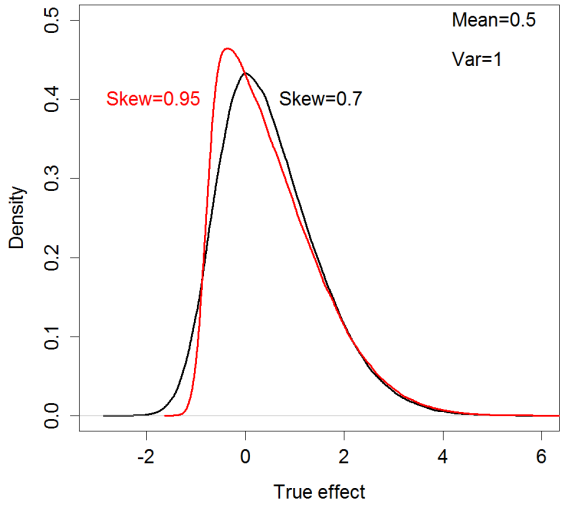


Figure 6.1: Probability density function of skew-normal distribution
Note that the variance differs depending on the simulation scenario, but this is only the scaling parameter.

6.6.5 Study sample sizes

Study sample sizes are generated in five ways to represent small-to-medium study sizes (scenario a), small equally-sized studies (scenario b), medium equally-sized studies (scenario c), small and large studies (scenario d) and large studies only (scenario e). These scenarios are chosen to represent meta-analyses with a range of study sizes and also a range of differences between study size. The systematic review of simulation studies in chapter 5 suggests that performance of heterogeneity variance estimators may be dependent on study sample sizes, but evidence is currently limited [55].

6.6.6 Average probability of event across study groups

Binary outcome meta-analyses are generated from a range of underlying event rates. In scenario a, the underlying average event rate is 0.5 to represent the ideal scenario with event rates sampled as far from the asymmetric tails of the binomial distribution as possible. Scenario b represents a situation where event rates are variable between studies but not so rare as to be considered a big contributing factor to poor estimates

of summary effects and standard errors. Scenarios c and d represent situations where the average underlying event rate is homogeneous and rare. It is not necessary to simulate meta-analyses where the event of interest is extremely common (e.g. 0.95) as the resulting odds ratios are the inverse of those obtained with extremely uncommon event rates.

6.7 An overview of the simulation study

The systematic review of previous simulation studies in the last chapter found conflicting recommendations. Four reasons for this were suggested, which we address in this new study. (1) There was conflict of interest in most studies because they compared existing methods with those newly proposed. To address this, we only compare pre-existing methods in our study. (2) Most studies only compared a small subset of the methods available, so we include a comprehensive list. (3) Simulations were often not representative of real meta-analyses, so we define parameter values for simulations based on meta-analyses seen in practice. (4) Studies don not address that all methods are very imprecise in typical meta-analyses. They failed to address this issue because their results were focused on the relative performance of methods. We consider both relative and absolute performance in this simulation study.

Meta-analyses are simulated with odds ratio and standardised mean difference study effects to capture properties of heterogeneity variance estimators for a representative range of outcome measures. Novianti et al. [78] was the only study identified in my systematic review that simulated both binary and continuous outcomes. participant-level data is simulated to ensure simulated data is representative of real meta-analyses. Generating participant-level data will also ensure the issues with heterogeneity estimation specific to certain types of outcome measures is captured. One issue is that estimated odds ratio and standardised mean difference study effects are correlated with their variances [3, 8]. This is a particularly large issue in all binary outcome meta-analyses with rare events [3].

Our methods for simulating meta-analysis data differ from most other previous simulation studies in two key ways. First, we define underlying τ^2 parameter values that correspond to a consistent range of underlying I^2 values. We define a range of I^2 between 0% and 95% to ensure the corresponding range of τ^2 represents zero, low, moderate and high inconsistency in study effects for all scenarios. Only Kontopantelis et al. [64] has previously taken a similar approach. No guidelines exist for interpreting τ^2 estimates because the measure cannot be compared between meta-analyses, but the Cochrane Collaboration have issued rough guidelines on interpreting I^2 values [51]. Second, all previous studies defined the event probability of the control group for simulating binary outcome meta-analyses. Conversely, we define the average event probability between both study groups. In doing so, the *rarity* of the event is more independent of the study effect sizes.

Results are presented from these simulated meta-analyses in the following two chapters. In the next chapter, we explore comprehensively the performance of all included heterogeneity variance estimators. Scenarios are identified where all estimators perform poorly, when they perform well and in such cases which estimators perform better than others. I then investigate how the findings from this analysis apply to real meta-analyses in chapter 8 by combining with empirical data. Methods for analysis of this simulated meta-analyses data are detailed in the two chapters that follow.

Chapter 7

Main simulation study results

7.1 Introduction

The last chapter detailed the design of a new simulation study to compare heterogeneity variance estimators in random-effects meta-analysis. Details included the methods for simulating meta-analysis data, which heterogeneity variance estimators are compared and the performance measures used for comparisons. The study is designed based on findings from a systematic review of previous simulation studies in chapter 5 and input from other collaborators. In this chapter, the results of this study are presented.

A number of heterogeneity variance estimators are excluded from the main results because they are clearly inferior to other estimators; section 7.2 explains the reasons for these exclusions. Also, given the scale of this study, it was only possible to present a subset of all simulated scenarios and performance measures. Reasons for choosing this subset are given in sections 7.3 and 7.4. These exclusions of estimators and results were based on a preliminary exploration of all study results, which are presented more fully in volume II of this thesis.

The main results are given in section 7.5 and split into three parts. First, results that compare estimators in terms of performance measures relating to point estimates of the heterogeneity parameter are presented in section 7.5.1. Mean bias and mean squared error performance measures in this section are plotted on the proportional scale to the heterogeneity variance parameter whenever $\tau^2 > 0$. In other words, mean bias is plotted as a proportion of the true parameter value rather than absolute difference from the truth. Similarly, for a proportional mean squared error of (for example) 100%, the average squared error is equal to τ^2. This is so that results can be compared more easily between scenarios of different τ^2 and to help interpretation. Raw mean bias and mean squared error is presented whenever $\tau^2 = 0$. After results from the primary performance measures, those relating to estimation of the summary effect are presented in section 7.5.3 and finally, those relating to the confidence interval for the summary effect are in section 7.5.4. Within each section,

selected results are presented to give a representative picture of all simulated scenarios and a summary explains how they can be generalised to all scenarios. Results are interpreted from two viewpoints; (1) as a relative comparison of the performance of heterogeneity variance estimators reveal those that perform best and (2) as a general comparison of performance between scenarios to summarise where all estimators perform well/poorly.

7.2 Heterogeneity variance estimators excluded from the main analysis

We excluded Rukhin's estimator with zero prior (B0), Rukhin's simple estimator (BP) [93] and that proposed by Malzahn, Böhning and Holling (MBH) [74] in a preliminary analysis. These estimators are not compared in the main results because they clearly have inferior properties and would distract the reader away from those with more reasonable properties.

To justify these exclusions, figure 7.1 presents two selected graphs from the simulation results including all heterogeneity variance estimators. The y-axes present proportional mean bias (left) and proportional mean squared error (right). These results are from scenarios of standardised mean difference meta-analyses with small-to-medium study sizes and heterogeneity variance of 0.0299 (which represents a mean I^2 of 60%). The figure shows B0 has considerable negative bias and BP has considerable positive bias when there are more than 5 studies. MBH has a higher mean squared error than all other estimators included in the main results, particularly when the number of studies is low.

The 12 remaining heterogeneity variance estimators are compared extensively in the main results in section 7.5 using graphs similar to those in figure 7.1.

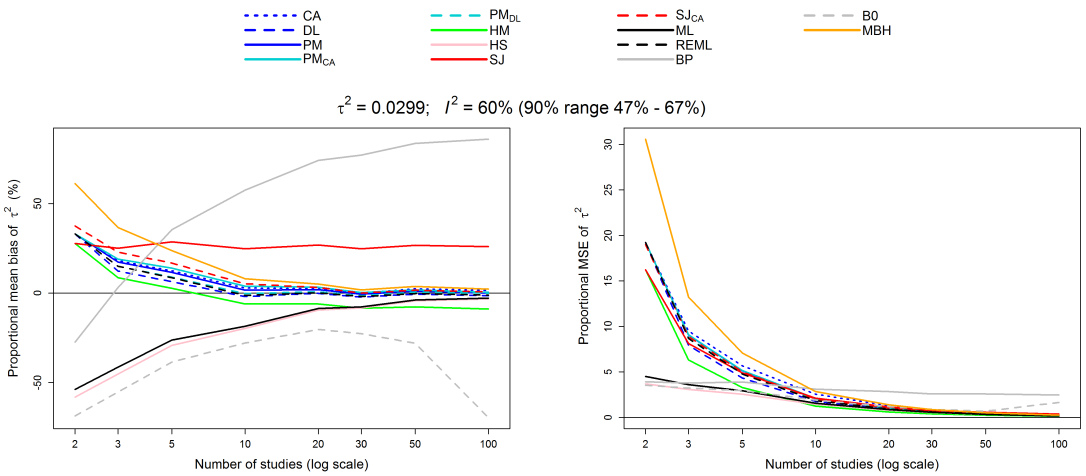


Figure 7.1: Proportional mean bias (left-hand-side) and proportional mean squared error (right-hand-side) in selected scenarios with B0, BP and MBH heterogeneity variance estimators included.
Scenarios containing standardised mean difference meta-analyses ($\theta = 0.5$) with small-to-medium study sizes and a mean I^2 of 60%.

7.3 Simulated scenarios not presented in full

Standardised mean difference and odds ratio meta-analyses data were generated for all combinations of parameter values as detailed in the last chapter. Results were produced from a total of 16,320 meta-analysis scenarios, of which only a representative subset could be presented in full in this chapter. The scenarios chosen are different for each performance measure; reasons for selection are detailed in each section of the results. Only scenarios with normally distributed study effects are presented in this chapter because results from skew-normal distributions were not meaningfully different indicating that heterogeneity variance estimators and confidence interval methods are robust to skew-normal effects. Scenarios with an effect size of 0.5 (standardised mean difference/log-odds ratio) are presented in most of this chapter given that results for the most part were representative of other effect sizes. In section 7.5.2, results from other effect sizes are presented to show where differences were observed.

7.4 Selected performance measures

Only a subset of the performance measures originally considered (as listed in the last chapter) are included in the main results. Preliminary analysis showed some failed to capture anything of interest and some showed comparable results. Therefore, five performance measures of those originally specified are included in this chapter. The mean bias, mean squared error and proportion of zero estimates are included that all relate directly to point estimates of the heterogeneity variance. Also included are bias of the summary effect estimate and coverage of 95% confidence intervals of the summary effect. Confidence intervals are calculated by Wald-type [25], t-distribution [28] and Hartung-Kanpp [38] methods.

Performance measures not reported are listed in table F.1 of the appendix along with the reasons for exclusion. Selected results to show why these performance measures are excluded are given in figures F.1 to F.6 in the same appendix.

7.5 Results

7.5.1 Properties of heterogeneity variance parameter estimates

In this section of the results, heterogeneity variance estimators are compared in terms of performance measures that relate directly to estimation of the heterogeneity parameter (τ^2). These performance measures are mean bias, mean squared error and the proportion of zero heterogeneity variance estimates.

A subset of results are presented from standardised mean difference and odds ratio meta-analyses containing small studies, small-to-medium sized studies and small and large studies. This subset of results were chosen as they represent both a range of study sizes and also a range of differences in study size. For odds ratio meta-analyses,

we present those containing studies with event probability simulated from a uniform distribution between 0.1 to 0.5. These scenarios represent meta-analyses where rare events impact the results but not so considerably that all estimators perform poorly. Scenarios of meta-analyses with mean I^2 values of 0%, 30%, 60% and 90%, which represent the full range of inconsistency between studies.

Figures 7.2 - 7.7 (pages 136 - 141) show the mean bias, mean squared error and proportion of zero heterogeneity variance estimates in standardised mean difference and odds ratio meta-analyses separately. The properties of each estimator are described in separate the sections that follow, based on these figures.

7.5.1.1 DerSimonian-Laird (DL)

In scenarios of standardised mean difference meta-analyses, DL is negatively biased when study effects have high I^2 and study sample sizes are small (as shown in figure 7.2, plot A4). This negative bias increases as the number of studies increases and reaches up to -20%. These scenarios have a τ^2 parameter of 0.991, so in absolute terms the mean bias is up to -0.19. In all the other standardised mean difference scenarios in this figure, DL is positively biased in meta-analyses containing fewer than 10-20 studies and roughly unbiased for those with more studies. DL has similar bias to many estimators including $\mathrm{PM}_{\mathrm{CA}}$, $\mathrm{PM}_{\mathrm{DL}}$ and REML in scenarios with small studies and small-to-medium studies. In meta-analyses with small and large studies (plots C1-C4), DL is one of the least biased estimators and distinctly lower than PM and $\mathrm{PM}_{\mathrm{CA}}$.

Mean bias in scenarios of odds ratio meta-analyses with event probabilities between 0.1 and 0.5 is shown in figure 7.3. In these scenarios, DL's negative bias is observed to a greater extent than in standardised mean difference scenarios and includes those with small-to-medium sized studies and high I^2. Results suggest that in odds ratio meta-analyses, larger sample sizes are required than in standardised mean difference meta-analyses to avoid DL's negatively biased estimates. Alternatively, study

event probabilities closer to the ideal 0.5. As with standardised mean difference meta-analyses, DL is one of the least biased estimators in odds ratio meta-analyses containing small and large studies (plots C1-C4).

DL is compared in terms of mean squared error in figures 7.4 and 7.5. DL has a relatively low mean squared error in the same scenarios as when the estimator is negatively biased. However, this is a consequence of how mean squared error is measured and so this isn't necessarily a good property. Also, DL also has relatively low mean squared error in scenarios containing small-to medium and small and large studies. In scenarios with small equally-sized studies, DL has mean squared error comparable with many other estimators including CA, PM, PM_{CA}, PM_{DL}, SJ_{CA} and REML.

DL consistently has one of the lowest proportions of zero heterogeneity variance estimates of all estimators that require truncation. However, this is a similar proportion as other weighted method of moments estimators including PM, PM_{CA} and PM_{DL}.

7.5.1.2 Cochran's ANOVA (CA)

CA tends to produce higher estimates of the heterogeneity variance than most other estimators in scenarios of both standardised mean difference and odds ratio meta-analyses. As such, CA is roughly unbiased in scenarios with typically high I^2 when most other estimators are negatively biased. However, CA is one of the most positively biased estimators for up to moderate I^2. CA's positive bias is particularly prominent in scenarios with small and large studies (figures 7.2 and 7.3, plots C1-C4). It is to be expected that CA performs poorly when there are large differences in study size, given that the estimator assigns equal study weights. CA's positive bias is slightly greater in odds ratio meta-analyses with event probability 0.1 to 0.5 than in standardised mean difference meta-analyses.

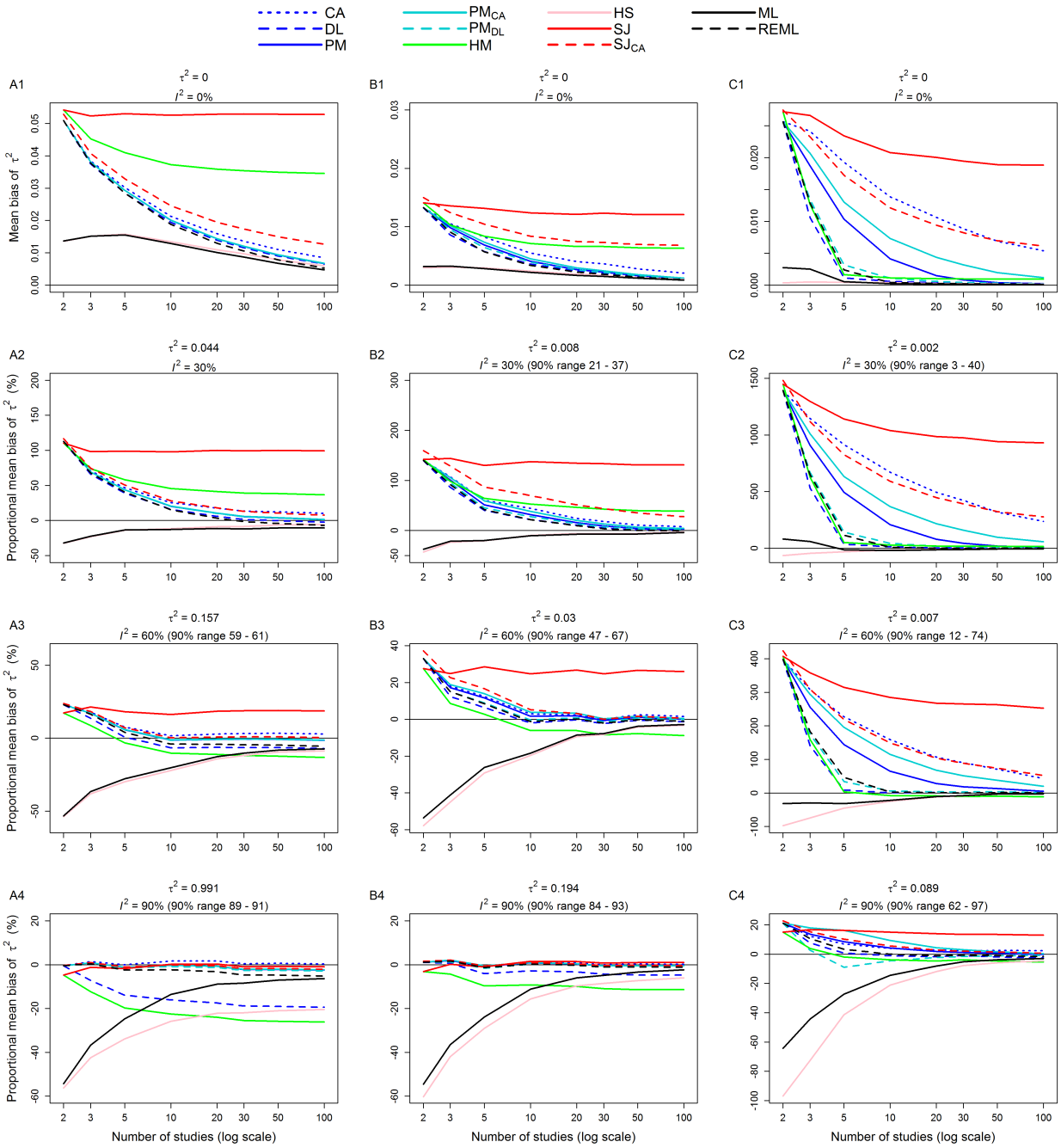


Figure 7.2: Mean bias of heterogeneity variance estimates in standardised mean difference outcome meta-analyses

Scenarios containing small studies (A1-A4), small-to-medium studies (B1-B4) and small and large studies (C1-C4). Effect size $\theta = 0.5$.

Bias is presented on the proportional scale when $\tau^2 > 0$.

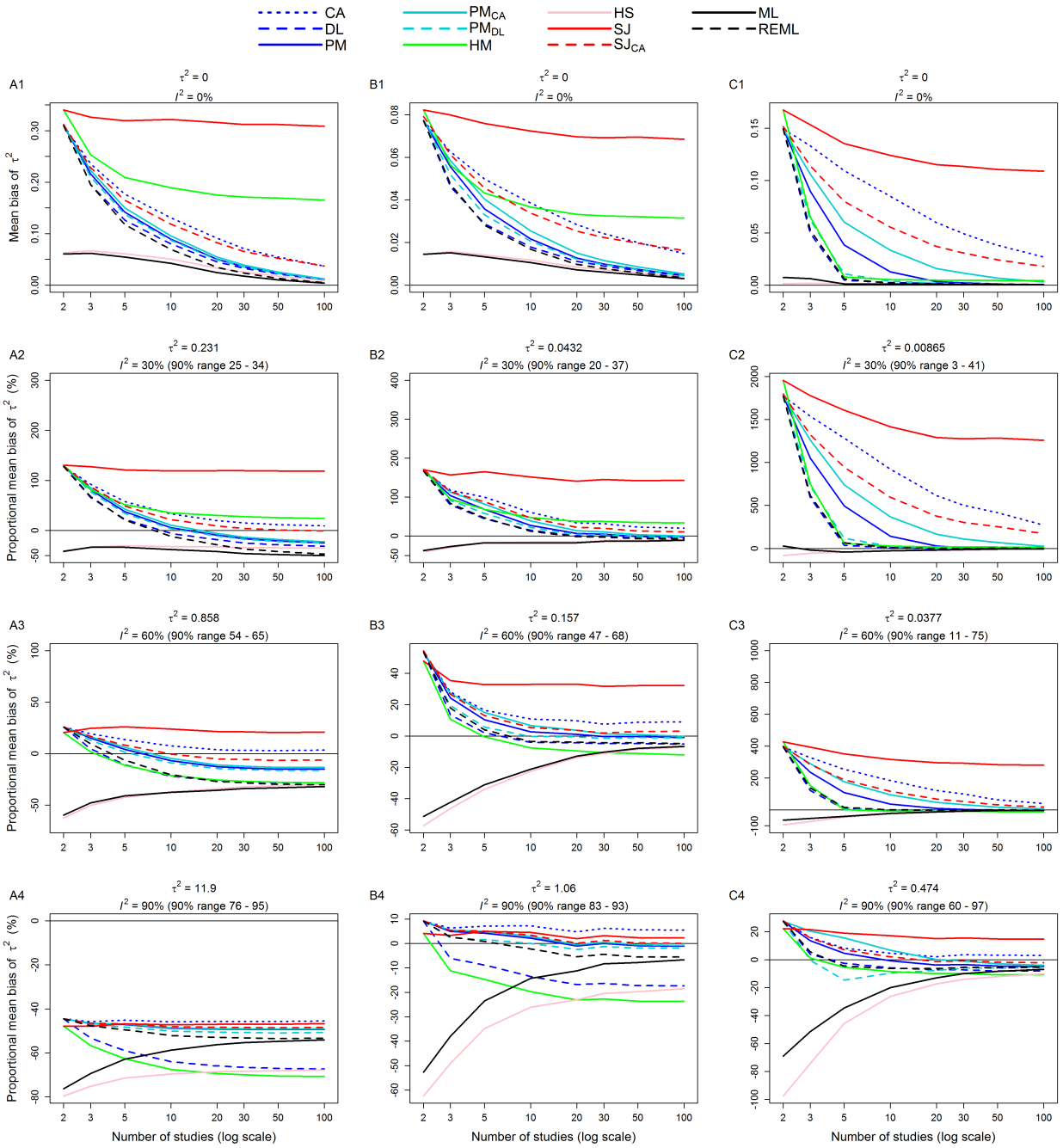


Figure 7.3: Mean bias of heterogeneity variance estimates in odds ratio meta-analyses with event probability 0.1 to 0.5

Scenarios containing small studies (A1-A4), small-to-medium studies (B1-B4) and small and large studies (C1-C4). Effect size $\theta = 0.5$.

Bias is presented on the proportional scale only when $\tau^2 > 0$.

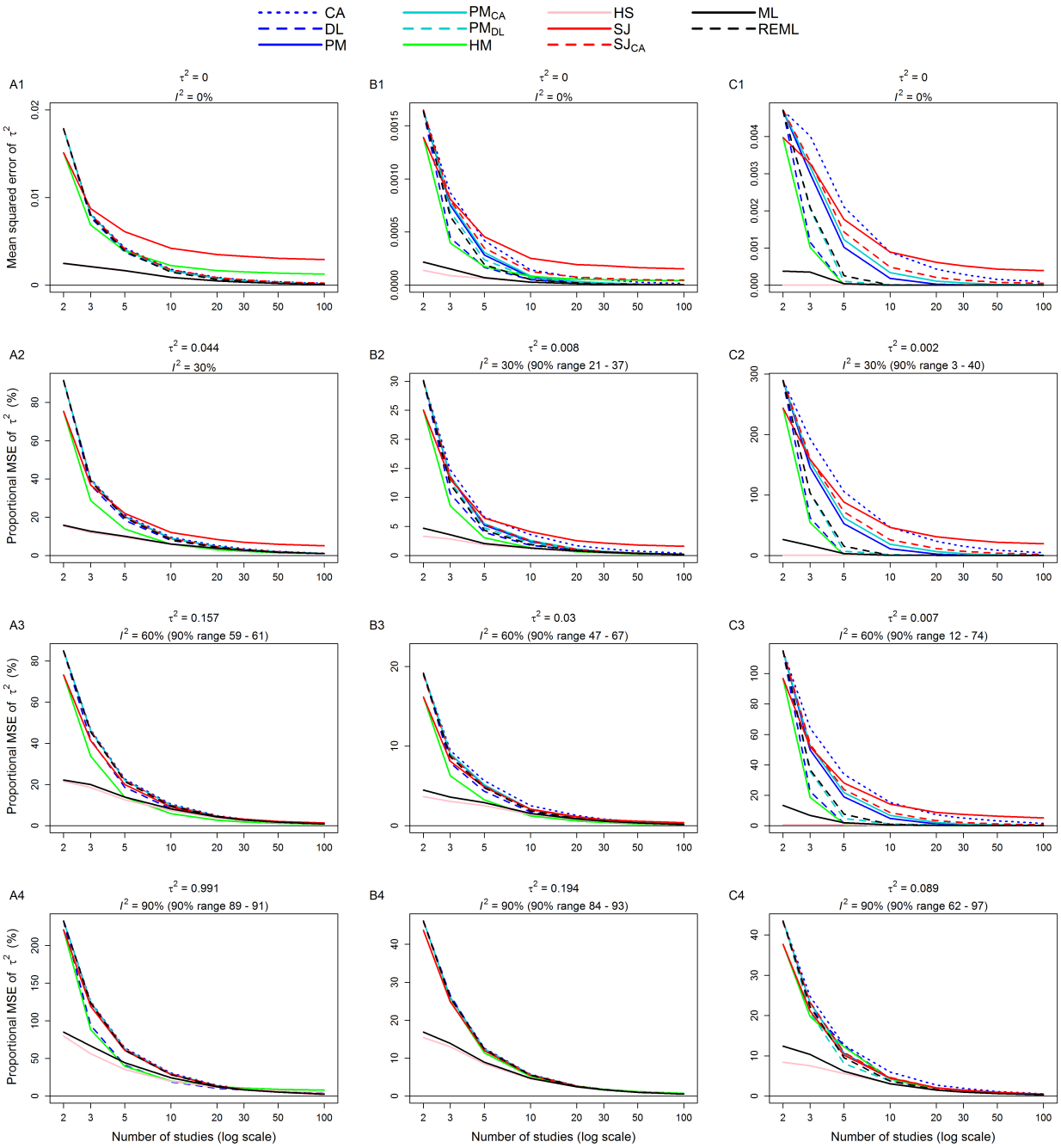


Figure 7.4: Mean squared error of heterogeneity variance estimates in standardised mean difference meta-analyses

Scenarios containing small studies (A1-A4), small-to-medium studies (B1-B4) and small and large studies (C1-C4). Effect size $\theta = 0.5$.

Mean squared error is presented on the proportional scale only when $\tau^2 > 0$.

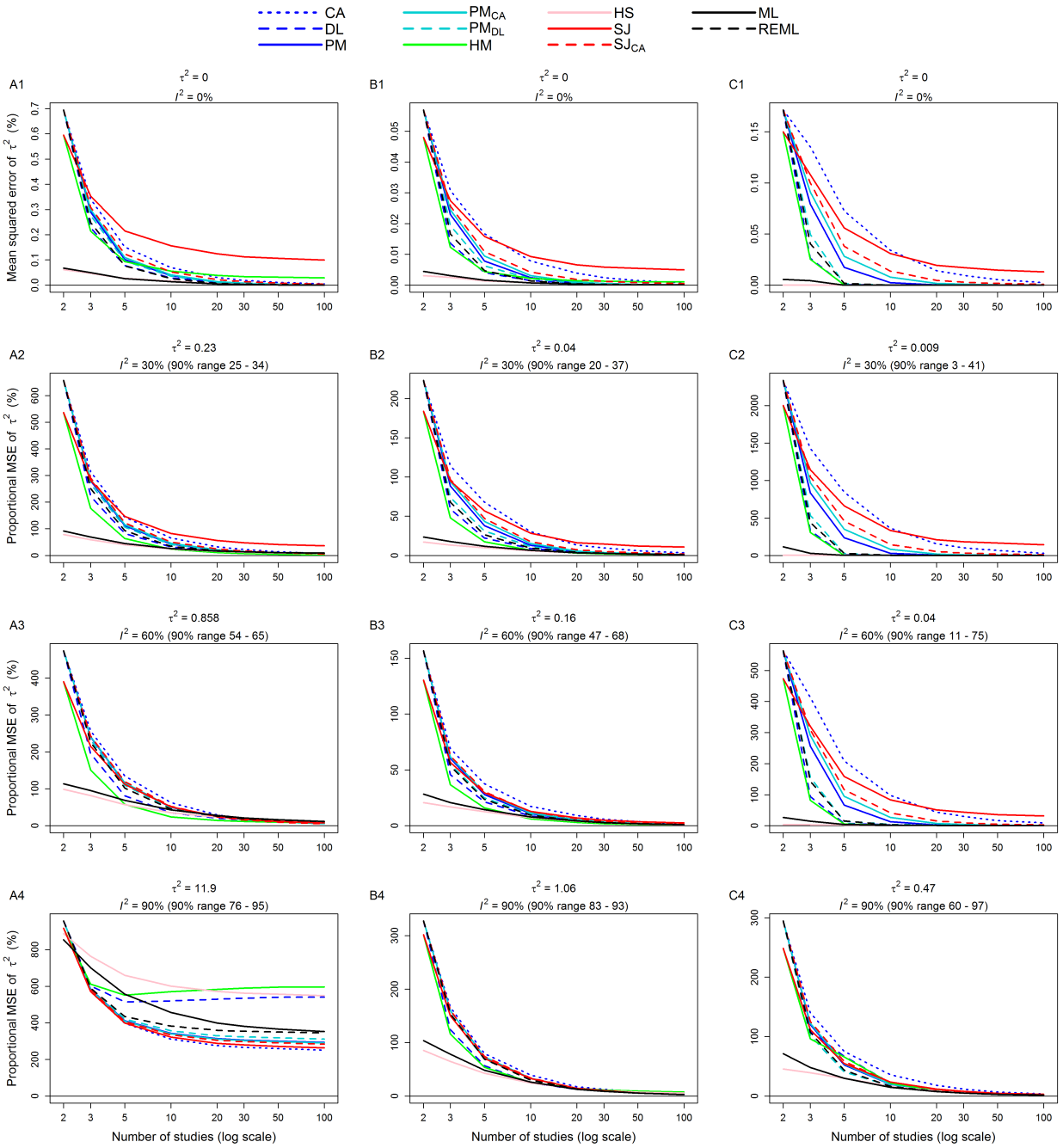


Figure 7.5: Mean squared error of heterogeneity variance estimates in odds ratio meta-analyses with event probability 0.1 to 0.5

Scenarios containing small studies (A1-A4), small-to-medium studies (B1-B4) and small and large studies (C1-C4). Effect size $\theta = 0.5$.

Mean squared error is presented on the proportional scale only when $\tau^2 > 0$.

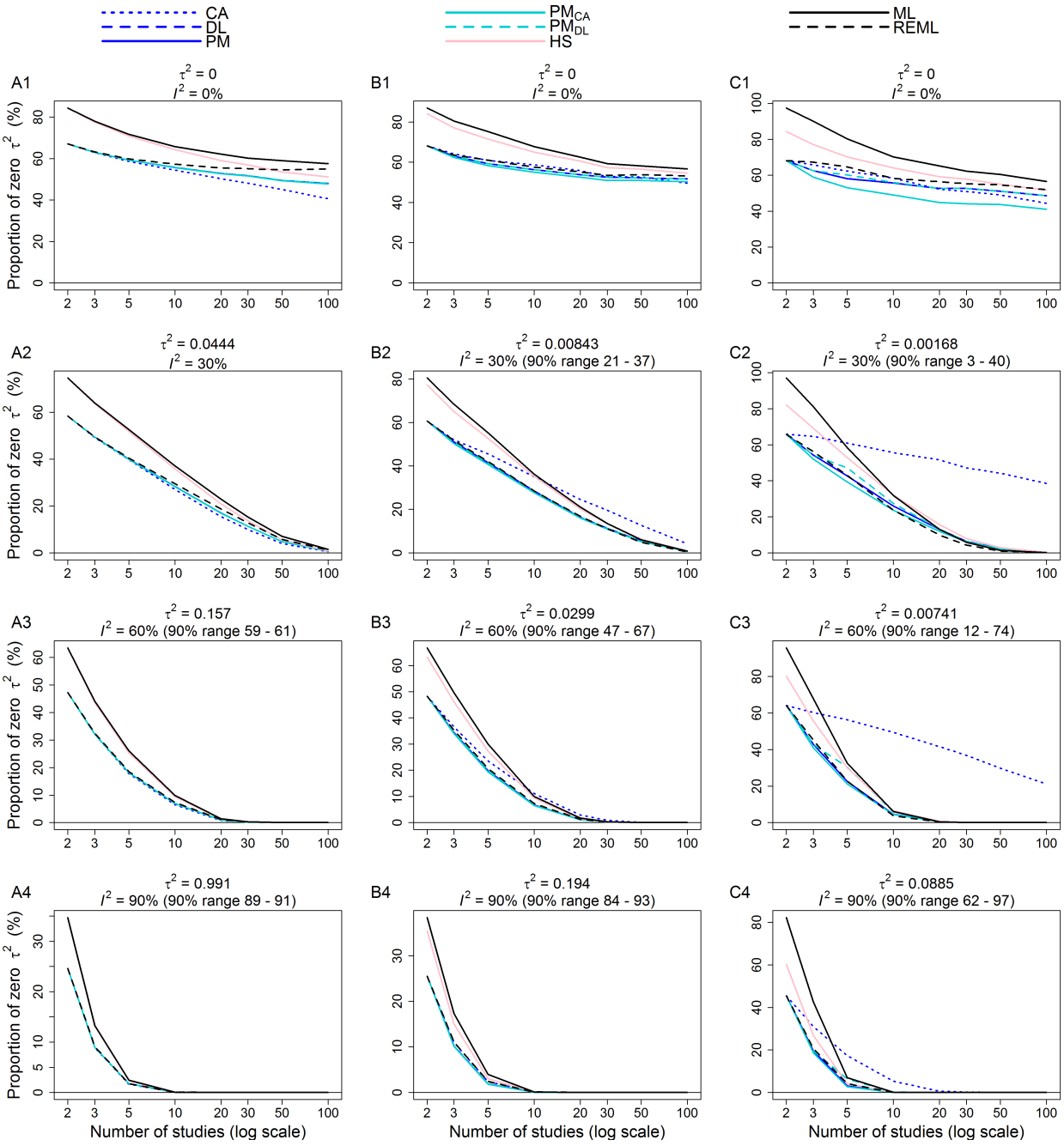


Figure 7.6: Proportion of zero heterogeneity variance estimates in standardised mean difference meta-analyses

Scenarios containing small studies (A1-A4), small-to-medium studies (B1-B4) and small and large studies (C1-C4). Effect size $\theta = 0.5$.

HM, SJ, SJ_{CA} are not included as they only produce positive τ^2 estimates.

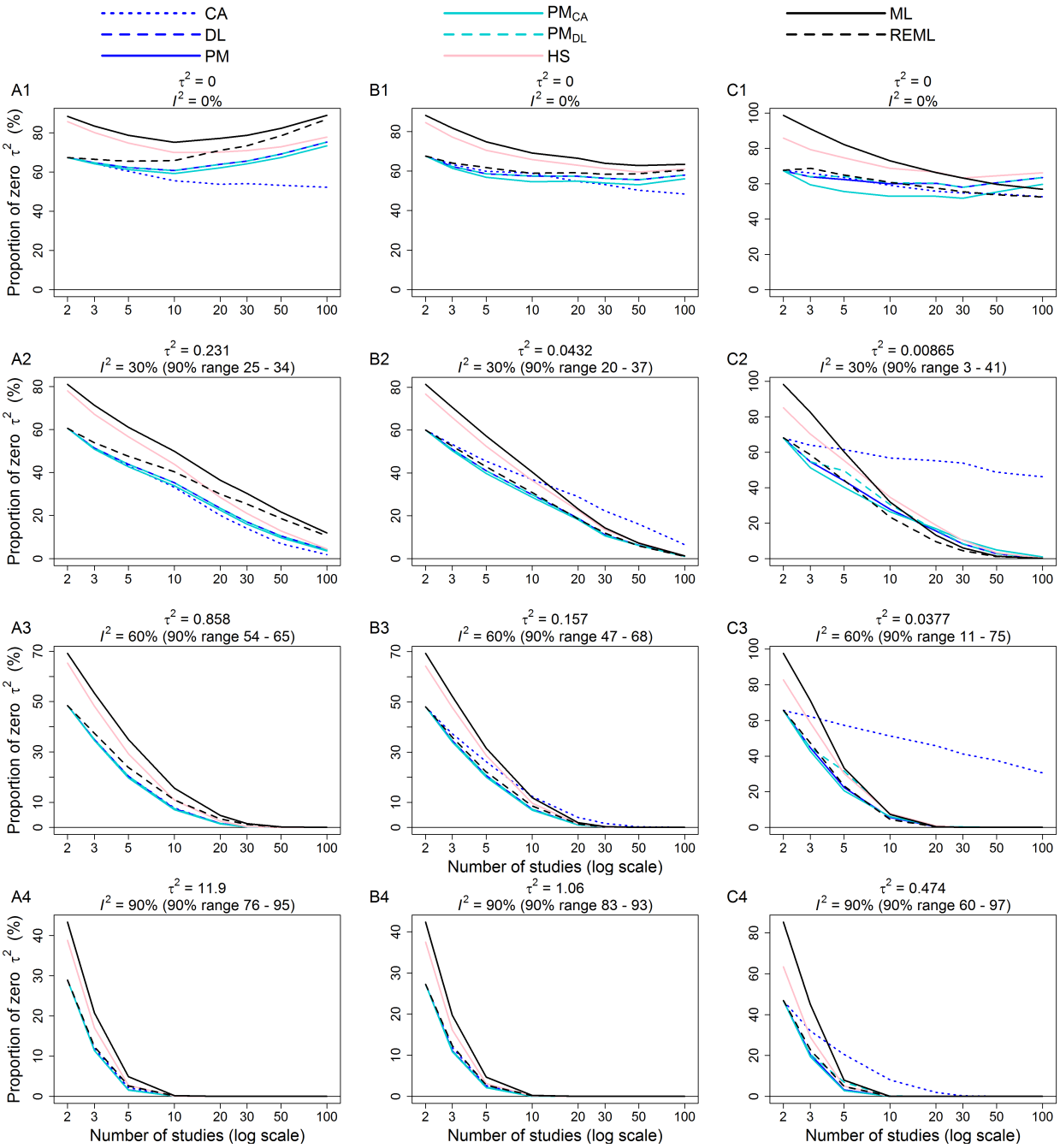


Figure 7.7: Proportion of zero heterogeneity variance estimates in odds ratio meta-analyses with event probability 0.1 to 0.5

Scenarios containing small studies (A1-A4), small-to-medium studies (B1-B4) and small and large studies (C1-C4). Effect size $\theta = 0.5$.

HM, SJ and SJ_{CA} are not included as they only produce positive τ^2 estimates.

CA also has a higher mean squared error than most other estimators when there are large differences in study sizes (7.4, plots C1-C4). Its mean squared error is comparable with most other estimators when study sizes are the same (7.4, plots A1-A4). These findings apply to both standardised mean difference and odds ratio meta-analyses (figure 7.5).

CA produces a comparatively high proportion of zero estimates in meta-analyses containing small and large studies. Only HS and ML have a higher proportion and only in meta-analyses with up to 5 studies.

7.5.1.3 Paule-Mandel (PM)

PM has a mean bias similar to PM_{CA}, PM_{DL} and REML in scenarios of standardised mean difference meta-analyses that contain small or small-to-medium sized studies (figure 7.2, plots A1-A4 and B1-B4). In these scenarios, PM is roughly unbiased when I^2 is high or the meta-analysis has more than 20 studies and positively biased otherwise. In scenarios with small and large studies (figure 7.2, plots C1-C4), PM's mean bias is distinct from all other estimators and has a higher positive bias than ML, HS, DL, PM_{DL}, HM and REML. PM's relatively high positive bias in these scenarios is particularly prominent where meta-analyses contain five or fewer studies.

PM has a similar mean bias relative to other estimators in odds ratio meta-analyses with study event probabilities between 0.1 to 0.5 (figure 7.3). The key differences are that PM has small levels of negative bias in scenarios containing small studies. For example, in plot A3 for moderate I^2 and $\tau^2 = 0.858$, PM has bias of up to -20% compared with DL's bias of up to -40%.

PM's mean squared error is comparable with PM_{CA}, PM_{DL} and REML in both standardised mean difference and odds ratio meta-analyses with small equally-sized studies (figures 7.4 and 7.5). In scenarios with small-to-medium and small and large studies, PM has a higher mean squared error than more than half of all estimators compared, namely ML, HS, DL, PM_{DL}, HM and REML.

PM produces an almost identical proportion of zero estimates as DL, PM_{CA}, PM_{DL} and REML (figures 7.6 and 7.7). When I^2 is low PM produces zero estimates of heterogeneity for meta-analyses containing up to 100 studies when I^2 is low. For meta-analyses with high I^2, PM produces no zero estimates when there are 10 or more studies.

7.5.1.4 Two-step Cochran's ANOVA (PM_{CA})

PM_{CA} is a two-step version of PM that uses CA as an initial estimate of heterogeneity. As such, PM_{CA}'s mean bias and mean squared error are equal to, or somewhere between, CA and PM in all scenarios.

PM_{CA} has similar bias to CA and PM (and also REML) in scenarios of standardised mean difference and odds ratio meta-analyses that contain small or small-to-medium sized studies (figure 7.2 and 7.3, plots A1-A4 and B1-B4). For standardised mean difference and odds ratio meta-analyses with small and large studies (plots C1-C4), PM_{CA}has a more distinct mean bias, slightly lower than CA and much higher than PM. Only CA, SJ and SJ_{CA} have more positive bias in these scenarios. PM_{CA} also has a distinct mean squared error in scenarios with small and large studies; slightly higher than PM and considerably less than CA.

PM_{CA} produces an almost identical proportion of zero estimates as DL, PM_{DL} and REML.

7.5.1.5 Two-step DerSimonian-Laird (PM_{DL})

In most scenarios, properties of PM_{DL} are similar to both DL and PM. Results differ in standardised mean difference and odds ratio meta-analyses with small studies and high I^2 in which PM_{DL} and PM are roughly unbiased and DL is negatively biased. Also, $PM_{DL}$and DL have relatively low positive bias and low mean squared error in scenarios containing small and large studies, where PM's positive bias and mean

squared error are higher. Thus, results overall show PM_{DL} has the best properties of DL and PM.

PM_{DL} produces an almost identical proportion of zero estimates as DL, PM_{CA} and REML.

7.5.1.6 Maximum likelihood (ML) and Hunter-Schmidt (HS)

ML has similar properties to HS in terms of all performance measures. ML and HS produce the lowest estimates of all the estimators compared, particularly when there are five or fewer studies in the meta-analysis. As such they are the least positively biased estimators for zero to low I^2 values and have the highest negative bias for moderate and high I^2. For example, in standardised mean difference meta-analyses with small-to-medium study sizes and a τ^2 range of 0 to 0.194 (figure 7.2, plots B1-B4), ML and HS have minimal positive bias for zero and low I^2 and a mean bias as low as -60% for moderate to high I^2. ML and HS have the lowest mean squared errors in all meta-analyses as a consequence of their comparatively low heterogeneity variance estimates (figures 7.4 and 7.5) and the highest proportion of zero heterogeneity variance estimates (figures 7.6 and 7.7). These findings apply to both standardised mean difference meta-analyses and odds ratio meta-analyses with study event probabilities 0.1 to 0.5.

7.5.1.7 REML

REML has similar properties to DL in most scenarios. In a small number of scenarios where DL is negatively biased, REML is also negatively biased but often to a much lesser extent. Recall, these scenarios include standardised mean difference meta-analyses with small studies and high I^2 (figure 7.2, plot A4) and to a greater extent in odds ratio meta-analyses containing up to medium-sized studies and from moderate I^2 (figure 7.3, plots A3, A4 and B4). REML has relatively low bias and low mean squared error, as does DL, in scenarios containing small and large studies.

REML has similar properties to DL and PM_{DL} in most scenarios. The main difference is in odds ratio meta-analyses where estimates are negatively biased, DL often has the highest negative bias, followed by REML and PM_{DL} has the least; this can be observed most prominently in figure 7.3 (plot A3). Differences in bias between REML and DL are also observed in standardised mean difference meta-analyses with small studies and high I^2 (figure 7.2, plot A4). REML has relatively low bias and low mean squared error, as does DL and PM_{DL}, in scenarios containing small and large studies (figures 7.2 and 7.3, plots C1-C4).

7.5.1.8 Hartung-Makambi (HM)

Recall that HM is a transformation of the DL estimator that only produces positive estimates of the heterogeneity. In meta-analyses with small or small-to-medium study sizes and zero or low I^2, HM tends to produce relatively high estimates of heterogeneity and therefore has relatively high positive bias. HM tends to produce comparatively low estimates when I^2 is moderate or high and has more negative bias DL in these scenarios. For example, in scenarios of standardised mean difference meta-analyses with small studies, high I^2 and a τ^2 parameter value of 0.991 (figure 7.2, plot A4), HM's negative mean bias is up to -25% and DL's negative mean bias reaches -20%. In contrast, HM is one of the least biased estimators in meta-analyses containing small and large studies, with similar bias as DL.

HM has a lower mean squared error than all estimators except HS and ML estimators, but these estimators have much more considerable negative bias. Surprisingly, HM has low mean squared error in scenarios with meta-analyses that have zero to low I^2, where HM has a relatively high positive bias. HM has a particularly low mean squared error, similar to DL, in meta-analyses with small and large studies because HM in these scenarios has relatively low bias.

7.5.1.9 Sidik-Jonkman (SJ)

SJ typically produces one of the highest estimates of the heterogeneity variance in both standardised mean difference and odds ratio meta-analyses. As such, SJ has considerable positive bias for meta-analyses with up to typically moderate I^2. For example, in standardised mean difference meta-analyses containing small-to-medium sized studies and low I^2 (figure 7.2, plot B2), SJ has mean bias of more than 100% when almost all other estimators are roughly unbiased. It is to be expected that SJ has positive bias for low I^2, given that it only produces positive heterogeneity variance estimates. However, SJ's positive bias is much higher than other positive estimators including SJ_{CA} and HM. In meta-analyses with high I^2 values, SJ has a relatively low bias similar to CA, SJ_{CA}, PM_{CA}, PM_{DL} and REML. SJ's bias remains constant as the number of studies in meta-analyses increase, while the bias of most other estimators converge to zero.

SJ also has a relatively high mean squared error in meta-analyses with up to moderate I^2 values and a mean squared error similar to most other estimators when I^2 is high.

7.5.1.10 Sidik-Jonkman (CA initial estimate) (SJ_{CA})

Recall that SJ_{CA} is a two-step heterogeneity variance estimator based on the same approach as SJ and as such only produces positive estimates. In standardised mean difference and odds ratio meta-analyses with up to moderate I^2, SJ_{CA} becomes more positively biased as typical study sizes increase. In meta-analyses with small studies (as shown in figures 7.2 and 7.3, plots A1-A4), SJ_{CA} is one of the least biased estimators, with bias similar to many of the truncated methods including DL, PM and REML. In meta-analyses with medium-sized studies, its bias is comparable with SJ and for meta-analyses with large studies SJ_{CA} has the highest positive bias of all estimators compared (the results of these scenarios are shown in the results in volume II of this thesis). SJ_{CA} is roughly unbiased in meta-analyses with high I^2 similar to CA, SJ, PM_{CA}, PM_{DL} and REML.

In scenarios where SJ_{CA} has positive bias, it also have relatively high mean squared error (i.e. in meta-analyses with large studies).

7.5.1.11 A summary of all simulated scenarios

Table 7.1 summarises the simulation results across all scenarios of standardised mean difference and odds ratio meta-analyses. The table is colour-coded to show scenarios where the properties of all estimators are similar in terms of all three performance measures reported thus far; (1) proportional mean bias, (2) proportional mean squared error and (3) proportion of zero heterogeneity variance estimates.

All estimators have substantial negative mean bias in odds ratio meta-analyses with an event probability of up to 0.05, except when all studies are large (i.e. those with sample sizes of 2000 per study group). All estimators also have considerable negative bias in odds ratio meta-analyses with common events and small studies (i.e. those with sample sizes of 20 per study group). In all other scenarios when there is a sufficient number of studies, many of the estimators have reasonable properties.

We derived two other key observations from table 7.1. First, heterogeneity variance estimators generally have worse properties in scenarios containing small study sizes and in odds ratio meta-analyses with low event probabilities. Second, the properties of heterogeneity variance estimators are similar between standardised mean difference meta-analyses and the equivalent odds ratio meta-analyses when events are common. The exception is in meta-analyses containing small studies; all heterogeneity variance estimators are considerably biased in odds ratio meta-analyses with high I^2 where many are unbiased in the equivalent standardised mean difference scenario (as shown in figures 7.2 and 7.3, plot A4).

<table>
<tr><td colspan="2"></td><td colspan="5">Study sizes</td></tr>
<tr><td colspan="2"></td><td>Small</td><td>Small-to-medium</td><td>Medium</td><td>Small and large</td><td>Large</td></tr>
<tr><td colspan="2">SMD meta-analyses</td><td>*</td><td>*</td><td rowspan="3"></td><td>*</td><td rowspan="3">All estimators have similar relative performance as those with equally-sized small studies and medium studies. Bias, mean squared error and proportion of zero estimates are proportionally reduced.</td></tr>
<tr><td rowspan="3">OR meta-analyses with average event probability:</td><td>0.5</td><td rowspan="2">All estimators have substantial negative bias for high I^2.
</td><td rowspan="2"></td><td rowspan="2">**</td></tr>
<tr><td>0.1 to 0.5</td></tr>
<tr><td>0.05</td><td colspan="2" rowspan="2"></td><td colspan="2">All estimators have substantial negative bias for moderate to high I^2 .</td><td></td></tr>
<tr><td>0.01</td><td colspan="3">All estimators have substantial negative bias for all I^2 and rarely estimate τ^2 above 0.</td><td></td></tr>
</table>

Table 7.1: A summary of the properties of heterogeneity variance estimators for all scenarios of standardised mean difference and odds ratio meta-analyses with effect size 0.5.

are scenarios where all estimators have considerable negative bias. Some estimators are reasonably unbiased in all other scenarios (if there are enough studies).

** Mean bias presented in figure 7.2, mean squared error in figure 7.4 and proportion of zero estimates in 7.6*

*** Mean bias presented in figure 7.3, mean squared error in figure 7.5 and proportion of zero estimates in 7.7*

7.5.2 Properties of heterogeneity variance estimates for varying effect sizes

The results presented thus far come from simulated scenarios with a log odds ratio and standardised mean difference of $\theta = 0.5$ (this corresponds to an odds ratio of 1.65). Meta-analyses were also simulated with other log odds ratio effect sizes; $\theta = 0$, 1.1, 2.3 (i.e. odds ratios of 1, 3 and 10). Results were generally consistent between effect sizes except in scenarios with meta-analyses containing only small studies (a selection of these results are presented in figure 7.8). In scenarios with rare events and small sample sizes, all methods have high negative bias regardless of the odds ratio effect size (therefore, the summary in table 7.1 can be generalised to all effect sizes in these scenarios).

In figure 7.8, mean bias is presented in meta-analyses with small studies and various underlying log odds ratios ($\theta = 0.5$, 1.1, 2.3). Odds ratio meta-analyses with $\theta = 0$ are not presented in this chapter because results are roughly consistent to $\theta = 0.5$ in all scenarios. All methods have more negative bias in the scenarios with a large effect size (C1-C3) than for low (A1-A4) and moderate (B1-B4) effect sizes. This may be partly due to a difference in τ^2 parameter values between these scenarios, given that results are consistent when $\tau^2 = 0$. REML is more negatively biased relative to other estimators in meta-analyses with a small effect size than with a large effect size. However, the difference in REML between effect sizes is marginal. HM is generally more robust to changes in effect size than other estimators, but still has considerable positive bias in scenarios of up to low heterogeneity.

Mean bias is the only performance measure presented in this section. However, this measure gives an understanding of the properties of methods according to other performance measures. Generally, results differ between effect sizes only in meta-analyses with small studies and where the event is not rare across both study groups (but could become rare in one of more study groups if the effect size is extreme enough). Appendix F.7 shows mean squared error of heterogeneity variance estimates

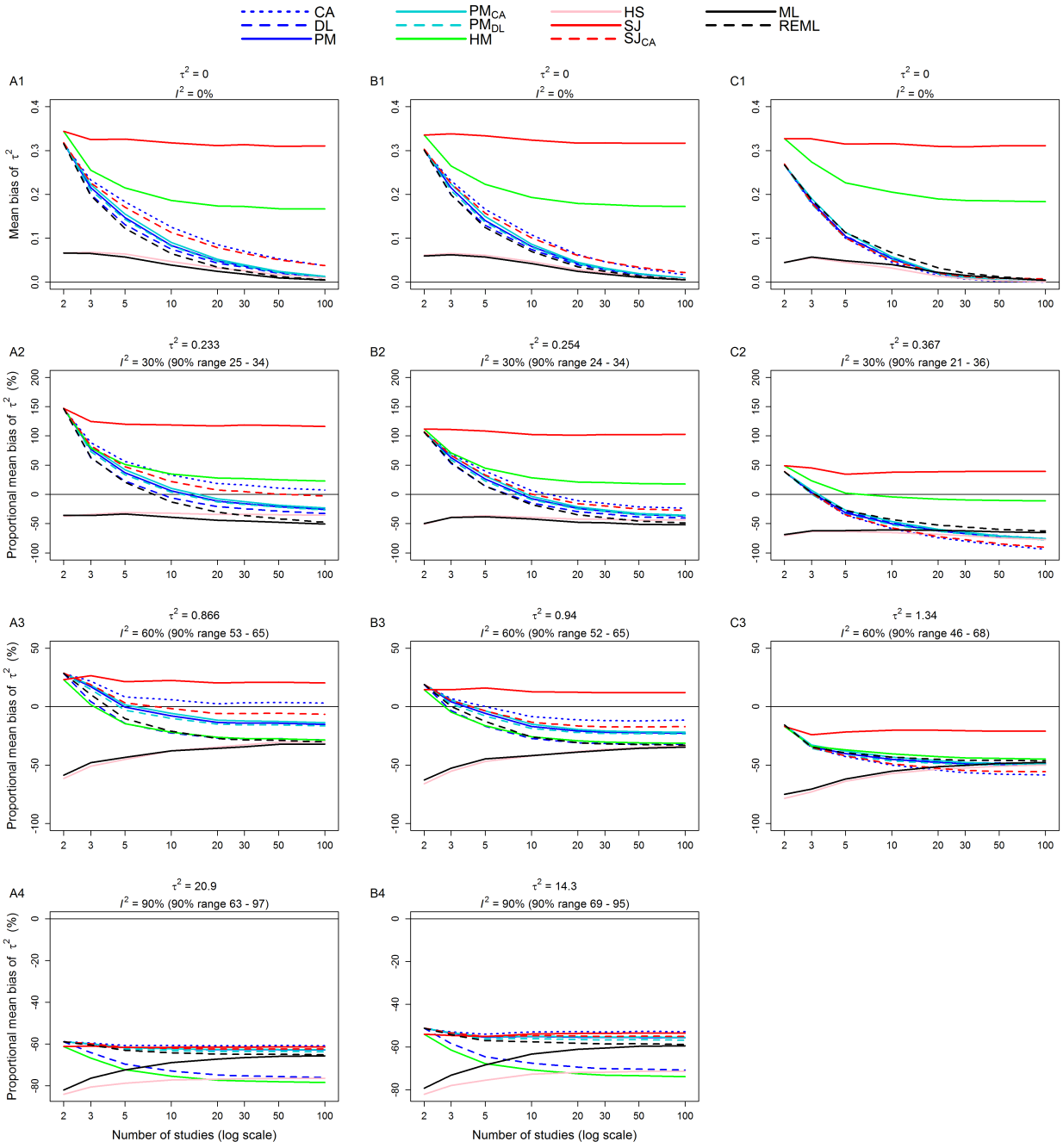


Figure 7.8: Mean bias of heterogeneity variance estimates in odds ratio meta-analyses containing small studies and with event probability 0.1 to 0.5

Scenarios with an underlying summary odds ratio of 1.65 (A1-A4), 3 (B1-B4) and 10 (C1-C3).

Bias is presented on the proportional scale only when $\tau^2 > 0$.
There was no such τ^2 that produced a mean I^2 of 90% when $\theta = 2.3$, so these scenarios are not presented.

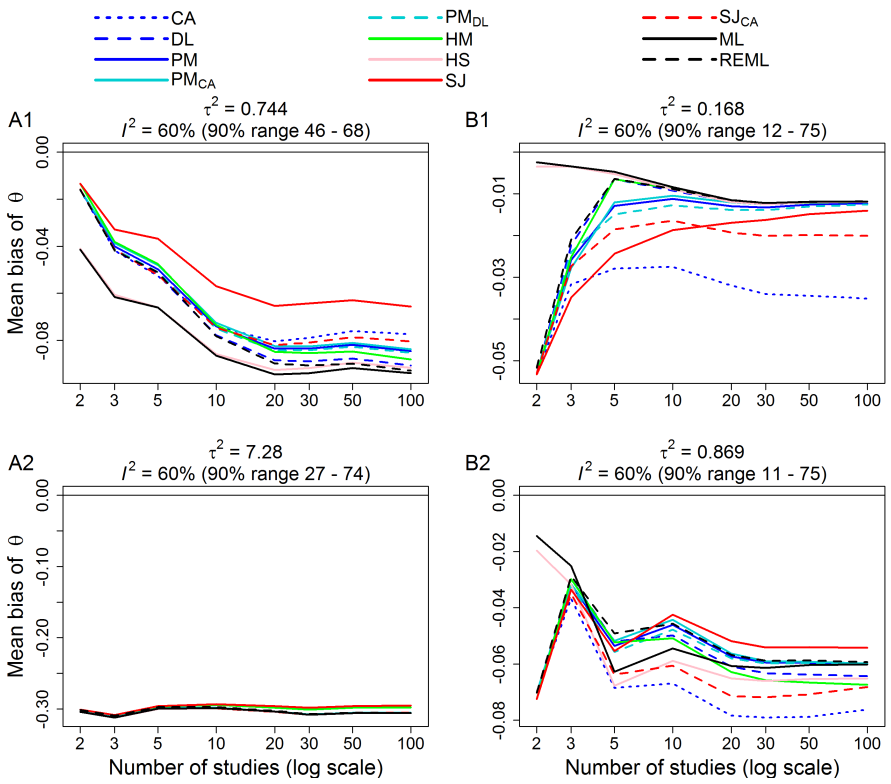


Figure 7.9: Mean bias of the summary effect estimates in odds ratio meta-analyses with rare events.
Scenarios presented are all with moderate I^2 (a mean of 60%) and (A1) Small-to-medium study sizes with an average event probability of 0.05; (B1) Small and large study sizes with an average event probability of 0.05; (A2) Small-to-medium studies with an average event probability of 0.01; (B2) Small and large studies with an average event probability of 0.01.
$\theta = 0.5$ and represents the underlying log odds ratio.

for the equivalent scenarios as in figure 7.8.

7.5.3 Properties of estimates of the summary effect

This section presents a comparison of heterogeneity variance estimators in terms of mean bias of the summary effect estimates. All meta-analyses were simulated with a true summary effect of 0.5, which represents either a standardised mean difference or a log odds ratio.

Results show that summary effect estimates of $\theta = 0.5$ are almost unbiased in all scenarios of standardised mean difference meta-analyses and odds ratio meta-analyses

with common events. However, summary effect estimates are negatively biased towards the null value of zero in odds ratio meta-analyses with rare events. In these scenarios, heterogeneity variance estimators also have considerable negative bias (see table 7.1). This indicates that the problem is not solely related to heterogeneity variance estimation, but problems with two-stage meta-analysis and perhaps the choice of continuity correction that affects all methods. Results from selected scenarios are presented in figure 7.9 of odds ratio meta-analyses with a rare event probability (0.01 and 0.05) and a mean I^2 of 60%. These scenarios were selected to show negative bias of summary effect estimates to varying degrees.

Estimators that produce larger estimates of the heterogeneity variance such as SJ, typically with positive bias, produce summary effect estimates with the lowest bias in scenarios with small-to-medium study sizes (plots A1 and A2). The opposite result is shown in plots B1 and B2 where studies are small and large; ML and HS produce the lowest estimates of heterogeneity and the least biased summary effect estimates. This is perhaps because larger heterogeneity variance estimates give studies more equal weight and this can reduce bias caused imprecise within-study variance estimates (as observed in plots A1 and A2). However, when there are large differences between study sizes, giving studies more equal weight can increase bias of the summary effect estimates (as observed in plots B1 and B2).

For scenarios with event probability 0.05, considerable negative bias in summary effect estimates is observed only when study sizes are small (as shown in plot A2). When the event probability is 0.01, considerable negative bias is observed in all odds ratio meta-analyses except when all studies are large. This can be observed in the full results in volume II of this thesis.

7.5.4 Coverage of 95% confidence intervals for the summary effect

Confidence intervals of the summary effect in this section are compared in terms of coverage. Confidence interval methods include Wald-type [25], t-distribution [28] and that proposed by Knapp and Hartung [38].

A representative subset of scenarios are presented before generalising the results to all scenarios. Results presented are from (1) standardised mean difference meta-analyses with small-to-medium studies (figure 7.10), (2) odds ratio meta-analyses with small-to-medium studies and 0.05 event probability (figure 7.11) and (3) odds ratio meta-analyses with small and large studies and event probability 0.1 to 0.5 (figure 7.12). The first scenarios represent ideal conditions and the final two scenarios represent conditions where methods generally perform more poorly. Scenarios of odds ratio meta-analyses with common events are excluded because results were consistent with the equivalent standardised mean difference meta-analyses. Results are plotted for τ^2 parameters that produce mean I^2 values of 0%, 30%, 60% and 90%. However, a mean I^2 of 90% was unattainable in some scenarios so these results are not included.

Results are given separately for each confidence interval method in sections 7.5.4.1 - 7.5.4.3 that follow, based on figures 7.10 to 7.12.

7.5.4.1 Wald-type confidence interval

The Wald-type 95% confidence interval is not robust to various simulated scenarios. Figure 7.10 shows coverage for scenarios of standardised mean difference meta-analyses with small-to-medium studies. In these scenarios, coverage can differ by up to 5% between heterogeneity variance estimators, up to 30% between numbers of studies and up to 20% between heterogeneity values. Coverage varies between 96-100% when studies are homogeneous and can be as low as 65% when the mean I^2 is 90% ($\tau^2 = 0.187$) and meta-analyses have two or three studies. When heterogeneity

is present, its coverage tends towards the nominal value of 95% as the number of studies increases.

In scenarios of odds ratio meta-analyses and an event probability of 0.05 (figure 7.11), coverage is above 90% when there are 20-30 studies. For meta-analyses with lower or higher numbers of studies, coverage is as low as 85%. In figure 7.12, derived from odds ratio meta-analyses with small and large studies, differences in coverage between heterogeneity variance estimators is up to 25%. For example, when $\tau^2 = 0.038$ (mean I^2 is 90%) and there are two or three studies in the meta-analysis, HS and ML has coverage as low as 60% while SJ and HM produce a confidence interval with coverage 85%.

In all scenarios, heterogeneity variance estimators that produce high estimates with positive bias (i.e. SJ, HM) tend to produce Wald-type confidence intervals with a higher coverage. Therefore these estimators work best with this confidence interval method when I^2 is high, given that coverage is typically low in these scenarios. By similar logic, HS and ML produce the lowest estimates of heterogeneity and generally work best with this method when I^2 is low. However, in meta-analyses with small and large studies (figure 7.12), CA produces the lowest coverage despite having positively biased heterogeneity variance estimates in these same scenarios. Perhaps this is because CA is the only estimator that assigns equal study weight and had high mean squared error in these scenarios.

7.5.4.2 t-distribution confidence interval

Coverage of the t-distribution 95% confidence interval is generally more robust to changes in the mean I^2, as shown in figure 7.10 in standardised mean difference meta-analyses. In these scenarios, however, coverage can differ by up to 5% depending on the heterogeneity variance estimator used and the number of studies. When there are 20 studies or more, 95% t-distribution confidence intervals have coverage 94-97%, but perform poorly with coverages close to 100% when there are fewer than 20

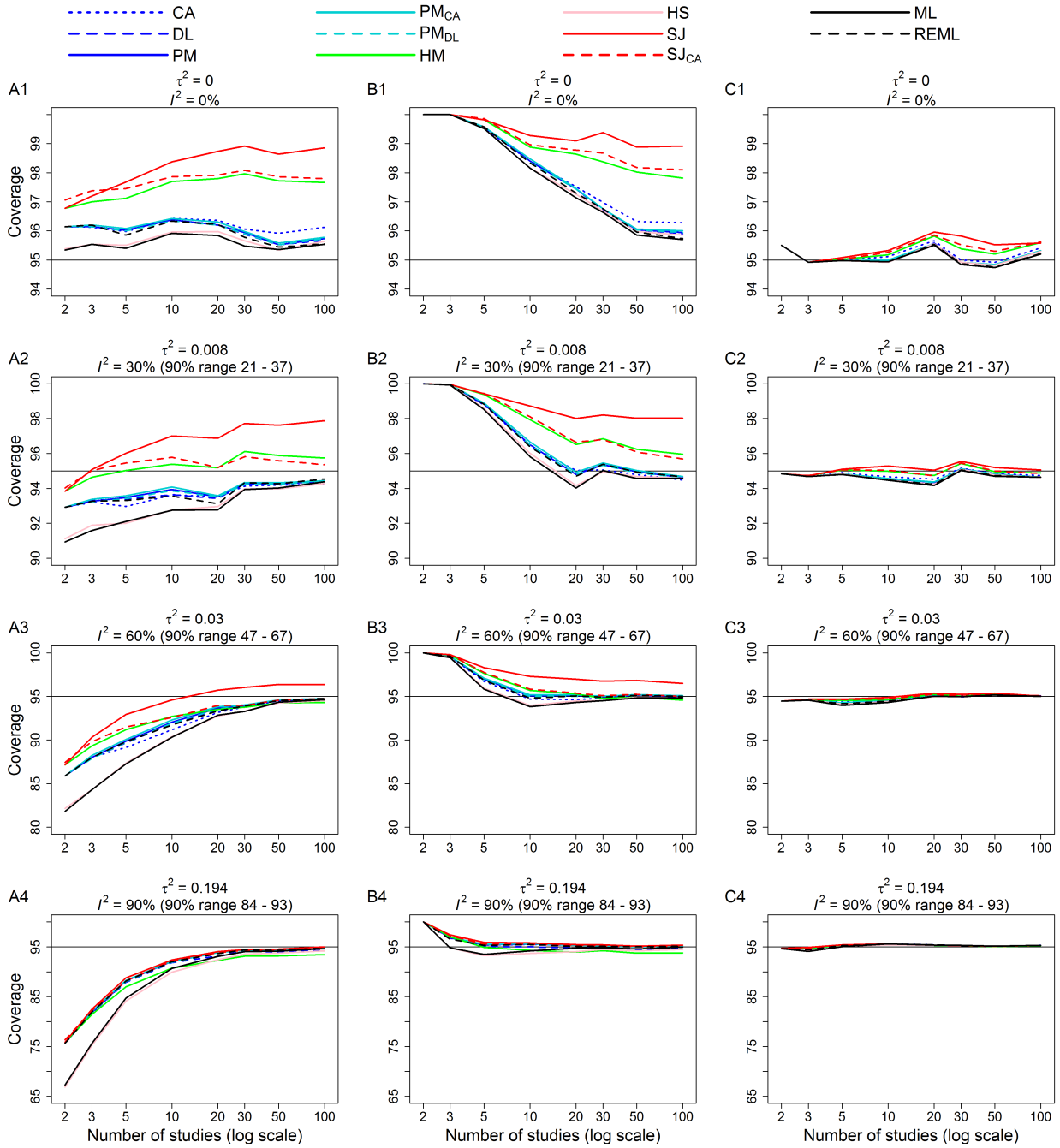


Figure 7.10: Coverage of 95% confidence intervals of the summary effect in standardised mean difference meta-analyses with small-to-medium studies
Coverage of Wald-type *(plots A1-A4), t-distribution (plots B1-B4) and Hartung-Knapp (plots C1-C4) confidence intervals presented. Effect size* $\theta = 0.5$.

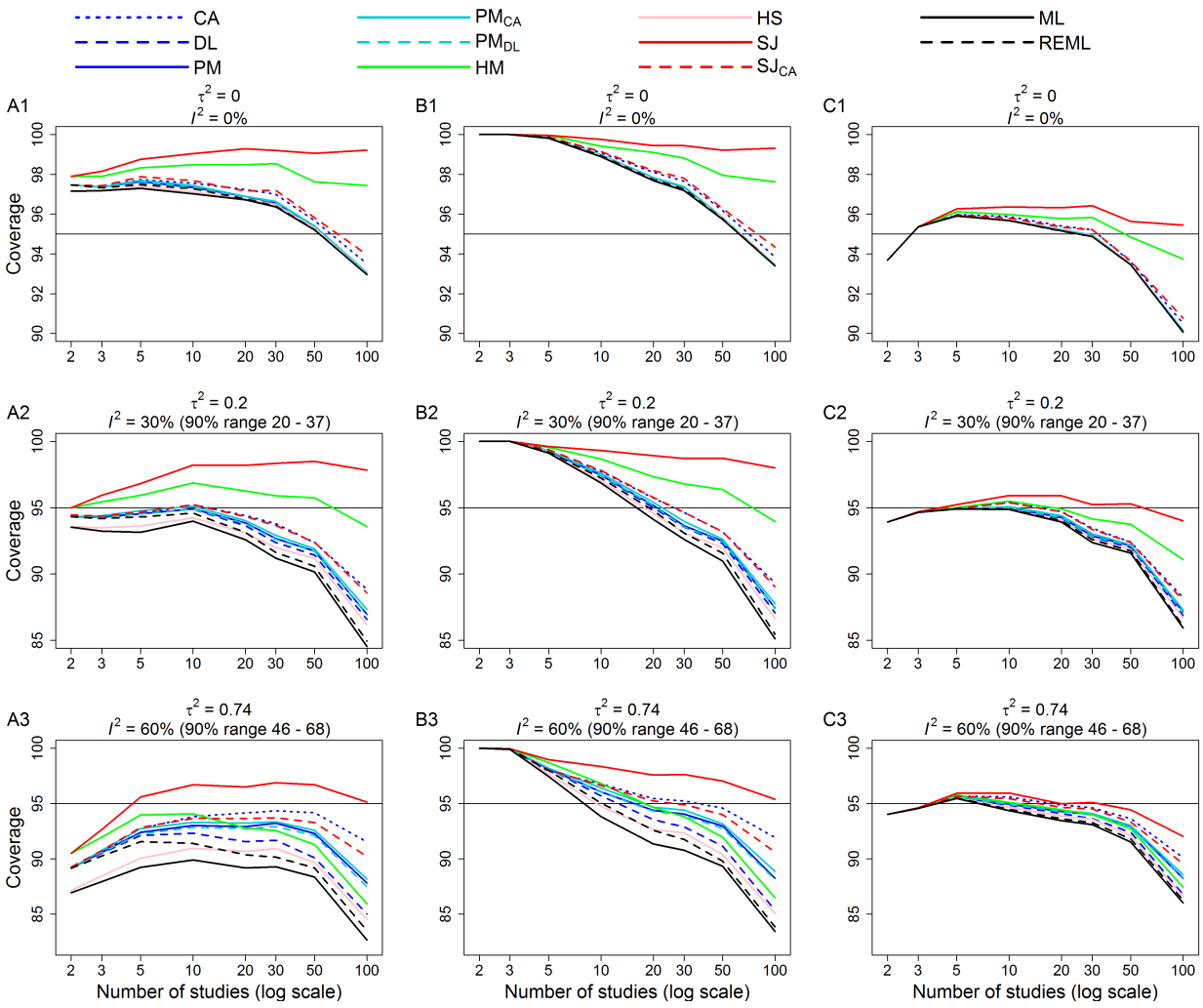


Figure 7.11: Coverage of 95% confidence intervals of the summary effect in odds ratio meta-analyses with small-to-medium studies and an average event probability of 0.05.

Coverage of Wald-type *(plots A1-A3), t-distribution (plots B1-B3) and Hartung-Knapp (plots C1-C3) confidence intervals presented. There was no such τ^2 that produced a mean I^2 of 90% so these scenarios are not presented. Effect size $\theta = 0.5$.*

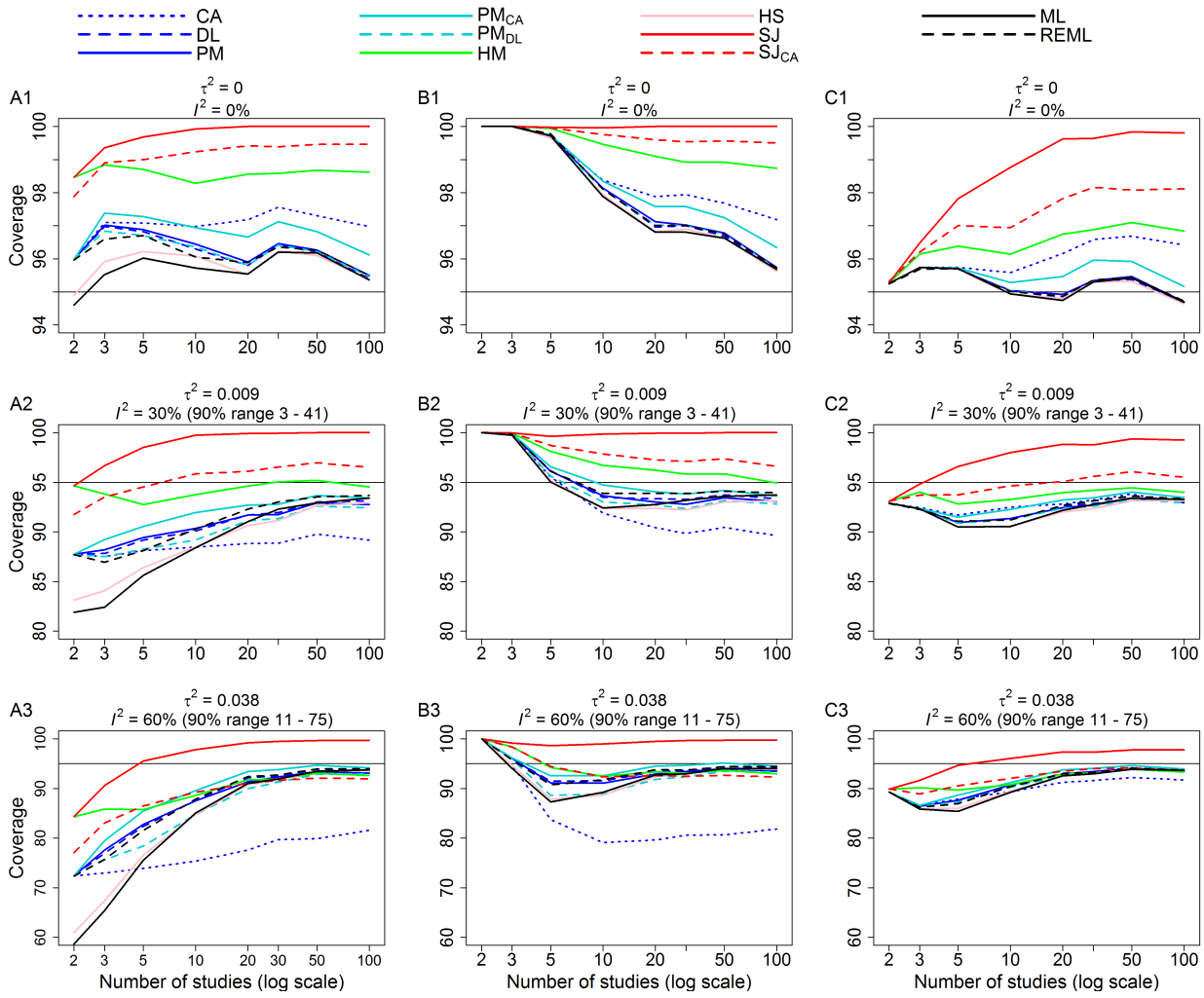


Figure 7.12: Coverage of 95% confidence intervals of the summary effect in odds ratio meta-analyses with small and large studies and an average event probability of 0.1 to 0.5.

Coverage of Wald-type *(plots A1-A3), t-distribution (plots B1-B3) and Hartung-Knapp (plots C1-C3) confidence intervals presented. There was no such* τ^2 *that produced a mean* I^2 *of 90% so these scenarios are not presented. Effect size* $\theta = 0.5$.

studies.

In odds ratio meta-analyses with event probability 0.05 (figure 7.12), the key difference is that coverage does not converge to 95% as the number of studies increases. Instead coverage remains close to 100% for meta-analyses with up to 10 studies and becomes as low as 85% in meta-analyses with 100 studies. In meta-analyses with small and large studies (figure 7.12), there is a greater difference between heterogeneity variance estimators; SJ has coverage close to 100% in all these scenarios and CA produce a confidence interval with coverage as low as 80%.

The heterogeneity variance estimator that works best with this confidence interval method varies considerably between scenarios, so it is difficult to select one overall.

7.5.4.3 Hartung-Knapp confidence interval

The Hartung-Knapp confidence interval for the summary effect has better coverage that the other two methods in all scenarios. This method has coverage 94-96% in standardised mean difference meta-analyses presented in figure 7.10 and insensitive to the choice of heterogeneity variance estimator. However, the Hartung-Knapp method's coverage can be far from optimal in other scenarios. In odds ratio meta-analyses with event probability 0.05 (figure 7.12), coverage decreases as the number of studies in the meta-analysis increases and can reach as low as 86%. In odds ratio meta-analyses with small and large studies, coverage is variable between estimators; HS and ML can produce coverage of 86% while SJ has coverage close to 95%.

The choice of heterogeneity has little impact on coverage in standardised mean difference meta-analyses with small-to-medium studies (figure 7.10), so all are equally good candidates to be used with this confidence interval method. In the other scenarios presented, coverage is too variable to select the best estimator overall.

<table>
<tr><td colspan="2" rowspan="2"></td><td colspan="5">Study sizes</td></tr>
<tr><td>Small</td><td>Small-to-medium</td><td>Medium</td><td>Small and large</td><td>Large</td></tr>
<tr><td colspan="2">SMD meta-analyses</td><td colspan="3">*
Z-type: Estimators have roughly 85% coverage when k=2. Coverage increases and is close to 95% when k ≥ 10.
t-distribution: All estimators have close to 100% coverage when k=2, some estimators converge to 95% as k increases.
Knapp-Hartung: Coverage very close to 95% in all scenarios and for all estimators.</td><td>Z-type and t-distribution: Similar results to the scenarios highlighted in white. The difference is there is a greater disparity between estimators. CA has low coverage and HM/SJ/BP have high coverage.

***

Knapp-Hartung: Coverage much poorer than in scenarios highlighted in white. 85-90% coverage when k=2 and converges closer to 95% when k≥20.</td><td rowspan="4"></td></tr>
<tr><td rowspan="3">OR meta-analyses with average event probability:</td><td>0.5</td></tr>
<tr><td>0.1 to 0.5</td></tr>
<tr><td>0.05</td><td>
**
For all confidence interval methods coverage fairly poor and lies somewhere between scenarios highlighted in white and dark grey(see figure).</td></tr>
<tr><td>0.01</td><td colspan="3">All estimators and all confidence interval methods have poor coverage.
Z-type and t-distribution: Both methods have similar coverage. All estimators have close to 100% coverage when k=2 and falls to below 70% for k=100.
Knapp-Hartung: Close to 95% coverage for all estimators and k=10, but <85% for low and high k.</td></tr>
</table>

Table 7.2: A summary of coverage for all scenarios of standardised mean difference and odds ratio meta-analyses with effect size 0.5
Scenarios with the same background shading show where coverage results are consistent for each confidence interval method
** Coverage presented in figure 7.10. ** Coverage presented in figure 7.11. *** Coverage presented in figure 7.12*
Recall, k denotes the number of studies in the meta-analysis

7.5.4.4 A summary of coverage in all simulated scenarios

Table 7.2 summarises coverage of all three confidence interval methods in all scenarios of standardised mean difference and odds ratio meta-analyses. The table is colour-coded to show where results are similar for each confidence interval method. All confidence interval methods could be summarised in one table because scenarios that can be grouped are consistent between methods.

All methods performed at their worst in scenarios of odds ratio meta-analyses with rare events (probability 0.05 or 0.01) and smaller study sizes. Methods also have poor coverage, but to a lesser extent, in meta-analyses with small and large studies. The

cause of poor coverage in these scenarios is that heterogeneity variance estimators have much greater variation in mean bias and therefore coverage also varies more between heterogeneity variance estimators. Coverage in standardised mean difference meta-analyses is consistent with the equivalent odds ratio meta-analyses if the event is common.

In all scenarios the Hartung-Knapp method has substantially better coverage and consistently produces confidence intervals with coverage close to 95% in standardised mean difference meta-analyses and most odds ratio meta-analyses with common events.

7.5.5 Convergence of ML and REML estimates of heterogeneity

ML and REML rarely failed to converge to a heterogeneity variance estimate. Fewer than 0.02% of meta-analyses failed to converge and only in meta-analyses with few studies and large differences in study sizes. A summary of these results are given in appendix D.3 (tables F.2 and F.3).

7.5.6 An overview of the results

Results showed that properties of estimates of the heterogeneity variance are dependent on the level of heterogeneity, number of studies in the meta-analysis, distribution of sample sizes, the probability of the event outcome and to a lesser extent the size of the effect in binary outcome meta-analyses. Estimates of the summary effect, and to a lesser extent, confidence intervals of the summary effect are more robust to changes in the heterogeneity variance estimate.

7.5.6.1 Properties of estimates of the heterogeneity variance

Three performance measures that related to estimation of the heterogeneity variance are presented in this chapter; mean bias, mean squared error and proportion of zero estimates.

B0, BP and MBH generally have the worst properties and as such were not presented in the main results of this chapter. B0 and BP have the highest positive bias of all estimators compared and MBH has the highest mean squared error. The main results also show that estimates calculated from SJ, SJ_{CA}, HS and ML generally have poor properties and should not be used in any meta-analysis. SJ also has considerable positive bias in meta-analyses for up to moderate levels of heterogeneity. SJ_{CA} has relatively low bias and mean squared error in meta-analyses with small studies but has considerable positive bias in meta-analyses with large studies. This is perhaps because SJ_{CA} is a non-truncated two-step estimator with a minimum initial τ^2 estimate of 0.01; this value represents high inconsistency in meta-analyses with large studies and is therefore not an appropriate cut-off. HS and ML have similar performance and are negatively biased in all meta-analyses when there are few studies in a meta-analysis, particularly when the mean I^2 is high.

CA, PM_{CA}, HM can also be excluded from consideration as there are alternative methods that have equal or better properties. CA and PM_{CA} have a bias and mean squared error comparable with many other estimators when study sizes are equal-sized but they have increasing positive bias and mean squared error as the difference between study sizes increase. HM is shown to have similar properties as DL but has more positive and negative bias when the I^2 is low and high respectively.

The remaining four methods with reasonable properties are DL, PM_{DL}, PM and REML. These are estimators that are shortlisted for possible recommendation in the conclusions of this chapter. DL is one of the best performing estimators in meta-analyses with large differences in study size. PM and PM_{DL} in most scenarios have similar properties and are more robust to imprecise within-study variances.

PM_{DL} performs better than PM with similar properties to DL when there are large differences in study size. REML generally has low mean squared error and a relatively low negative bias where DL has considerably more. However, in very few meta-analyses, the iterative method failed to produce a REML estimate (see section 7.5.5).

7.5.6.2 Properties of estimates of the summary effect

Mean bias and mean squared error of the inverse-variance summary effect estimates were presented in this chapter. In scenarios of standardised mean difference meta-analyses and of odds ratio meta-analyses with common events, results show summary effect estimates are approximately unbiased for all heterogeneity variance estimators. Estimates of the odds ratio summary effect are biased towards the null value when studies are simulated with a low event probability of 0.05 and 0.01, where all heterogeneity variance estimators also have considerable negative bias.

7.5.6.3 Properties of 95% confidence intervals of the summary effect

In section 7.5.4, coverage is presented as the only performance measure that relates to estimation of 95% confidence intervals for the summary effect. Coverage of Wald-type [25], t-distribution [28] and Hartung-Knapp [38] confidence interval methods were presented.

The Hartung-Knapp confidence interval has more optimal coverage than Z-type and t-distribution confidence intervals in nearly all scenarios. This method has coverage close to the nominal 95% in most scenarios of standardised mean difference meta-analyses or odds ratio meta-analyses with common events and robust to the choice of heterogeneity variance estimator. However, in odds ratio meta-analyses with rare events or when differences between study sizes are large, coverage probabilities of all confidence interval methods decrease to as low as 85%.

7.6 Discussion

The DerSimonian-Laird (DL) estimator cannot be recommended for wide-spread use in random effects meta-analysis, given that it has substantial negative bias in odds ratio meta-analyses with small studies and rare events. This finding can perhaps be explained by DerSimonian-Laird's fixed-effect study weights that are based solely on estimated within-study variances; these variances are imprecise and likely to be biased under such conditions. This negative bias of DerSimonian-Laird estimates has also been observed in previous simulation studies [3, 74, 78, 102] as identified in the systematic review in chapter 5. Viechtbauer [124] and Böhning et al. [7] stated that DerSimonian-Laird is asymptotically unbiased when within-study variances are known. Finding from this study also show DerSimonian-Laird has good properties in meta-analyses with large differences in study size and could be recommended, among other estimators, in this setting.

One of the primary aims was to investigate when it is appropriate to rely on one estimate of the heterogeneity variance. Results show all estimators are imprecise and often fail to detect high levels of heterogeneity in meta-analyses containing fewer than 10 studies. Chapter 4 reported that only 14% on meta-analyses in the *Cochrane Database of Systematic Reviews* (CDSR) contain 10 studies or more, so it is rarely appropriate to rely on one estimate of heterogeneity in this setting. All estimators have poor properties even in meta-analyses containing high numbers of studies when study sizes are small or the event of interest is rare (as shown in table 7.1). How frequently these scenarios occur in practice is investigated in the next chapter.

Estimates of the summary effect and its Hartung-Knapp confidence interval are of less cause for concern, performing well even for low numbers of studies. However, caution must still be applied when dealing with meta-analysis datasets with rare events, where summary effects are biased and any of the included confidence interval methods can have coverage as low as 85%. These findings agree with a previous simulation study [55], in which the Hartung-Knapp method (using the DL heterogen-

eity variance estimate only) was compared with other confidence interval methods for both continuous and binary outcome measures. The results in this chapter also show the Hartung-Knapp method is robust to changes in the heterogeneity variance estimate, except in meta-analyses with large differences in study size.

Results are in disagreement with some previous simulation studies. In all cases, this can be attributed to differences in parameter values and other differences in study design. SJ, MBH, B0 and BP performed well in previous simulations and have been recommended by their respective authors, yet this study shows they have poor properties. SJ performed well in simulations conducted by Sidik and Jonkman [101], yet simulations in this study shows they have considerable positive bias in meta-analyses of up to moderate I^2. This was not observed by Sidik and Jonkman [101] because meta-analyses were only simulated with high I^2. MBH has high mean squared error in meta-analyses with few studies, but Malzahn et al. [74] only simulated meta-analyses with 15 studies. B0 and BP were recommended because they have low mean squared error in meta-analyses with few studies. These methods are considerably biased, but bias was not included as a performance measure in the original study [93]. Furthermore, PM has been recommended based on the results of three previous simulation studies [3, 78, 79], but these studies did not simulate meta-analyses with moderate-large differences in study size, where PM has considerable positive bias.

Meta-analysis data were simulated from five distributions of study sample sizes. These distributions produced small, medium and large equally-sized studies and studies with moderate and large differences in size. DL and HM use fixed-effect study weights and have the best properties in meta-analyses with large differences in study size. PM, which uses random-effects weights, has better properties when studies are equal-sized. A possible explanation of these findings is that random-effects weights can be unduly similar in small and large studies when the heterogeneity variance estimate is large [46]. The two-step PM_{DL} estimator can be considered a robust compromise between DL and PM, since it uses fixed-effect study weights in

the first step and random-effects study weights in the second. Other estimators were recommended over $\mathrm{PM}_{\mathrm{DL}}$in two previous simulation studies [3, 64], but neither simulated meta-analyses with large differences in study size and neither reported the mean squared error of heterogeneity variance estimates.

There are two main advantages to the design of this simulation study over previous studies. First, a comprehensive set of heterogeneity variance estimators are compared in a wide range of scenarios and reported a wide range of performance measures. Previous simulation studies gave conflicting recommendations because their results only gave a limited picture. Second, meta-analyses were simulated using τ^2 parameter values that varied between scenarios and defined such that meta-analyses represent a consistent and wide range of I^2 values. Results suggest that properties of heterogeneity variance estimators are more comparable between scenarios with the same I^2, rather than the same τ^2. The I^2 statistic also takes into account the 'typical' within-study variances and is a measure of inconsistency between studies. Previous simulation studies set τ^2 parameter values in many cases with little knowledge of whether they constitute low, moderate or high levels of heterogeneity.

The limitations of this simulation study are as follows. First, only a subset of all confidence interval methods for the summary effect are included. Results show the Hartung-Knapp method is a more robust than the Z-type method to changes in the heterogeneity variance estimator, but no conclusive recommendations can be made going forward. Other methods exist such as the profile likelihood method [37], which has also been shown as a better alternative to the Z-type method in simulated meta-analysis data [45] and recommended elsewhere [20]. Second, a continuity correction of 0.5 was applied wherever simulated studies with a binary outcome contained zero events, but other better-performing methods are available [113]. This choice may have impacted the results in scenarios where the event is rare, but was chosen in this study because it is widely used. For each scenario, the probability that a study has zero events was calculated retrospectively and shown in the table F.4 of the appendix. Finally, the five distributions from which sample sizes were drawn can't

be considered representative of all distributions observed in practice; study sample sizes are unlikely to conform to a defined distribution.

Summarising the properties of a comprehensive list of heterogeneity variance estimators, compared over many combinations of parameter values was the biggest challenge of this study. By simulating meta-analyses from a wide range parameter values, inevitably there are scenarios that reflect meta-analyses rarely observed in practice. For example, most meta-analyses contain very few studies [21], but meta-analyses with up to 100 studies were simulated in order to show results over the full range of possible meta-analysis sizes. When interpreting results and drawing conclusions, equal consideration was given to rare and common scenarios. In the next chapter, results from a secondary analysis of the simulation data is presented using novel analysis methods that take into account the characteristics of meta-analyses in the *Cochrane Database of Systematic Reviews* (CDSR).

7.7 Conclusions

The DerSimonian-Laird two-step estimator (PM_{DL}) and REML have similar properties in both standardised mean difference and odds ratio two-stage meta-analyses. REML is recommended over PM_{DL} on the basis of these results because it's already widely known, available in most statistical software packages, and rarely fails to converge using Fisher's scoring algorithm. PM_{DL} is recommended as an alternative when convergence fails. The Hartung-Knapp confidence interval for the summary effect is generally recommended over other Wald-type and t-distribution methods compared in this study, but other methods not included may have better coverage in meta-analyses with rare events. To be consistent, we recommend the same REML estimate of the heterogeneity variance to calculate this confidence interval. However, this is inconsequential given how robust this confidence interval is to changes in the heterogeneity variance method in most scenarios. REML, or indeed any other single estimate of heterogeneity, should not be relied on to gauge the extent of heterogen-

eity in most meta-analyses. However, this single estimate can be used calculate a reliable Hartung-Knapp confidence interval for the summary effect.

Chapter 8

Properties of heterogeneity variance estimators in meta-analyses of Cochrane reviews

8.1 Introduction

In the last chapter, I simulated meta-analysis data that represented a wide range of meta-analyses occurring in practice. Results from all these simulations were presented and considered when drawing conclusions. However, this analysis approach did not account for the possibility that some simulation scenarios may be more representative of real meta-analyses than others. Those that represent meta-analyses more frequently occurring in practice should arguably have more bearing on the conclusions.

In this chapter, I implement a novel and systematic method of focusing more on these scenarios representative of real meta-analyses. I combine the findings from my simulated meta-analysis data with empirical data from the *Cochrane Database of Systematic Reviews* (CDSR) [21]. This CDSR dataset was used to perform an empirical comparison of heterogeneity variance estimators in chapter 4. I include the same 12,894 meta-analyses as I did in chapter 4. Recall that CDSR meta-analyses containing fewer than three studies are excluded; studies in these 'meta-analyses' are unlikely to have been synthesised and therefore it's unlikely an estimate of the heterogeneity variance was presented.

The principle aim of this analysis is to provide a clear and concise summary of the simulation results to lead into the concluding chapter. I also aim to: (1) describe the absolute performance of heterogeneity variance estimators in meta-analyses in practice; (2) distinguish between the heterogeneity variance estimators that I identified in the last chapter as having reasonable but similar properties; and (3) show the potential consequence of using heterogeneity variance estimators with poor properties.

8.2 Methods

To summarise the performance of heterogeneity variance estimators expected in CDSR meta-analyses, analysis was carried out in three steps. First, I mapped each

CDSR meta-analysis to a simulated scenario with the closest matching characteristics. The methods for this process are detailed in the next section. Second, I calculated the total number of CDSR meta-analyses matched to each scenario. Finally, by combining these frequencies and the simulation results, I derive a predicted distribution of each estimators performance in CDSR meta-analyses. The performance measures included in this chapter are given in section 8.2.2, the heterogeneity variance estimators included are detailed in section 8.2.3 and analysis methods are in section 8.2.4.

8.2.1 Mapping empirical to simulated meta-analyses

I mapped every included CDSR meta-analysis to a simulated scenario with the closest matching characteristics. Six meta-analysis characteristics are considered in this process: (1) the type of outcome measure, (2) the number of studies, (3) the level of inconsistency between study effects (estimated by I^2), (4) the summary effect, (5) the distribution of study sample sizes and (6) the average event probability in each study (binary outcome meta-analyses only). I mapped all meta-analyses to scenarios with normally distributed study effects because results show that all heterogeneity variance estimators are robust to non-normal effects. CDSR meta-analyses with a binary outcome were matched with one of 2,560 simulated scenarios of OR meta-analyses. Those with a continuous outcome were matched with one of 160 simulated scenarios of SMD meta-analyses. Matching criteria for all other characteristics are given in table 8.1.

I used the Sidik-Jonkman (SJ) estimate of I^2 in CDSR meta-analyses (see section 2.3.2). SJ was chosen because it only produces positive heterogeneity variance estimates, which is advantageous for two reasons. First, the distribution of SJ estimates of I^2 is likely to be more realistic of the underlying distribution, given that truncated methods produce an unrealistic proportion of zero estimates (see chapter 4). Second, it minimises the number of meta-analyses matching with scenarios where $I^2 = 0\%$. The reason this is beneficial is made clear in section 8.3.2.

Parameter	Parameter value/distribution	Empirical matching criteria
Number of studies in the meta-analysis (k)	2, 3, 5, 10, 20, 30, 50, 100	The closest value. When k is equidistant between two scenarios (i.e. if $k = 4$), the meta-analysis is matched to one of the two closest at random.
Mean I^2 for each scenario	0%, 30%, 60%, 90%	The closest I^2 estimate. SJ is used to estimate I^2 in CDSR meta-analyses using formula 1.6 for I^2 (chapter 1). e.g. SJ estimates of I^2 from 15% to 45% mapped to the scenario with $I^2 = 30\%$
Summary effect (θ)	In SMD meta-analyses, $\theta = 0.5$. In log-odds ratio meta-analyses, $\theta = 0$, 0.5, 1.1, 2.3	The closest absolute θ estimate. All SMD meta-analyses matched to $\theta = 0.5$.
Distribution of true study effects (θ_i)	*(a)* Normal distribution, *(b)* normal distribution with moderate skew and *(c)* normal distribution with high skew	All CDSR meta-analyses matched to scenario *(a)*.
Study sample sizes (n_{1i}, n_{2i})*	*(a)* Small studies: $n_{1i} = 20$	$n_{1i} + n_{2i} < 50$ for all studies
	(b) Small to medium sized studies: $n_{1i} \sim U(20, 200)$	$n_{1i} + n_{2i} < 500$ for all studies; and $n_{1i} + n_{2i} < 50$ for at least one study
	(c) Medium sized studies: $n_{1i} = 200$	$50 \leq n_{1i} + n_{2i} < 500$ for all studies
	(d) Small and large studies: $n_{11}, ..., n_{1m} = 20$ and $n_{1m}, ..., n_{1k} \sim U(1000, 2000)$ where m is the integer half way between 1 and k (when k is odd, one study is be generated from one of the two distributions at random)	$n_{1i} + n_{2i} \geq 500$ for at least one study; and $n_{1i} + n_{2i} < 50$ for at least one study
	(e) Large studies: $n_{1i} \sim U(1000, 2000)$	$n_{1i} + n_{2i} \geq 50$ for all studies; and $n_{1i} + n_{2i} \geq 500$ for at least one study
Parameters only applying to **odds ratio meta-analyses**		
Average event probability in study ($\overline{p_i}$)	*(a)* $\overline{p_i} = 0.5$	$\overline{p_i} \geq 0.1$ and $sd(\overline{p_i}) < 0.05$ ** †
	(b) $\overline{p_i} \sim U(0.1, 0.5)$	$\overline{p_i} \geq 0.1$ and $sd(\overline{p_i}) \geq 0.05$ ** †
	(c) $\overline{p_i} = 0.05$	$0.025 \leq \overline{p_i} < 0.1$ **
	(d) $\overline{p_i} = 0.01$	$\overline{p_i} < 0.025$ **

Table 8.1: Matching criteria for simulated and empirical CDSR meta-analysis data

**In all scenarios, sample sizes are equal between groups ($n_{1i} = n_{2i}$)*

*** In CDSR meta-analyses, $\overline{p_i}$ is estimated by the proportion of events in both groups combined (i.e. $p_i = (a_i + c_i) / (n_{1i} + n_{2i})$ using the notation from section 1.3.2 in the introduction chapter)*

† The cut-off value of 0.05 for the standard deviation is roughly half way between the standard deviations of scenarios (a) and (b) respectively

Study sample sizes and event probabilities are simulated from various distributions, so matching CDSR meta-analyses to these is more difficult. Empirical study sample sizes are unlikely to come from some natural distribution and I generated study sample sizes from a limited number of distributions. I took a pragmatic approach to address this issue and define matching criteria in table 8.1. These criteria were simple to implement and I believe lead to reasonably unbiased results. Nevertheless, I applied caution when interpreting the results because of the limitations of these methods.

8.2.2 Performance measures

I predict the performance of CDSR meta-analyses according to four of the five performance measures reported in the previous chapter of simulation results, namely:

- Proportional bias of heterogeneity variance estimates
- Proportional mean squared error (MSE) of heterogeneity variance estimates
- Mean bias of the summary effect estimates
- Coverage of 95% confidence intervals of the summary effect

Performance measures relating directly to the heterogeneity variance parameter are presented on the proportional scale so that results can be combined between scenarios with different parameter values. The scenarios with homogeneous study effects (i.e. $\tau^2 = 0$) cannot be presented on the proportional scale, so they are excluded from the analysis of these measures. Coverage is presented for all confidence interval methods included in the previous chapter of simulation results; Wald-type, t-distribution and Hartung-Knapp methods. The proportion of zero estimates of the heterogeneity variance is not reported in this analysis but was reported in the previous chapter. I excluded this measure because results would be analogous with the proportion of observed of zero estimates from CDSR meta-analyses in chapter 4.

8.2.3 Included estimators of the heterogeneity variance

I compared six heterogeneity variance estimators in this analysis that were chosen based on simulated results in the last chapter. DerSimonian-Laird (DL), Paule-Mandel (PM), the two step Paule-Mandel (PM_{DL}) and restricted maximum likelihood (REML) were included because they have the best properties overall. Sidik-Jonkman (SJ) and maximum likelihood (ML) were also included because these produce heterogeneity variance estimates with the most positive and negative bias respectively. SJ and ML were included to show the potential consequence of using estimators with poor properties and highlight the added benefit of using estimators with more reasonable properties.

8.2.4 Analysis methods

I present the results in two parts. First, I summarise how many CDSR meta-analyses are matched to each scenario based on the criteria defined in section 8.2.1. This summary is in the form of a heat map that highlights the key scenarios that are likely to be most representative of CDSR meta-analyses. The heat map is presented in such a way that it can be directly compared with results tables 7.1 and 7.2 in the last chapter. These tables summarise the scenarios that cause problems with heterogeneity variance estimation.

Second, I present the distribution of performance in CDSR meta-analyses according to each of the included four measures. These could alternatively be described as a weighted distribution of performance, with weights defined as the number of CDSR meta-analyses matched to each scenario. Distributions were derived for the six heterogeneity variance estimators and compared in the same plot and results of OR and SMD meta-analyses are presented separately. These distributions are not naturally smooth given the finite number of scenarios they are based on. Therefore, to plot them clearly, I defined appropriate intervals for each measure and calculate

the proportion of CDSR meta-analyses mapped to scenarios within each interval. The proportion in each interval are shown on the y-axis of each plot. Performance measures are presented on the x-axes on appropriate log-scales to focus on the region that represents optimal performance (e.g. where MSE is close to zero).

8.3 Results

I present the number of CDSR meta-analyses matched to each scenario in section 8.3.1 to show which simulated scenarios are most representative of real meta-analyses. The main results of the analysis follow in section 8.3.2 onwards.

8.3.1 The proportion of CDSR meta-analyses matched to each simulated scenario

The proportion of CDSR meta-analyses that match to each simulated scenario are given in figure 8.1 in the form of a heat map. Each combination of sample size distribution and probability of event (in OR meta-analyses only) are presented in separate blocks. Within each block I present combinations of simulated I^2 values ($I^2 = 0\%$, 30%, 60%, 90%) and numbers of studies (3,5,10,20,30). Scenarios of meta-analyses containing 2 studies are excluded because these empirical meta-analyses were from the results (as they were in chapter 4, when the same data was used). Scenarios of meta-analyses containing 50 and 100 studies were excluded because they only account for 0.5% of SMD and OR meta-analyses. The number of meta-analyses matched to each summary effect parameter are not presented in the heat map to make the figure more concise. Overall, 4092 (45.6%) odds ratio meta-analyses were matched to $\theta = 0$, 3354 (37.4%) to $\theta = 0.5$, 1249 (13.9%) to $\theta = 1.1$ and 284 (3.2%) to $\theta = 2.3$.

CDSR meta-analyses are distributed fairly uniformly between scenarios with mean I^2 values of 0%, 30%, 60% and 90%. However, fewer meta-analyses are matched

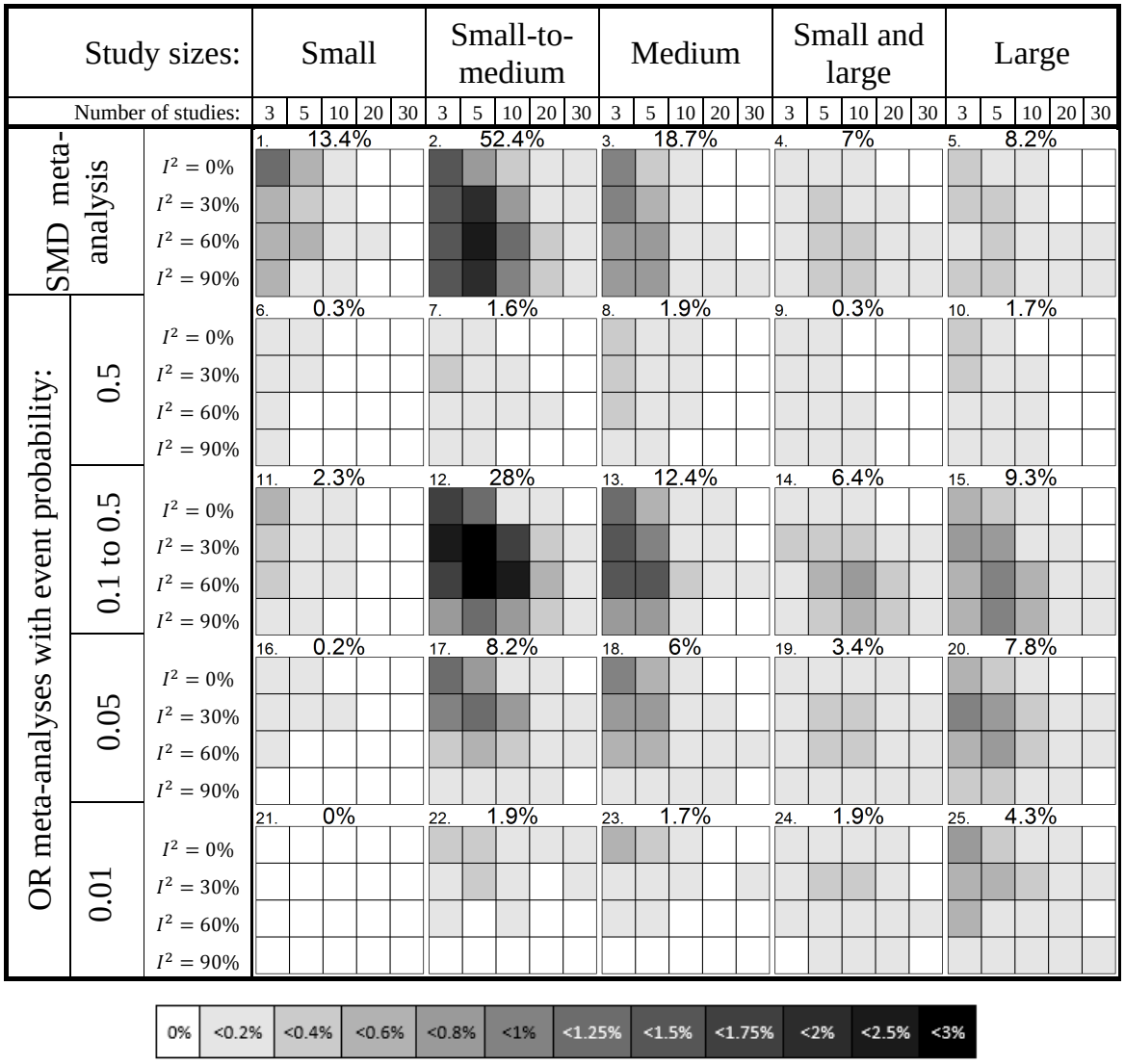


Figure 8.1: A heat map of CDSR meta-analyses falling into each simulated scenario

Percentages represent the proportion of meta-analyses matched to each block, out of the total of SMD or OR meta-analyses.

with scenarios where $I^2 = 0\%$ (i.e. where the SJ method produces $I^2 < 15\%$) because the Sidik-Jonkman only produces positive heterogeneity variance estimates; 1911 (21.3%) OR meta-analyses and 727 (18.6%) SMD meta-analyses are matched to these scenarios. I^2 are estimated in CDSR meta-analyses, so this is only a rough representation of the distribution of underlying I^2.

In the last chapter of results, I found that all heterogeneity estimation methods have considerable negative bias in meta-analyses with rare events, except where all study sizes are large. The heat map shows these scenarios represent 2,094 (23.3%) odds ratio meta-analyses (as shown in blocks 16-19 and 21-24 on the heat map). 229 (2.6%) of odds ratio meta-analyses are matched to scenarios with small studies and common events (blocks 6 and 11), where heterogeneity variance estimators have considerable negative bias when there is a high level of inconsistency between study effects. The remaining scenarios represent all 3,915 (100%) SMD meta-analyses and 6,656 (74.1%) OR meta-analyses, where most heterogeneity variance estimators have low bias, at least when the effect size is not extreme and there are sufficient numbers of studies. However, of these meta-analyses, 1650 (42.1%) SMD meta-analyses and 2739 (39.8%) OR meta-analyses are represented by the scenarios with only three studies. All heterogeneity variance estimates are imprecise and most have small to moderate positive bias in these scenarios.

8.3.2 Performance of heterogeneity variance estimators in CDSR meta-analyses

In this section, I present the predicted distributions of performance in CDSR meta-analyses. I present proportional bias and MSE of the heterogeneity variance estimators in sections 8.3.2.1 and 8.3.2.2. For these analyses, I excluded scenarios where $I^2 = 0\%$ as results cannot be presented on the proportional scale. Only a small number of CDSR meta-analyses matched with these scenarios, as shown in section 8.3. I present bias of the summary effect in section 8.3.2.3 and coverage of 95%

confidence intervals of the mean effect in section 8.3.2.4.

8.3.2.1 Predicted bias of the heterogeneity variance

Figure 8.2 shows the predicted distribution of proportional bias of heterogeneity variance estimators in CDSR meta-analyses. The x-axis is plotted on a log scale so that it expands around the point where bias is zero, i.e. the optimal bias. Results for other performance measures in this analysis are presented with x-axes on similar scales.

Figure 8.2 predicts that reasonably unbiased estimates of the heterogeneity variance are produced in few CDSR meta-analyses using any of the estimators compared. DL, PM_{DL} and REML are likely to produce the highest proportion of reasonably unbiased estimates; these methods are predicted to derive estimates with less than 10% bias (positive or negative) in roughly 40% of OR meta-analyses and 60% of SMD meta-analyses. Slightly fewer PM estimates are predicted to have bias under 10%; 35.9% of OR meta-analyses and 48.5% of SMD meta-analyses. However, these results suggest PM would produce the least negatively biased estimates of these four estimators in OR meta-analyses. This is because PM has marginally less bias in scenarios with rare events, where it is not recommended to rely on a single estimate of heterogeneity.

As expected from the results of the last chapter, results predict that ML and SJ have considerable bias in most CDSR meta-analyses. results predict that ML estimates are negatively biased by more than 10% in 95.9% of OR meta-analyses and 90.8% of SMD meta-analyses. Similarly, SJ is predicted to be positively biased (>10%) in 82.2% of OR meta-analyses and 70.4% of SMD meta-analyses.

Results predict that all estimators except ML would produce a much higher proportion of positively biased estimates than negatively biased. These estimators have positive bias in meta-analyses containing few studies, which represent most CDSR meta-analyses. It is widely noted that DL has negative bias in certain scenarios,

but the figure shows that in OR meta-analyses, 40.4% of DL estimates would be positively biased and only 15.2% negatively biased by more than 10%. Less than 1% of DL estimates in SMD meta-analyses have negative bias more than 10%.

8.3.2.2 Predicted mean squared error of the heterogeneity variance

The predicted distributions of proportional MSE are given in figure 8.4 for each of the six included heterogeneity variance estimators. The means of these distributions are given in table 8.2. The proportional MSE of heterogeneity variance estimates in OR meta-analyses is typically around 0.4-0.5, which shows estimates in these meta-analyses are usually imprecise. The proportional MSE in OR meta-analyses is typically four times higher than in SMD meta-analyses; given the difference in scale between OR and SMD outcome measures, they are expected to be only 1.81 times higher [15]. This 'additional' error can be attributed to scenarios with rare events.

These results confirm what was already noted in the last chapter; methods that produce a higher proportion of negatively biased heterogeneity variance estimates (i.e. ML and to a lesser extent DL) have a lower MSE. Of the four estimators included with reasonable properties (DL, PM, PM_{DL} and REML), PM has the highest MSE because it produces the least negatively biased estimates.

8.3.2.3 Predicted bias of summary effect estimates

Figure 8.3 shows the predicted distribution of bias of the summary effect in CDSR meta-analyses. Recall, the summary effect represents a log odds ratio in binary outcome meta-analyses and a standardised mean difference in continuous outcome meta-analyses. As expected from the results of the last chapter, bias of the summary effect is consistent between all heterogeneity variance estimators compared. Results predict that roughly 10% of OR meta-analyses produce summary effects that have small to moderate bias towards the null value. All SMD meta-analyses produce reasonably unbiased summary effect estimates.

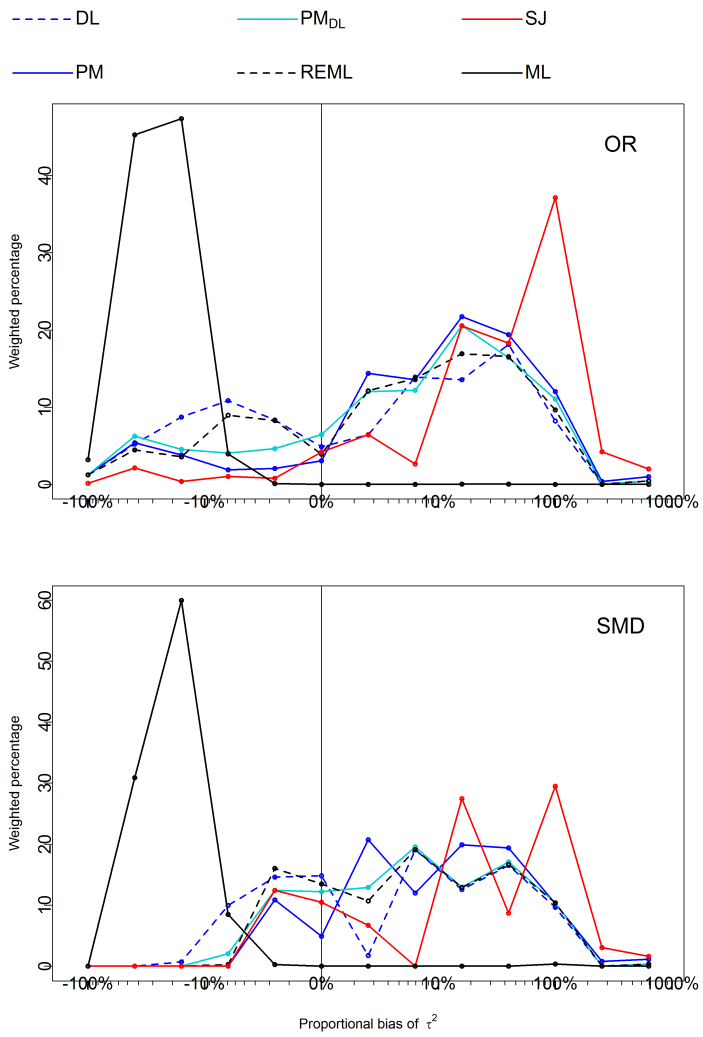


Figure 8.2: Predicted distribution of proportional bias of the heterogeneity variance estimators

x-axis presented on the log scale for bias >0 and the reverse-log scale for bias<0. Log scales are in base 10.

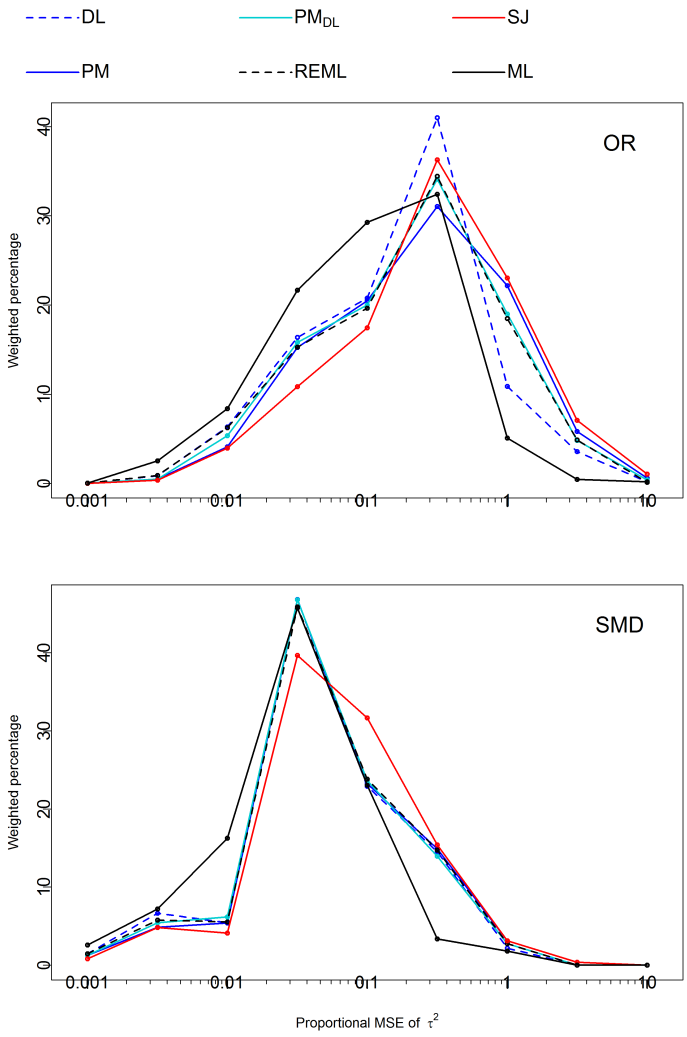


Figure 8.4: Predicted distribution of proportional MSE of heterogeneity variance estimates

x-axis presented on the log scale with base 10.

	Heterogeneity variance estimator					
Outcome type	DL	PM	PM_{DL}	REML	SJ	ML
OR	0.358	0.500	0.429	0.408	0.638	0.203
SMD	0.103	0.129	0.118	0.118	0.133	0.060

Table 8.2: The average proportional MSE of heterogeneity variance estimates in CDSR meta-analyses

These summary statistics are derived from the same results as in figure 8.4

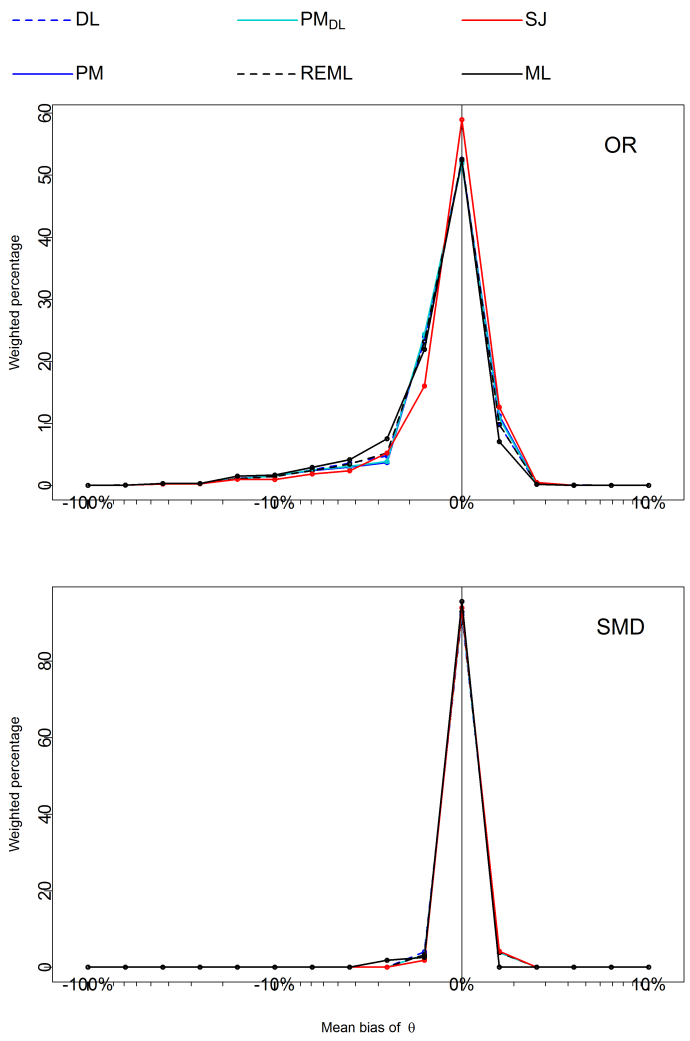


Figure 8.3: Predicted distribution of bias of the summary effect (θ)
x-axis presented on the log scale for bias >0 and the reverse-log scale for bias<0. Log scales are in base 10.

8.3.2.4 Predicted coverage of 95% confidence intervals of the summary effect

Coverage of 95% confidence intervals of the summary effect are shown in figure 8.5. Confidence intervals are calculated by Wald-type, t-distribution and Hartung-Knapp methods are presented in the same figure, separately for OR and SMD meta-analyses.

Figure 8.5 shows the Wald-type confidence interval method rarely produces confidence intervals with coverage close to 95%. ML produces a lower Z-type coverage than

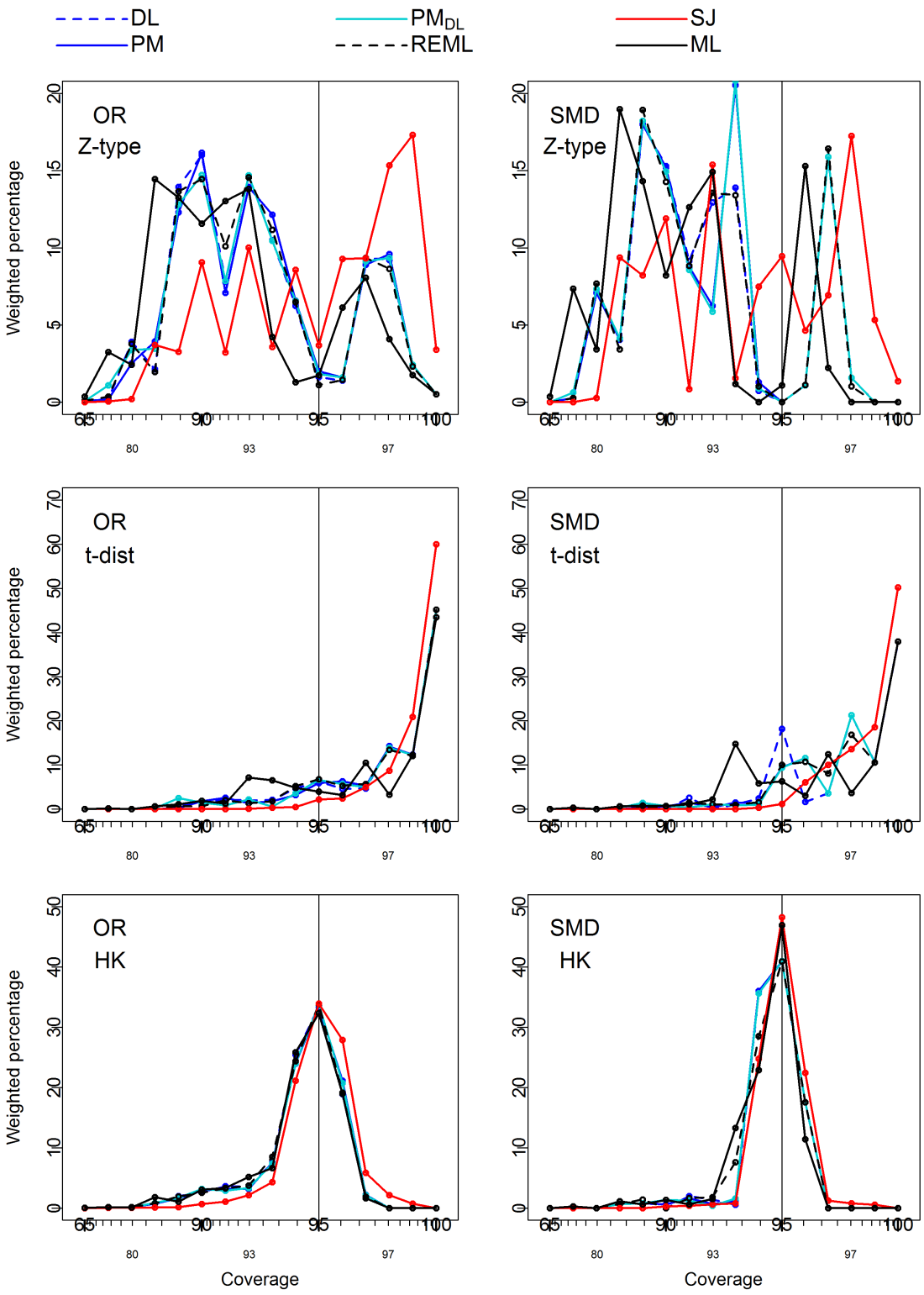


Figure 8.5: Predicted distribution of the coverage of summary effect confidence intervals; Wald-type, t-distribution and Hartung-Knapp.
x-axis presented on the log scale for coverage >95% and the reverse-log scale for coverage<95%. Log scales are in base 5.

SJ because of a large difference in bias between these methods. All other included heterogeneity variance estimators with reasonable properties (DL, PM, PM_{DL} and REML) show almost identical results. For these estimators, coverage is between 93% and 97% in 30% of OR meta-analyses and 35% of SMD meta-analyses. Coverage is below 93% in 57% of OR meta-analyses and 67% of SMD meta-analyses and substantially low (less than 85%) in 6.3% of OR meta-analyses and 11.4% of SMD meta-analyses. As shown in the last chapter of results, low Z-type coverage tends to come from meta-analyses with low numbers of studies and high underlying I^2.

For t-distribution confidence intervals, the predicted distribution of coverage is also similar between the heterogeneity variance estimators compared. Most meta-analyses produce t-distribution confidence intervals far above the nominal 95% level. 68% of OR meta-analyses and up to 65% of SMD meta-analyses have coverage above 97%. High coverage of the t-distribution confidence interval is observed in meta-analyses with low numbers of studies (as shown in the last chapter).

Hartung-Knapp confidence intervals typically perform much better than the other two confidence interval methods. Coverage is almost identical for the four heterogeneity variance estimators. 88% of OR meta-analyses and 95% of SMD meta-analyses have coverage between 93% and 97%. A small proportion of OR meta-analyses have coverage below 93%; these results come from scenarios with small studies and rare events.

8.4 Discussion

One of the main criticisms of the DL method is that it produces negatively biased heterogeneity variance estimates. This has been shown in the last chapter and in previous simulation studies identified in chapter 5. Results from this analysis suggests this negative bias is of concern in many meta-analyses from Cochrane reviews. However, less expectedly, the DL method is is predicted to be positively biased in

twice as many meta-analyses; results show 40% of CDSR meta-analyses have characteristics that produce DL heterogeneity variance estimates with positive bias of more than 10%. This is because DL is positively biased in meta-analyses containing few studies and these simulated scenarios constitute most CDSR meta-analyses. The predicted bias of other heterogeneity variance methods is only marginally better. I included three other methods in this analysis that showed reasonable properties in the last chapter; Paule-Mandel (PM), two-step DerSimonian-Laird (PM_{DL}) and REML. These methods would produce heterogeneity variance estimates with only marginally improved properties in meta-analyses in practice.

I compared heterogeneity variance estimators in terms of bias of the summary effect and my conclusions are in agreement with those in the last chapter. The predicted level of bias is consistent between all included heterogeneity variance estimators. Results suggest that summary effect estimates are approximately unbiased in all CDSR meta-analyses with a SMD outcome measure. 12% of meta-analyses with a log odds ratio outcome measure have negative bias greater than 0.1.

Finally, I compared heterogeneity variance estimators in terms of coverage of 95% confidence intervals of the summary effect, where confidence intervals were calculated from Wald-type, t-distribution and Hartung-Knapp methods. Predicted coverage is fairly consistent between heterogeneity variance estimators for all three confidence interval methods. Coverage of Wald-type and t-distribution confidence intervals are typically much further from the nominal 95% than anticipated in the last chapter. Wald-type confidence intervals are predicted to have coverage below 85% in up to 15% of SMD meta-analyses and up to 8% of OR meta-analyses. This confidence interval method typically performs better in OR meta-analyses, perhaps because more of these meta-analyses were matched to simulated meta-analyses with larger studies. t-distribution confidence intervals are predicted to have very poor performance in most CDSR meta-analyses, with coverage of over 97%. Hartung-Knapp confidence intervals showed results that were anticipated from the last chapter; they would perform well in most meta-analyses but should be used with caution in OR meta-

analyses with small studies and rare events.

This analysis provides a clear summary or the simulation results but can never replace the comprehensive exploration of results in chapter 7. The analysis method I used here can be considered practical and pragmatic but not without criticisms. First, matching the CDSR distributions of study sample sizes and event probabilities to simulation scenarios was problematic. It is unlikely that the simulations are representative of the CDSR meta-analyses in this respect. Second, by presenting bias and MSE of heterogeneity variance estimates on the proportional scale, I made the following implausible assumption: the consequence of a heterogeneity variance estimate with 100% error is the same regardless of the underlying parameter value. Finally, I excluded scenarios where studies were homogeneous ($\tau^2 = 0$) because results could not presented on the proportional scale. This is likely to have lead to an underestimate of the proportion of positively biased heterogeneity variance estimates. Furthermore, ML performs well in meta-analyses with homogeneous study effects, so these results will have exaggerated MLs negative bias. These issues are not likely to have affected my conclusions, given that I made them while also considering my results from the last chapter. Limitations of the simulation study in chapter 7 also apply here, given it is based on the same data. Most notably, I used a continuity correction of 0.5 for odds ratio meta-analyses with zero events, though other methods have been shown to perform better [113]. Other continuity corrections or methods could have improved the predicted performance of heterogeneity variance estimators in these results.

8.5 Conclusions

The REML heterogeneity variance estimator, recommended based on the results from the last chapter, has a similar predicted performance to the other three estimators with reasonable properties (DL, PM and PM_{DL}). The overriding conclusion of this analysis is that heterogeneity variance estimates in meta-analyses of Cochrane

reviews are most likely imprecise and biased. Meta-analyses rarely have characteristics in practice that allow for a single reliable point estimate of the heterogeneity variance. A more substantial improvement in the 95% confidence interval of the summary effect is usually possible if the Hartung-Knapp method is used over Wald-type or t-distribution methods. The method used to calculate the heterogeneity variance estimate is unlikely make a substantial impact on coverage of the Hartung-Knapp confidence interval. A single estimate of the summary effect and its random-effects confidence interval is usually sufficient even in meta-analyses with few studies and sensitivity analysis is usually not required in this respect. In random-effects meta-analyses, conclusions should not be drawn directly from a single point estimate of heterogeneity without first considering its uncertainty and likely level of bias.

Chapter 9

Discussion and conclusions

9.1 Introduction

There is often heterogeneity across studies in a meta-analysis that cannot be explained by known study characteristics. It is therefore common to assume a random-effects model, which includes an additional study variance component known as the heterogeneity variance parameter. This parameter is commonly estimated by the DerSimonian-Laird method [25]. Prior to conducting this research, simulation studies found that DerSimonian-Laird produces negatively biased estimates in certain scenarios [78, 79, 102, 124]. This estimator continues to be the default method for random-effects meta-analysis, partly because there is no consensus over which method, if any, should be used in its place. In this thesis, I reviewed available methods for heterogeneity variance estimation, investigated their properties in empirical and simulated meta-analysis data and made recommendations for future meta-analyses in health research.

A chapter-by-chapter summary of the content and main findings of this thesis is given in section 9.2. In section 9.3, I discuss the applicability of the findings from my research and its limitations. I make final conclusions in section 9.5.

9.2 Thesis summary

Chapters 1 to 3 are introductory chapters that detail all statistical methods for random-effects meta-analysis that are relevant to the rest of the thesis. In chapter 1, I introduced the concept of meta-analysis and methods for statistically combining studies to provide a summary effect. In chapter 2, I presented a comprehensive methodological review of heterogeneity variance estimators. I drew attention to methodological connections between methods and, in the case of Paule-Mandel and empirical Bayes, found they are identical and only expressed in different terms. In chapter 3, I introduced a number of methods for estimating the confidence interval of

the mean effect; it was deemed relevant to introduce these methods because I compare heterogeneity variance estimators in terms of their impact on these confidence intervals in many later chapters.

After collating and reviewing the relevant methods, I investigated empirically whether choosing a method other than DerSimonian-Laird significantly changes the heterogeneity variance estimate and conclusions of a meta-analysis. I compared a wide selection of methods in 12,894 meta-analyses from the *Cochrane Database of Systematic Reviews* (CDSR). Results showed high discordance between heterogeneity variance estimates between most methods, with differences on the scale of the I^2 statistic of up to 50%. I investigated whether meta-analysis characteristics, such as study sizes and sparsity of data, could have an impact on these differences. I found no convincing patterns, which suggests that differences are related to differences in the methods that apply regardless of the meta-analysis characteristics. Estimated summary effects derived from different heterogeneity estimation methods showed a much higher level of agreement. However, there was discordance in the level of statistical significance of the mean effect between methods in a small percentage of meta-analyses. Findings from this empirical analysis gave motivation for the rest of the thesis.

Next, in chapter 5, I presented the results of a systematic review of simulation studies that compare heterogeneity variance estimators. I identified twelve simulation studies that matched the inclusion criteria, but only four could be considered comprehensive and unbiased [78, 79, 96, 124]. The other eight simulation studies recommended their own newly proposed estimator and often compared them with very few other methods. I found the Paule-Mandel (PM) estimator performed well in both binary and continuous outcome meta-analyses and was recommended based on the findings of three simulation studies [3, 78, 79]. However, inconsistency between recommendations of other simulation studies, and a number of limitations in their designs, meant that a new simulation study was justified.

The design of a new simulation study is presented in chapter 6 and addresses the

limitations found in other previous simulation studies. To minimise the conflict of interest that was present in many previous studies, many collaborators gave input into the study design and only pre-existing estimators are compared. The main results of this simulation study are presented in chapter 7. Findings confirm the DerSimonian-Laird is negatively biased in binary outcome meta-analyses with rare events and/or where meta-analyses contain small studies; within-study variances are imprecise and often biased in these scenarios. The Paule-Mandel estimator, recommended most frequently in previous simulation studies, has better properties than DerSimonian-Laird overall and is negatively biased only in the most extreme cases where all methods perform poorly. However, results of this study revealed that Paule-Mandel estimates have higher positive bias in meta-analyses with moderate to large differences in study size. This can be attributed to Paule-Mandel's random-effects study weights, which can assign a relatively large weight to small studies. The two-step DerSimonian-Laird estimator or REML are a good compromise between these two methods.

A secondary analysis of the simulated data is presented in chapter 8. Results of this analysis predicted the likely properties of heterogeneity variance estimators in meta-analyses from the CDSR dataset. Findings from this analysis suggest that heterogeneity variance estimates are likely to be biased and imprecise in most meta-analyses in practice regardless of which method is used. The two-step DerSimonian-Laird estimator offers only a minimal improvement over DerSimonian-Laird. More promisingly, in most meta-analyses, estimates of the mean effect are unbiased and its Hartung-Knapp confidence interval has coverage close to the nominal 95%.

9.3 Discussion

I consider my thesis to have thoroughly examined the properties of heterogeneity variance estimators in frequentist meta-analyses. I assessed their properties in a wide range of scenarios in both binary and continuous outcome data, summarised using

odds ratio and standardised mean difference measures respectively. These measures make up only 37% and 24% of binary and continuous outcome CDSR meta-analyses respectively (see chapter 4), but I believe my findings apply to other measures. The relative risk is used in 62% of CDSR meta-analyses with a binary outcome, and my review of previous simulation studies (chapter 5) suggests properties in these meta-analyses are comparable with odds ratio meta-analyses. My findings can also apply to meta-analyses that use a (unstandardised) mean difference measure, which make up 76% of continuous outcome CDSR meta-analyses. Viechtbauer [124] conducted a simulation study of both standardised and unstandardised mean difference meta-analyses and found properties were reasonably consistent between the two measures.

Findings in this thesis may suggest how heterogeneity variance estimators perform in meta-analyses of other types of data. For time-to-event outcomes, study results may be expressed in terms of hazard ratios, which can be interpreted as the relative risk of an event occurring per unit of time [112, 118]. Therefore, they share many of the same properties of relative risks in non-time-to-event data [65]. Standard errors of hazard ratios are large when few events are observed and therefore, the size of the study is correlated with the hazard ratio. It is not possible to identify the number of meta-analyses in the CDSR dataset with time-to-event outcomes, however, Davey et al. [21] suggests the proportion could be up to 4%. It is possible that the issues identified for odds ratio meta-analyses with rare events, as observed in my simulation study, are also present for meta-analyses of hazard ratios. A new simulation study would be required to confirm this.

Methods for meta-analyses of diagnostic accuracy studies are more diverse because test performance depends on the defined threshold value [23, 112]. However, for a given threshold, study results can be presented as a binary 2x2 contingency table that includes the number of true and false-positives and negatives [23]. Likelihood ratios or diagnostic odds ratios can be derived from these tables. Findings in this thesis from binary outcome meta-analyses may be applied to meta-analyses of these summary statistics. Studies can be summarised in other ways, such as sensitivity,

specificity, or through the whole Receiver Operating Curve (ROC), in which properties of the heterogeneity variance are likely to be different. The number of diagnostic accuracy meta-analyses in the CDSR dataset is likely to be small [70]. No simulation studies that compare heterogeneity variance estimators in time-to-event or diagnostic accuracy meta-analyses were identified in the systematic review in chapter 5.

I compared methods to estimate the heterogeneity variance in meta-analyses of aggregate data throughout this thesis. Individual participant data (IPD) can also be combined in a meta-analysis in one or two stages. A two-stage approach calculates study-level aggregate data from IPD, so my results can trivially be applied in this setting. Other IPD meta-analyses use a one-stage approach [104], which involves multi-level modelling and the calculating the heterogeneity variance simultaneously with all other parameters in the model. This may be preferred over the aggregate two-stage approach because it allows subject-level covariates to be added into the model and a more thorough investigation into the causes of heterogeneity [105]. Of the heterogeneity variance estimators mentioned in this thesis, only the maximum likelihood, REML and Bayesian methods can be applied in this setting. Methods are also available to combine study-level 2x2 contingency tables in binary outcome meta-analyses [103, 121], which generally use REML methods for heterogeneity variance estimation. This approach may lead to improved estimates of the heterogeneity variance in meta-analyses with sparse data, but there currently been little simulation research in this area.

I introduced a number of Bayesian approaches to heterogeneity variance estimation in chapter 2. Those that require a subjective prior distribution were not compared in further chapters because of difficulties in defining them in simulated data and empirical meta-analysis data out of context. Bayesian methods naturally avoid zero heterogeneity variance estimates and may also increase precision in meta-analyses with few studies, which constitute most meta-analyses in Cochrane reviews. Turner et al. [120] and Rhodes et al. [88] define informed prior distributions for binary and continuous outcome meta-analyses respectively. These priors are based on previous

meta-analyses from the CDSR dataset and defined separately for each disease area. A full Bayesian approach is likely to lead to improved estimates of the heterogeneity variance when reliable and informative priors are available. However, this is not always the case, particularly in disease areas with few previous meta-analyses [88].

Random-effects meta-analysis, and most heterogeneity variance methods, are built on the assumption of normally distributed effects [48]. However, my simulation results, and those from Kontopantelis et al. [64], show heterogeneity variance methods are robust to all but the most extreme distributions of study effects. Publication bias is potentially more of an issue for heterogeneity variance estimation. In meta-analyses with publication bias, the size of the study effects are correlated with study size. This issue was deemed beyond the scope of this thesis. Assessing the properties of heterogeneity variance methods in simulated meta-analyses with publication bias is problematic. Studies could be systematically excluded to simulate publication bias but this would not preserve the parameters of the underlying distribution. Methods will inevitably perform poorly in the presence of significant publication bias, but this is understandable.

A continuity correction of 0.5 was applied to all binary outcome meta-analyses with zero events in my simulation study. This correction factor was chosen because it's widely used and the default method in the software Revman [87]. Other methods for dealing with zero events are available [10, 30]. In particular, a one-stage logistic regression modelling approach has been shown to produce less biased odds ratio estimates than the methods I used [113]. The decision to use this correction factor may have affected results in scenarios with rare outcomes. However, it is unlikely that using a different correction factor would have affected conclusions, particularly in scenarios with extremely rare events where all heterogeneity variance estimates had considerable negative bias.

Table 9.1 summarises the heterogeneity variance estimation methods available in the main statistical software packages at the time of writing. I include the four estimators that have the most reasonable properties (DerSimonian-Laird, two-step

DerSimonian-Laird, Paule-Mandel and REML) and full Bayes. WinBUGS [87] is the only software in which the DerSimonian-Laird estimator is not available; this is only because WinBUGS is software that specialises in Bayesian methods. DerSimonian-Laird is the only available estimator in Revman [87] and is the software used to conduct all Cochrane systematic reviews. The two-step DerSimonian-Laird estimator (PM_{DL}) is not readily available in any statistical software. The packages *meta* [99] and *metafor* [126] in R [85] can produce PM_{DL} estimates only by restricting the Paule-Mandel iterative process to two steps. PM_{DL} is not available in any software package because it is widely considered as a simplified version of Paule-Mandel, and therefore assumed to have inferior properties.

9.4 Further work

I have identified several limitations that came to light during the conduct of my research and discussed them in the last section. They were not addressed in this thesis mainly because of time and length constraints and were arguably inevitable given the scale of the problem of heterogeneity variance estimation in meta-analysis. I now suggest potential areas for further research to address many of these limitations.

9.4.1 Logistic regression models for meta-analysis

I discussed logistic regression methods in the last section that can be used to combine 2x2 contingency table data in binary outcome meta-analyses [103, 121]. This method makes full use of study data that is often readily available from these study's published results. There are a limited number of heterogeneity variance estimation methods available for use in combination with this method, but one of which is REML, which I showed has reasonable properties in aggregate data meta-analyses. A simulation study would be of benefit to compare the properties heterogeneity

	License Type	DL	PM_{DL}	PM	REML	FB
RevMan [87]	Freeware	✓	-	-	-	-
R [85]	Freeware	✓ *(meta, metafor)*	*	✓ *(meta, metafor)*	✓ *(meta, metafor)*	✓ *(R2WinBUGS, BRugs, rjugs)*
SAS [97]	Commercial	✓ *(marandom.sas)*	-	-	✓ *(PROC IML, PROC MIXED, PROC GLIMMIX)*	✓ *(SASBUGS, RASmacro, PROC MCMC)*
SPSS [54]	Commercial	✓ *(meanes.sps, metaf.sps, metareg.sps)*	-	-	-	-
Stata [108]	Commercial	✓ *(metareg, metan, metaan, mvmeta)*	-	✓ *(metareg)*	✓ *(metareg, metaan, mvmeta)*	-
WinBUGS [72]	Freeware	-	-	-	-	✓

Table 9.1: Heterogeneity variance estimators available in popular statistics software

**PM_{DL} not a readily available method in packages metafor and meta, but functions can be adapted.*

The table is taken and condensed from a recent published review of heterogeneity variance methods, which I co-authored [122].

variance estimates derived from this method (using REML) and aggregate data estimation methods (using REML and the two-step DerSimonian-Laird heterogeneity variance estimators).

9.4.2 Distributions of study size

In the simulation study in chapters 6 - 8, meta-analyses were generated with study sizes derived from five different distributions representing a variety of sizes and also a wide variety of differences in study size. These could not be considered a comprehensive selection of distributions, but had a substantial impact on the the properties of estimators. A new simulation study would be of benefit for further exploration in a wider variety of distributions. Distributions that are yet to be explored and may reveal interesting results include; (1) few small studies and many large studies, (2) many large studies and few small studies, (3) uniformly distributed from small to large, (4) negatively skewed (producing more large studies than small), and (5) positively skewed (producing more small studies than large).

9.4.3 Wider strategies for heterogeneity variance estimation in *problem* meta-analyses

I showed in the last chapter that heterogeneity variance estimates are usually imprecise and biased in meta-analyses in practice. Therefore, we can rarely rely on a single estimate when making inference on the degree of heterogeneity and a wider strategy in these scenarios is required. I recommend sensitivity analysis in these meta-analyses, but further research may be required to investigate the potential impact of different sensitivity analysis strategies.

9.4.4 Confidence intervals for the heterogeneity variance

Finally, it is imperative that confidence intervals for the heterogeneity variance are reported as standard in meta-analyses to express the uncertainty around their estimates. Many confidence intervals are available for this purpose [122], so a systematic review may be required to find if there is consensus over which confidence interval method has the best properties. Recommendations from this research would form part of the wider strategy for addressing the issue imprecise estimates, as I mentioned above.

9.4.5 Implementation in statistical software

Finally, to encourage the use of the two-step DerSimonian-Laird estimator in future meta-analyses, its code must be implemented into statistical software packages and ideally be the default option. My recommendation differs from the recommendations of other comprehensive simulation studies [78, 79, 124], and an editorial letter to Cochrane [123]; these recommend the iterative Paule-Mandel estimator and/or REML for use in practice. Therefore, it may take time for the dissemination of my research to impact on meta-analysis methods in statistical software.

9.5 Conclusion

DerSimonian-Laird is the most commonly used method to estimate the heterogeneity variance in meta-analysis, and produces negatively biased estimates in meta-analyses of binary data with rare events and/or meta-analyses containing only small studies. The Paule-Mandel estimator produces estimates with negative bias only in extreme cases where all meta-analysis method fail, and where conducting the meta-analysis at all is questionable. However, Paule-Mandel produces estimates with a higher positive bias than DerSimonian-Laird in meta-analyses with moderate to large differences in

study size. The two-step DerSimonian-Laird estimator and REML are shown in my simulation study to have the best properties of both these methods. I recommend REML given that it is already widely known and available in most statistical software packages. I recommend the Hartung-Knapp confidence interval for the summary effect and advise caution when making inference on this only in binary outcome meta-analyses with rare events.

More importantly, heterogeneity variance estimates derived from any method in a two-stage meta-analysis are usually imprecise and either negatively or positively biased in practice. I recommend the reporting of confidence intervals for the heterogeneity variance estimate and I^2. Recent studies have found these confidence intervals are rarely reported in practice [56], which can mislead researchers into thinking the level of heterogeneity is known. I recommend sensitivity analyses, particularly if the researcher believes conclusions could change solely based on a change is the level of heterogeneity present. Sensitivity analysis are rarely required for inference on the the summary effect alone if the Hartung-Knapp confidence interval method is used.

My thesis has demonstrated many of the problems inherent in estimating heterogeneity in a meta-analysis. The DerSimonian-Laird approach has been criticised in the past, and I recommend that the REML or two-step DerSimonian-Laird estimators be used instead. The use of the Hartung-Knapp confidence interval could also provide a more realistic interpretation of uncertainly of the summary effect in heterogeneous meta-analyses. None of these methods are perfect however, and caution should always be exercised when estimating heterogeneity or I^2, particularly when there are few studies or events are rare. In such circumstances comparing several estimates of heterogeneity may be useful. This work highlights the fact that our response to heterogeneity should not begin and end with performing a single random effects analysis; we should always seek to investigate potential causes of any identified heterogeneity.

Recommendations:

- There is no method likely to produce an accurate estimate of the heterogeneity variance in most two-stage meta-analyses in Cochrane reviews. Therefore, a confidence interval for this estimate should always be reported.
- REML heterogeneity variance estimates generally have the most reasonable properties, so this method is recommended. Two-step DerSimonian-Laird is a good alternative when iteration for REML fails to converge.
- The Hartung-Knapp confidence interval method for the summary effect is recommended over Wald-type methods.

Appendix A: Supplementary material from chapter 2

A.1 Search strategy

The following search criteria was designed to identify methods to estimate the heterogeneity variance parameter and its confidence interval in Veroniki et al. [122]. My thesis does not compare confidence interval methods for this parameter, so papers that relate solely to these methods were excluded from my review. PubMed was searched to identify research articles and references of each article were scanned for additional relevant literature. The following search criteria was used:

> ((heterogen*[Title/Abstract]) OR (*consisten*[Title/Abstract]) OR (between - study variance*[Title/Abstract]) OR (between - trial variance*[Title/Abstract])) AND (meta - analys*[Title/Abstract]) AND ((random effect*[Title/Abstract]) OR (mixed effect*[Title/Abstract]) OR (meta - regress*[Title/Abstract])) AND ((distribution) OR (prior) OR (prediction) OR (estimat*) OR (overall treatment effect*) OR (summary treatment effect*) OR (pooled effect*) OR (confidence interval*) OR (bias*) OR (error*) OR (power) OR (simulation*) OR (coverage probability*) OR (mean square* AND error*))

Appendix B: Supplementary material from chapter 4

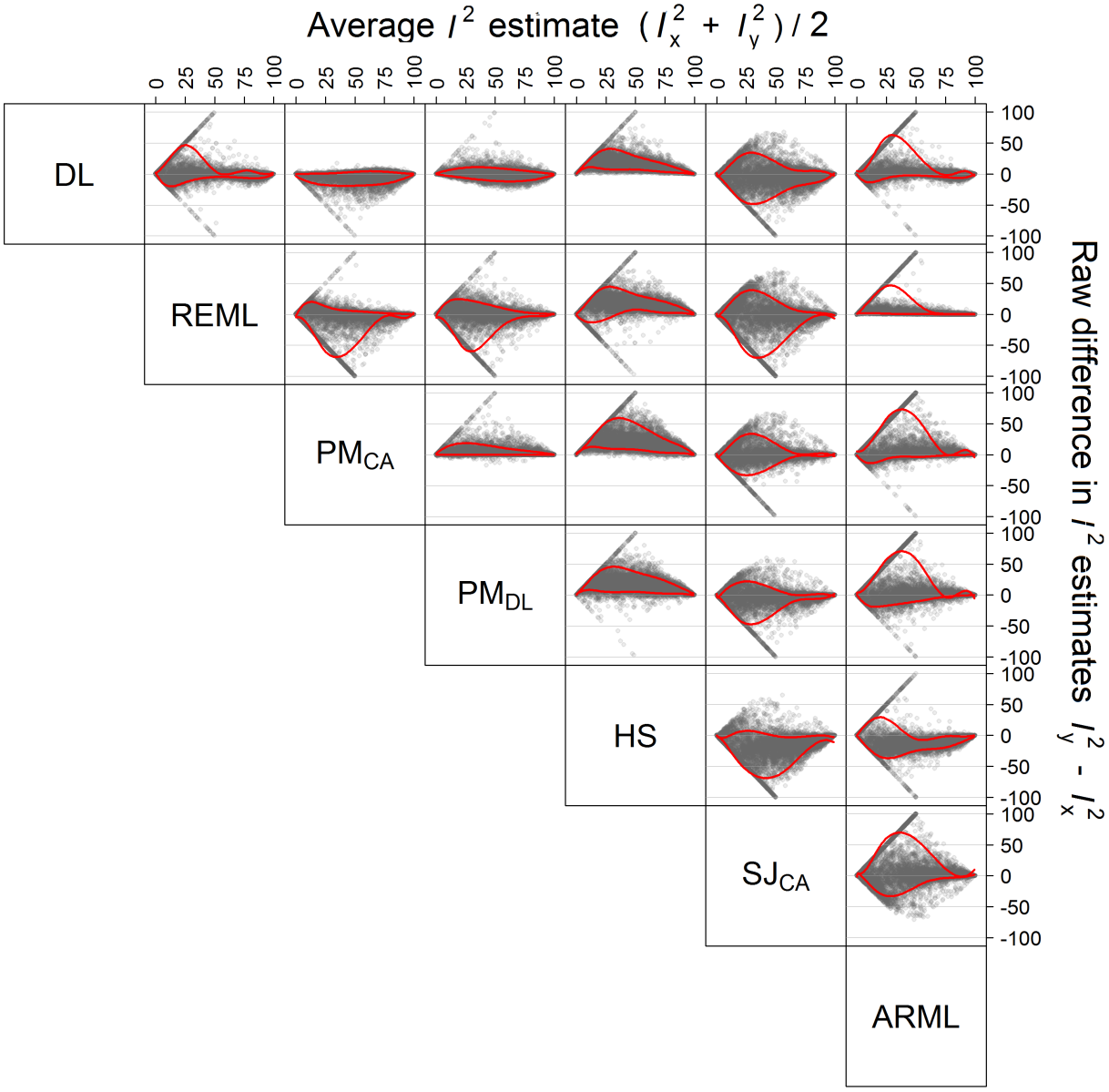


Figure B.1: Bland-Altman scatter plots comparing I^2 estimates from different heterogeneity variance methods excluded from the main results

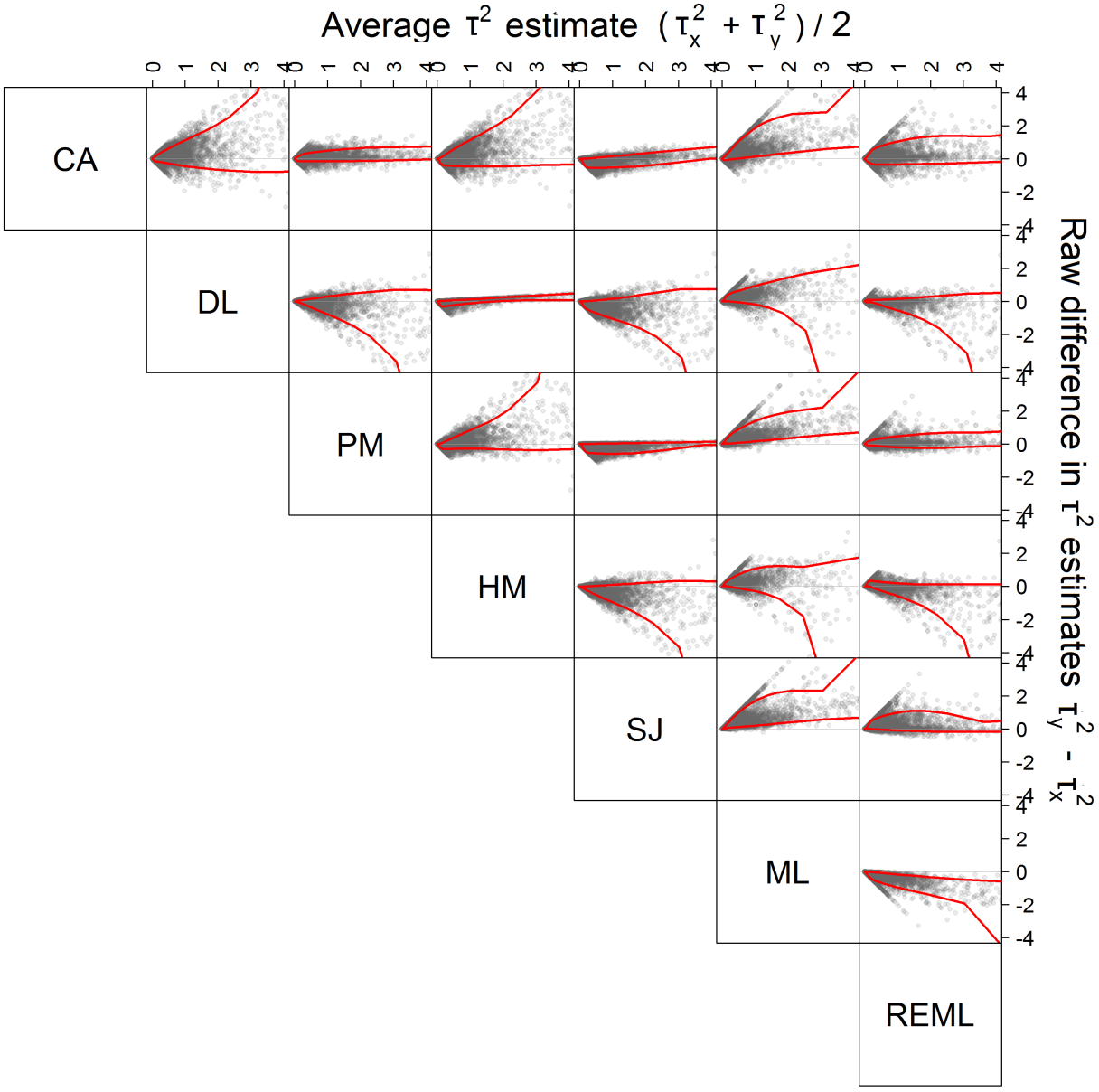


Figure B.2: Bland-Altman scatter plots comparing τ^2 estimates from different heterogeneity variance methods.

Appendix C: Supplementary material from chapter 5

C.1 Search strategy

We searched MEDLINE and the Web of Science core collection using the following strategy:

1. meta-* OR heterogen* [Title]
2. random effect* OR random-effect* OR mixed effect* OR meta-regress* OR sequential meta-analys* [Title/Abstract]
3. compar* OR simulat* OR mean square* OR bias* OR estimat* [Title/Abstract]
4. between-trial OR (between-study OR heterogen* OR *consisten* OR DerSimonian* [Title/Abstract]
5. cancer OR stroke OR blood OR arthritis OR alcohol OR depress* OR infect* OR diabetes OR disease* OR illness* OR surviv* OR smok* OR risk OR vitamin* OR therapy OR surgery [Title]
6. (#1 AND #2 AND #3 AND #4) NOT #5

To reduce the number of applied meta-analyses from the search results, publications with selected common medical terms in the title were excluded (#5). The search strategy needed to be adapted for JSTOR for two reasons: (1) There is a limit to the number of search terms that can be used and (2) Only 10% of publications in the database include abstracts and so "Title/Abstract" terms were searched within the full text instead. The JSTOR search strategy was as follows:

1. meta anal* [Title]
2. random effect* [Full Text]
3. compar* OR simulat* OR mean square* OR bias* [Full Text]

4. heterogen* [Full Text]

5. #1 AND #2 AND #3 AND #4

C.2 Performance measures in simulated data

Measure	Notation	Details
Measures relating to τ^2		
Bias	$bias\left(\hat{\tau}^2\right)$	$E\left[\hat{\tau}^2\right] - \tau^2$
MSE	$MSE\left(\hat{\tau}^2\right)$	$E\left[\left(\hat{\tau}^2 - \tau^2\right)^2\right]$
Variance	$Var\left(\hat{\tau}^2\right)$	$E\left[\left(\hat{\tau}^2 - E\left[\hat{\tau}^2\right]\right)^2\right]$
Efficiency	$e\left(\hat{\tau}^2\right)$	$\frac{I_F^{-1}-}{var\left(\hat{\tau}^2\right)}$ where $I_F^{-1}-$ is Fishers information: $I_F^{-1}- = 2\left[\sum_{i=1}^{k} \frac{1}{(\tau^2 + \sigma_i^2)}\right]^{-1}$
Proportion of zero estimates	$P(\hat{\tau}^2 = 0)$	The proportion of zero estimates of $\hat{\tau}^2$
Measures relating to θ		
Bias	$bias\left(\hat{\theta}\right)$	$E\left[\hat{\theta}\right] - \theta$
MSE	$MSE\left(\hat{\theta}\right)$	$E\left[\left(\hat{\theta} - \theta\right)^2\right]$
Measures relating to confidence intervals for θ		
Coverage	$cov\left(CI_x\right)$	The proportion of confidence intervals for $\hat{x}$ that contain x
Error interval estimation	$EI\left(CI_x\right)$	A ratio between the observed and the true confidence interval widths. For example, in the case when $x = \theta$, then $EI\left(CI_\theta\right) = \frac{upperCI_\theta - lowerCI_\theta}{3.92\sqrt{\left(\sum_{i=1}^{k}\left(\tau^2 + \hat{\sigma}_i^2\right)^{-1}\right)^{-1}}}$

Table C.1: Performance measures in simulated data

Appendix D: Supplementary material from chapter 6

D.1 Tables of τ^2 parameter values

Outcome measure	Study sizes	Event prob	$I^2 = 15\%$	$I^2 = 30\%$	$I^2 = 45\%$	$I^2 = 60\%$	$I^2 = 75\%$	$I^2 = 90\%$	$I^2 = 95\%$
			τ^2						
OR	small	0.5	0.067	0.178	0.344	0.633	1.33	4.5	15.6
	small-to-medium	0.5	0.0144	0.0333	0.0655	0.122	0.244	0.78	1.67
	medium	0.5	0.0067	0.0174	0.0333	0.056	0.122	0.367	0.78
	small and large	0.5	0.0025	0.0066	0.0144	0.023	0.0756	0.356	0.78
	large	0.5	0.0001	0.0023	0.00456	0.0082	0.0166	0.045	0.01
	small	0.1 to 0.5	0.0944	0.233	0.445	0.856	1.89	20	-
	small-to-medium	0.1 to 0.5	0.0178	0.0433	0.0855	0.1545	0.322	1.11	2.3
	medium	0.1 to 0.5	0.0089	0.0233	0.0433	0.078	0.156	0.45	1.11
	small and large	0.1 to 0.5	0.0036	0.0084	0.0178	0.0356	0.0945	0.456	1.22
	large	0.1 to 0.5	0.0012	0.0023	0.00589	0.0107	0.0222	0.0645	0.134
	small	0.05	0.422	1.156	2.56	7.56	-	-	-
	small-to-medium	0.05	0.0755	0.189	0.378	0.745	1.78	-	-
	medium	0.05	0.034	0.0967	0.189	0.356	0.756	3.44	-
	small and large	0.05	0.0144	0.0345	0.0745	0.167	0.433	2.3	-
	large	0.05	0.0053	0.0133	0.0255	0.0445	0.089	0.23	0.56
	small	0.01	2.78	14.5	-	-	-	-	-
	small-to-medium	0.01	0.378	1.11	2.45	6.7	-	-	-
	medium	0.01	0.12	0.45	1.067	2.44	7.8	-	-
	small and large	0.01	0.0656	0.178	0.34	0.1	3.67	-	-
	large	0.01	0.0245	0.0622	0.122	0.233	0.478	1.78	-
SMD	small	-	0.0178	0.0444	0.0845	0.156	0.322	0.1	2.44
	small-to-medium	-	0.0035	0.00856	0.0156	0.023	0.056	0.12	0.34
	medium	-	0.0018	0.00444	0.00844	0.01545	0.0311	0.089	0.12
	small and large	-	0.0007	0.00156	0.00344	0.00744	0.0189	0.089	0.12
	large	-	0.0002	0.0006	0.0011	0.0021	0.0042	0.0133	0.0256

Table D.1: τ^2 parameter values for each simulated scenario.
τ^2consistent between numbers of studies and distributions of study effects. $I^2 = 0\%$ is not included in this table, because it always corresponds to $\tau^2 = 0$. Parts of the table marked with a dash are where there is no such τ^2 that produces meta-analyses with the given mean I^2

Outcome measure	Study sizes	Event prob	$I^2 = 15\%$	$I^2 = 30\%$	$I^2 = 45\%$	$I^2 = 60\%$	$I^2 = 75\%$	$I^2 = 90\%$	$I^2 = 195\%$
			90% reference range of underlying I^2						
OR	small	0.5	14 - 15	30 - 31	44 - 46	59 - 61	73 - 76	88 - 92	91 - 97
	small-to-medium	0.5	10 - 20	20 - 36	33 - 53	48 - 67	64 - 81	86 - 93	92 - 96
	medium	0.5	15 - 15	30 - 30	45 - 45	59 - 60	74 - 75	90 - 90	95 - 95
	small and large	0.5	1 - 20	3 - 40	6 - 58	13 - 75	27 - 88	63 - 97	78 - 99
	large	0.5	14 - 17	27 - 32	42 - 48	57 - 63	72 - 77	89 - 91	94 - 95
	small	0.1 to 0.5	12 - 17	25 - 34	40 - 49	54 - 65	69 - 80	60 - 97	-
	small-to-medium	0.1 to 0.5	9 - 20	20 - 36	32 - 54	48 - 68	62 - 81	84 - 93	92 - 97
	medium	0.1 to 0.5	12 - 17	26 - 34	40 - 49	55 - 64	70 - 78	88 - 92	93 - 96
	small and large	0.1 to 0.5	1 - 22	3 - 40	6 - 59	10 - 74	24 - 88	59 - 97	78 - 99
	large	0.1 to 0.5	12 - 18	24 - 34	39 - 50	53 - 65	70 - 79	87 - 92	94 - 96
	small	0.05	13 - 16	25 - 34	35 - 52	28 - 72	-	-	-
	small-to-medium	0.05	10 - 19	19 - 36	34 - 53	47 - 68	60 - 82	-	-
	medium	0.05	15 - 16	29 - 31	43 - 47	58 - 62	71 - 77	84 - 93	-
	small and large	0.05	1 - 20	3 - 39	6 - 57	12 - 75	25 - 88	58 - 98	-
	large	0.05	13 - 17	27 - 33	42 - 49	57 - 63	72 - 77	89 - 92	93 - 96
	small	0.01	10 - 19	3 - 45	-	-	-	-	-
	small-to-medium	0.01	9 - 19	18 - 38	28 - 55	30 - 74	-	-	-
	medium	0.01	14 - 16	27 - 32	38 - 50	47 - 67	47 - 84	-	-
	small and large	0.01	1 - 20	3 - 40	6 - 58	12 - 77	19 - 91	-	-
	large	0.01	13 - 16	27 - 33	41 - 48	56 - 63	70 - 78	86 - 92	-
SMD	small	-	15 - 15	29 - 30	44 - 46	60 - 61	74 - 76	89 - 91	94 - 96
	small-to-medium	-	9 - 19	20 - 36	33 - 53	46 - 67	64 - 80	85 - 93	92 - 96
	medium	-	15 - 15	30 - 30	45 - 45	60 - 60	75 - 75	89 - 90	95 - 95
	small and large	-	1 - 21	3 - 39	6 - 57	12 - 74	26 - 88	63 - 97	79 - 99
	large	-	13 - 16	27 - 33	42 - 48	57 - 63	73 - 77	89 - 91	94 - 95

Table D.2: 90% reference range of underlying I^2 values for each simulated scenario.

τ^2consistent between numbers of studies and distributions of study effects. $I^2 = 0\%$ is not included in this table, because it corresponds only to $\tau^2 = 0$. Parts of the table marked with a dash are where there is no such τ^2 that produces meta-analyses with the given mean I^2

D.2 R code for all heterogeneity variance estimators

```
#########################################################################
# R code for calculating all heterogeneity estimates in one function  #
#########################################################################

########################################
#list of arguments and their meanings #
########################################
#xi - vector of effect estimates for each study.  If the outcome is
# odds ratio (for example), we assume that xi is already converted to
# log odds-ratios. log argument can be used to convert output back onto
# the original scale after all heterogeneity estimates have been
# calculated
#
#sei - vector of standard errors for each study.
#
#hetest - vector of heterogeneity estimators that you would like to be
# calculated.  The default is NULL, which means all estimates are
# calculated.
#
#signiftau2 - number of significant figures to round tau2 estimates
# (inc confidence intervals)
#
#maxit - maximum number of iterations allowed where the process of
# estimating tau2 involved iteration
#
#output - TRUE if output is displayed, FALSE otherwise (stops too much
# output when we are running the program iteratively)
#
#tau2.0 - starting value of iterative estimators
#
###PARAMETERS SPECIFIC TO AB...note that 2 out of 3 are required to
# calculate the estimate:
#eta - shape parameter of the prior distribution
#lambda - spread parameter of the prior distribution
#tau2prior - prior estimate of heterogeneity
#
###PARAMETERS SPECIFIC TO HS2
#Ntot - total sample size in meta-analysis (over both treatment groups)
#
###PARAMETERS SPECIFIC TO IPM (nci and nti are also used for the MBH
###estimator)
#nci - sample size of the control group
#nti - sample size of the treatment group
```

```
#eci - number of events in the control group
#
###PARAMETERS SPECIFIC TO DLp
#DLpos - truncation value as an alternative to zero with the original
# DL estimator
#
###PARAMETERS SPECIFIC TO DLb
#bsamp - number of bootstrap samples
#
#PARAMETERS SPECIFIC TO MBH
#corrbias - corrects xi for bias if TRUE, only functional for MBH
# estimator (using the method of Malzahn et al 2000)
#
#######################################
#list of estiamtors and their acornyms#
#######################################
#####General MoM approaches
#CA    - Cochran's ANOVA
#DL    - DerSimonian-Laird
#PM    - Paule Mandel
#CA2   - Two step PM with CA initial estimate
#DL2   - Two step PM with DL initial estimate
#IPM   - Improved Paule-Mandel (binary outcome data only) - uses arguments
#        eci, nci and nti
#DLp   - Positive DerSimonian-Laird estimate, with truncation at 0.01
#DLb   - bootstrap version of DerSimonian-Laird

####Other approaches
#HM    - Hartung Makambi
#HS    - Hunter Schmidt (original estimator using FE weightings)
#SJ    - Sidik Jonkman
#SJ2   - An improvement on Sidik Jonkman
#MBH   - Malzahn, Bohning and Holling (from original paper Malzahn 2000)

####Maximum Likelihood approaches
#ML    - Maximum Likelihood
#REML  - Restricted Maximum Likelihood
#ARML  - Approximate Restricted Maximum Likelihood

####Bayesian Approaches
#EB    - Empirical Bayes
#AB    - Approximate Bayes
#B0    - Rukhin Bayes with zero prior (with correction for sum(n))
#BP    - Rukhin Bayes with simple prior
```

```
hetest <- function(xi=lSS, sei=seSS, Ntot=NULL, nci=NULL, nti= NULL,
  eci=NULL, eta=NULL, lambda=NULL, tau2prior=NULL, DLpos=0.01,
  bsamp=5000, SMD=FALSE, hetest=NULL, signiftau2=6, maxit=100,
  tau2.0=NULL, trunc=TRUE, output=TRUE) {

  #if no specific set of estimates is required, calculate them all...
  if (is.null(hetest)) hetest <- c("CA","DL","PM","IPM","CA2",
    "DL2","DLp","DLb","HM","HS","SJ","SJ2","MBH","ML","REML",
    "ARML","AB","B0","BP")

  #clear the variables that may have been defined previously when this
  #function was run so that we can start again fresh
  CA_est<-DL_est<-PM_est<-IPM_est<-CA2_est<-DL2_est<-as.numeric(NA);
  DLp_est<-DLb_est<-HM_est<-HS_est<-SJ_est<-SJ2_est<-as.numeric(NA);
  MBH_est<-ML_est<-REML_est<-ARML_est<-AB_est<-B0_est<-as.numeric(NA);
  B0K_est<-BP_est<-as.numeric(NA);

  #assume equal sample sizes in arms
  if (!is.null(Ntot) & is.null(nci) & is.null(nti)) {
    Ntot <- nci + nti
  }
  if (!is.null(Ntot) & is.null(nci) & is.null(nti)) {
    nci <-  nti <- round(Ntot/2,digits=0)
  }



  Kest <- length(hetest) # number of estimates to be calculated
  esti <- 1
  #^a counter so that we can create a dataset with a separate estimate
  #on each row the first specified estimate will be in row 1 ... etc.

  ####specifying all output vectors before replacing the values with
  ####actual estimates.
  name <- rep(NA,Kest)
  tau2 <- rep(NA,Kest)
  #theta not needed for output, just for the process of calculating some
  #of the tau2 estimates
  theta <- rep(NA,Kest)

  #bias correction if the meta-analysis has an SMD outcome measure
  #keep unadjusted SMDs for MBH estimator
  if (SMD) {
    xi_unadj <- xi
    sei_unadj <- sei
```

```
  J <- 1- (3/(4*(Ntot-2) - 1))
  xi <- xi*J
  sei <- sqrt((sei^2)*(J^2))
}

K <- length(xi) #K=number of studies in the meta-analysis
vi <- sei^2 #variance of each study
wFEi <- 1/vi
FEtheta <- sum(xi*wFEi)/sum(wFEi)

#DerSimonian Laird
if ('DL' %in% hetest) {
  name[esti] <- "DL"
  DLw <- 1/vi
  theta[esti] <- sum(xi*(DLw))/sum((DLw))
  DLtausq1 <- sum(DLw*((xi-theta[esti])^2)) - (sum(DLw*vi)) +
    (sum((DLw^2)*vi)/sum(DLw))
  DLtausq2 <- sum(DLw) - (sum(DLw^2)/sum(DLw))

  if (trunc) DL_est <- tau2[esti]<- max(0,DLtausq1/DLtausq2)
  else DL_est <- tau2[esti]<- DLtausq1/DLtausq2
  esti <- esti + 1
}

#DerSimonian Laird
if ('DLp' %in% hetest) {
  name[esti] <- "DLp"
  DLw <- 1/vi
  theta[esti] <- sum(xi*(DLw))/sum((DLw))
  DLtausq1 <- sum(DLw*((xi-theta[esti])^2)) - (sum(DLw*vi)) +
    (sum((DLw^2)*vi)/sum(DLw))
  DLtausq2 <- sum(DLw) - (sum(DLw^2)/sum(DLw))

  if (trunc) DLp_est <- tau2[esti]<- max(DLpos,DLtausq1/DLtausq2)
  else DLp_est <- tau2[esti]<- DLtausq1/DLtausq2
  esti <- esti + 1
}

#DerSimonian Laird
if ('DLb' %in% hetest) {
  name[esti] <- "DLb"
  DLw <- 1/vi
  theta[esti] <- sum(xi*(DLw))/sum((DLw))

  #number of possible combinations given the number of studies
```

```
  #if K=9 then there are 24310 possible combinations, much more
  #samples than would ever be needed. so no need to calculate perm
  if (K<9) perm <- factorial(2*K - 1) / (factorial(K)*factorial(K-1))
  else perm <- bsamp-1
  #if the number of combinations is small, then no need to do all
  #bootstraps, just take complete sample...
  if (perm<=bsamp) {
    comb_DLb <- combinations(n=K, r=K, repeats.allowed=TRUE)
    no_samples <- perm
  }
  #if the number of combinations is large then just do a sample...
  else {
    comb_DLb <- t(replicate(bsamp,sample(1:K,K,replace = TRUE)))
    no_samples <- bsamp
  }

  DLb_est2 <- rep(NA,times=no_samples)

  for (i in 1:no_samples) {
    studycomb <- comb_DLb[i,]
    theta_b <- sum(xi[studycomb]*(DLw[studycomb]))/
      sum((DLw[studycomb]))
    DLtausq1_b <- sum(DLw[studycomb]*((xi[studycomb]-theta_b)^2)) -
      (sum(DLw[studycomb]*vi[studycomb])) +
      (sum((DLw[studycomb]^2)*vi[studycomb])/sum(DLw[studycomb]))
    DLtausq2_b <- sum(DLw[studycomb]) - (sum(DLw[studycomb]^2)/
      sum(DLw[studycomb]))
    if (trunc) DLb_est2[i] <- max(0,DLtausq1_b/DLtausq2_b)
    else DLb_est2[i] <- DLtausq1_b/DLtausq2_b
  }

  DLb_est <- tau2[esti]<- mean(DLb_est2)
  esti <- esti + 1

}

#Cochran ANOVA
if ('CA' %in% hetest | 'ML' %in% hetest | 'REML' %in% hetest) {
  #to calculate REML, we need a starting value of tau2_ML, or else
  #there may be more than 1 solution.
  CAw <- rep(1/K,times=K)
  theta[esti] <- sum(xi*CAw)/sum(CAw)
  CAtausq1 <- sum(CAw*((xi-theta[esti])^2)) - (sum(CAw*vi)) +
    (sum((CAw^2)*vi)/sum(CAw))
  CAtausq2 <- sum(CAw) - (sum(CAw^2)/sum(CAw))
```

```
  CA_est <- max(0,CAtausq1/CAtausq2)
  if ('CA' %in% hetest){
    if (trunc) tau2[esti]<- max(0,CAtausq1/CAtausq2)
    else tau2[esti]<- CAtausq1/CAtausq2
    name[esti] <- "CA"
    esti <- esti + 1
  }
}

#Paule Mandel
if ('PM' %in% hetest) {

  quant <- df <- K-1
  #degrees of freedom and expected mean under the fixed effects
  #assumption
  PMtau2out <- 1
  # just set an initial value for PM estimate for output

  if (is.null(tau2.0)) PMtausq <- 0 #initial estimate of tau2
  else PMtausq <- tau2.0
  PMit <- 1 #iteration number
  PM_F <- 1 #just to get the iteration started. F=0 => convergence

  while (PM_F!=0){

    #first calculate the the pooled effect based on present
    #estimate of tausq
    PMw = 1/(sei^2+PMtausq)
    PMyW = sum(xi*PMw)/sum(PMw)

    #equation comes from DerSimonian and Kacker 2007
    Q1 <- sum(PMw*(xi-PMyW)^2) #generalised Q statistic
    Q2 <- sum((PMw^2)*(xi-PMyW)^2) #denominator from delta
    #quant=statistic coming from the chisq dist regardless of data.
    #mean/CI bound etc
    if (trunc) PM_F <- max(Q1-quant,0)
    else PM_F <- Q1-quant
    delta <- PM_F/Q2 #what to add onto the next tausq estimate
    if (PM_F!=0) PMtausq <- PMtausq + delta

    PMit <- PMit + 1

    if (PM_F==0) {
      PMtau2out <-PMtausq
    }
```

```
    if (PMit==maxit) {
      PM_F<- 0
      if (output==TRUE)
        cat("PM estimator: Maximum Number of iterations reached
            without convergence\n")
    }
  }

  name[esti] <- "PM"
  if (PMit==maxit) PM_est <- tau2[esti] <- NA
  else PM_est <- tau2[esti] <- PMtau2out
  PMw <- 1/(vi + tau2[esti])
  theta[esti] <- sum(xi*PMw)/sum(PMw)
  esti <- esti + 1
}

#Paule Mandel (with improved standard errors)
if ('IPM' %in% hetest) {

  quant <- df <- K-1
  #degrees of freedom and expected mean under the
  #fixed effects assumption

  if (is.null(tau2.0)) IPMtausq <- 0 #initial estimate of tau2
  else IPMtausq <- tau2.0
  IPMdiff <- 1
  IPMit <- 1 #iteration number
  #counter for number of negative estimates
  negcount <- 0

  #calculations needed to calculate standard errors, but that don't
  #change for each iteration
  oddsc <- log(eci/(nci-eci)) # odds in control group
  thetaCA <- sum(xi)/K # un-weighted average


  while (IPMdiff!=0){

    IPMtausq_prev <- IPMtausq

    #first calculate the standard errors according to the alternative
    #formula proposed by Bhaumik (depends on tau2 estimate so needs to
    #be calculated for each iteration)
    sei_IPM <-
```

```
    ((exp(-oddsc - thetaCA + (IPMtausq/2)) + 2 +
    exp(oddsc + thetaCA + (IPMtausq/2)))/(nci + 1)) +
    ((exp(-oddsc) + 2 + exp(oddsc))/(nti + 1))

  #calculate the the pooled effect based on present estimate of
  #tausq
  IPMw = 1/(sei_IPM^2+IPMtausq)
  IPMyw = sum(xi*IPMw)/sum(IPMw)

  #equation comes from DerSimonian and Kacker 2007
  Q1 <- sum(IPMw*(xi-IPMyw)^2) #generalised Q statistic
  Q2 <- sum(IPMw*(sei_IPM^2)) - (sum((IPMw^2)*(sei_IPM^2)) /
    sum(IPMw))
  Q3 <- sum(IPMw) - (sum(IPMw^2) / sum(IPMw))
  IPMtausq <- (Q1 - Q2) / Q3

  if (trunc) {
    if(IPMtausq>=0) IPMdiff <- round(abs(IPMtausq - IPMtausq_prev),
      digits=signiftau2)
    else {
      negcount <- negcount + 1
      #if iteration is negative more than once then REML=0
      #final est
      if (negcount>=2) IPMdiff<-0
      IPMtausq<-0
    }
  }
  else IPMdiff <- round(abs(IPMtausq - IPMtausq_prev),
    digits=signiftau2)

  IPMit <- IPMit + 1

  if (IPMit==maxit) {
    IPMdiff<- 0
    if (output==TRUE)
      cat("IPM estimator: Maximum Number of iterations reached
        without convergence\n")
  }

}

name[esti] <- "IPM"
if (IPMit==maxit) IPM_est <- tau2[esti] <- NA
else IPM_est <- tau2[esti] <- IPMtausq
IPMw <- 1/(vi + tau2[esti])
```

```
  theta[esti] <- sum(xi*IPMw)/sum(IPMw)
  esti <- esti + 1
}

#Cochran ANOVA initial estimate with PM weightings
if ('CA2' %in% hetest) {
  name[esti] <- "CA2"
  if (trunc) CAtau2 <- max( 0 , (1/(K-1))*sum((xi-(sum(xi)/K))^2) -
    (1/K)*sum(vi) )
  else CAtau2 <- (1/(K-1))*sum((xi-(sum(xi)/K))^2) - (1/K)*sum(vi)
  CA2w <- 1/(CAtau2 + vi)
  theta[esti] <- sum(xi*CA2w)/sum(CA2w)
  CA2wtausq1 <- sum(CA2w*((xi-theta[esti])^2)) - (sum(CA2w*vi)) +
    (sum((CA2w^2)*vi)/sum(CA2w))
  CA2wtausq2 <- sum(CA2w) - (sum(CA2w^2)/sum(CA2w))
  if (trunc) CA2_est <- max(0,tau2[esti]<- CA2wtausq1/CA2wtausq2)
  else CA2_est <- tau2[esti]<- CA2wtausq1/CA2wtausq2
  esti <- esti + 1
}

#DerSimonian Laird initial estimate with PM weightings
if ('DL2' %in% hetest) {
  name[esti] <- "DL2"
  DLw <- 1/(vi)
  DLtheta <- sum(xi*DLw)/sum(DLw)
  if (trunc)  DLtau2 <- max( 0 , (sum(DLw*((xi-DLtheta)^2)) - K + 1)/
      ( sum(DLw) - (sum(DLw^2)/sum(DLw)) ) )
  else DLtau2 <- (sum(DLw*((xi-DLtheta)^2)) - K + 1) /
    ( sum(DLw) - (sum(DLw^2)/sum(DLw)) )
  DL2w <- 1/(DLtau2 + vi)
  theta[esti] <- sum(xi*DL2w)/sum(DL2w)
  DL2tausq1 <- sum(DL2w*((xi-theta[esti])^2)) - (sum(DL2w*vi)) +
    (sum((DL2w^2)*vi)/sum(DL2w))
  DL2tausq2 <- sum(DL2w) - (sum(DL2w^2)/sum(DL2w))
  if (trunc)  DL2_est <- max(0,tau2[esti]<- DL2tausq1/DL2tausq2)
  else DL2_est <- tau2[esti]<- DL2tausq1/DL2tausq2
  esti <- esti + 1
}

#Hartung Makambi
if ('HM' %in% hetest) {
  name[esti] <- "HM"
  HMQ<- sum((1/vi)*((xi-FEtheta)^2))
  HM_est <- tau2[esti] <- (HMQ^2) / ((2*(K-1)+HMQ)*(sum(1/vi)-
    (sum((1/vi)^2)/sum(1/vi))))
```

```
  esti <- esti + 1
}

#Hunter Schmidt (original estimator using FE weightings)
if ('HS' %in% hetest) {
  name[esti] <- "HS"
  if (trunc)  HS_est <- tau2[esti] <- max(0 ,
    (sum(wFEi*(xi - FEtheta)^2) - K) / (sum(wFEi)) )
  else   HS_est <- tau2[esti] <- (sum(wFEi*(xi - FEtheta)^2) - K) / (sum(wFEi))
  esti <- esti + 1
}

#Sidik Jonkman
if ('SJ' %in% hetest) {

  name[esti] <- "SJ"

  ####ESTIMATE OF TAU2

  #calculate the pooled estimate
  SJtheta_0 <- sum(xi)/K

  #Cochrans equally weighted estimate of the pooled result
  SJtau2_0 <- (1/K)*sum((xi - SJtheta_0)^2)

  #if all estimates are identical then we cannot go any further in
  #the calculation and our estimate is zero
  if (SJtau2_0>0) {
    #SJ weightings based on initial estimate of tau2 (SJtau2_0)
    SJw <- 1/( (vi/SJtau2_0)+1)
    #Random effects pooled estimate based on the above weightings
    SJtheta_1 <- sum(xi*SJw)/sum(SJw)
    SJ_est <- tau2[esti] <- (1/(K -1)) * sum( SJw * (xi-SJtheta_1)^2 )
  }

  else SJ_est <- tau2[esti] <- 0

  #pooled effect estimate
  SJw2 <- 1/( (vi/tau2[esti])+1)
  theta[esti] <- sum(SJw2 * xi) / sum(SJw2)

  esti <- esti + 1
}

#Improved Sidik Jonkman
```

```
if ('SJ2' %in% hetest) {

  name[esti] <- "SJ2"

  ####ESTIMATE OF TAU2

  #calculate the pooled estimate
  SJ2theta_0 <- sum(xi)/K

  #if all estimates are identical then tau2 is zero
  if (sum((xi - SJ2theta_0)^2)>0) {
    #variance components method (general form of hedges olkin)
    SJ2tau2_0 <- max( 0.01 , ((1/(K-1))*(sum((xi - SJ2theta_0)^2))) -
      ((1/K)*(sum(vi))) )

    #SJ2 weightings based on initial estimate of tau2 (SJ2tau2_0)
    SJ2w <- 1/( (vi/SJ2tau2_0)+1)
    #Random effects pooled estimate based on the above weightings
    SJtheta_1 <- sum(xi*SJ2w)/sum(SJ2w)

    SJ2_est <- tau2[esti] <- (1/(K -1)) *
      sum( SJ2w * (xi-SJtheta_1)^2 )
  }
  else SJ2_est <- tau2[esti] <- 0

  #pooled effect estimate
  SJ2w2 <- 1/( (vi/tau2[esti])+1)
  theta[esti] <- sum(SJ2w2 * xi) / sum(SJ2w2)

  esti <- esti + 1
}


#Malzahn, Bohning, Holling estimator (as given in the original
#Malzahn et al 2000 paper) only for SMD effects
if ('MBH' %in% hetest) {
  name[esti] <- "MBH"
  if (SMD) {
    Ni <- nci + nti - 2
    Hi <- sqrt(Ni/2)*(gamma((Ni/2)-0.5)/gamma(Ni/2))
    di <- xi/Hi
    thetaMBH <-  sum(di)/K #equal weighted mean effect
    Ki <- 1 - ((Hi)^2*((Ni-2)/Ni))
    MBH_est <- (sum((1-Ki)*((di-thetaMBH)^2))/(K-1)) -
      ((1/K)*sum((nci + nti)/(nci*nti))) -
```

```
      ((1/K)*sum(Ki*(di^2)))

    if (trunc) MBH_est <- tau2[esti]<- max(0,MBH_est)
    else tau2[esti]<- MBH_est
  } else MBH_est <- tau2[esti] <- NA
  esti <- esti + 1
}

#Maximum Likelihood
if ('ML' %in% hetest) {

  name[esti] <- "ML"

  #difference between this iteration and previous to assess when we
  #have convergence set MLdiff!=0 initially to get the process of
  #iteration going
  MLdiff <- 1

  #counter for number of iterations
  MLit<-0
  #counter for number of negative estimates
  negcount <- 0

  #first set initial estimate of tau2 and theta
  #(fixed effect estimates)
  if (is.null(tau2.0)) MLtau2 <- CA_est
  else MLtau2 <- tau2.0
  MLtheta <- sum(xi*wFEi)/sum(wFEi)

  while(MLdiff!=0){

    #estimate of between study heterogeneity
    MLtau2_prev <- MLtau2 #record of previous step
    if (-min(vi)>=MLtau2) MLtau2 <- -min(vi)+(10^-signiftau2)
    MLtau2 <- sum( ((xi-MLtheta)^2 - vi) / (vi+MLtau2)^2 ) /
      sum( 1 / (vi+MLtau2)^2 )
    #estimate for pooled effect
    MLtheta_prev <- MLtheta #record of previous step
    MLtheta <- sum( xi / (vi+MLtau2) ) / sum( 1 / (vi+MLtau2) )

    if (trunc) {
      if(MLtau2>=0) MLdiff <- round(abs(MLtau2 - MLtau2_prev),
        digits=signiftau2)
      else {
        negcount <- negcount + 1
```

```
        # if iteration is negative more than once then REML=0 final est
        if (negcount>=2) MLdiff<-0
        MLtau2<-0
        MLtheta<-sum(xi*wFEi)/sum(wFEi)
      }
    }
    else MLdiff <- round(abs(MLtau2 - MLtau2_prev),digits=signiftau2)

    MLit <- MLit + 1

    if (MLit==maxit) {
      MLdiff<- 0
      if (output==TRUE)
        cat("ML estimator: Maximum Number of iterations reached
            without convergence\n")
    }

  }

  if (MLit==maxit) ML_est <- tau2[esti] <- NA
  else ML_est <- tau2[esti] <- MLtau2

  #pooled effect estimate
  MLw2 <- 1/( (vi/tau2[esti])+1)
  theta[esti] <- sum(MLw2 * xi) / sum(MLw2)

  esti <- esti + 1

}

#Restricted Maximum Likelihood
if ('REML' %in% hetest) {

  name[esti] <- "REML"

  #first set initial estimate of tau2 and theta
  #(fixed effect estimates)
  if (is.null(tau2.0)) REMLtau2 <- CA_est
  else REMLtau2 <- tau2.0
  REMLtheta <- FEtheta


  #difference between this iteration and previous to assess when we
  #have convergence set diff!=0 initially to get the process of
  #iteration going
```

```
REMLdiff <- 1

#counter for number of iterations
REMLit <- 0
#counter for number of negative estimates
negcount <- 0

#process of iteration, stop when there is no difference between the
#last two steps
while(REMLdiff!=0){

  #estimate of between study heterogeneity
  REMLtau2_prev <- REMLtau2 #record of previous step
  tau2_p1 <- sum((1/((vi+REMLtau2_prev)^2))*(((xi-REMLtheta)^2)-vi))
  tau2_p2 <- sum(1/((vi+REMLtau2_prev)^2))
  tau2_p3 <- sum(1/(vi+REMLtau2_prev))
  REMLtau2 <- (tau2_p1/tau2_p2)+(1/tau2_p3)

  if (trunc) {
    if(REMLtau2>=0) REMLdiff <- round(abs(REMLtau2 - REMLtau2_prev),
      digits=signiftau2)
    else {
      negcount <- negcount + 1
      #if iteration is negative more than once then REML=0 final est
      if (negcount>=2) REMLdiff <- 0
      REMLtau2 <- 0
      REMLtheta <- FEtheta
    }
  }
  else REMLdiff <- round(abs(REMLtau2 - REMLtau2_prev),
    digits=signiftau2)

  REMLit <- REMLit + 1

  if (REMLit==maxit) {
    REMLdiff<- 0
    if (output==TRUE)
      cat("REML estimator: Maximum Number of iterations reached
          without convergence\n")
  }

  #this is just to update theta, rather than because this has
  #anything to do with convergence of this outcome
  #estimate for pooled effect
  REMLtheta_prev <- REMLtheta #record of previous step
```

```
    REMLtheta <- sum( xi / (vi+REMLtau2) ) / sum( 1 / (vi+REMLtau2) )

  }

  if (REMLit==maxit) REML_est <- tau2[esti] <- NA
  else REML_est <- tau2[esti] <- REMLtau2


  #pooled effect estimate
  theta[esti] <- REMLtheta

  esti <- esti + 1
}

#Approximate Restricted Maximum Likelihood
if ('ARML' %in% hetest) {
  name[esti] <- "ARML"

  #first set initial estimate of tau2 and theta
  #(fixed effect estimates)
  ARMLtau2 <- 0
  ARMLtheta <- FEtheta


  #difference between this iteration and previous to assess when we
  #have convergence
  #set diff!=0 initially to get the process of iteration going
  ARMLdiff <- 1

  #counter for number of iterations
  ARMLit<-0

  #process of iteration, stop when there is no difference between
  #the last two steps
  while(ARMLdiff!=0){

    #estimate of between study heterogeneity
    ARMLtau2_prev <- ARMLtau2 #record of previous step
    tau2_p1 <- sum(1/((vi+ARMLtau2_prev)^2))
    tau2_p2 <- sum( (1/((vi+ARMLtau2_prev)^2)) *
      ( ((K/(K-1)) * (xi-ARMLtheta)^2) - vi ) )
    ARMLtau2 <- tau2_p2/tau2_p1

    if (trunc) {
      if(ARMLtau2>=0) ARMLdiff <-
```

```
        round(abs(ARMLtau2 - ARMLtau2_prev),digits=signiftau2)
      else {
        ARMLdiff <- 0
        ARMLtau2 <- 0
        ARMLtheta <- FEtheta
      }
    }
    else ARMLdiff <- round(abs(ARMLtau2 - ARMLtau2_prev),
      digits=signiftau2)

    ARMLit <- ARMLit + 1

    if (ARMLit==maxit) {
      ARMLdiff<- 0
      cat("Maximum Number of iterations reached without
          convergence\n")
    }

    #this is just to update theta, rather than because this has
    #anything to do with convergence of this outcome
    #estimate for pooled effect
    ARMLtheta_prev <- ARMLtheta #record of previous step
    ARMLtheta <- sum( xi / (vi + ARMLtau2) ) /
      sum( 1 / (vi + ARMLtau2) )

  }

  ARML_est <- tau2[esti] <- ARMLtau2

  #pooled effect estimate
  theta[esti] <- ARMLtheta

  esti <- esti + 1
}


#Approximate Bayes
if ('AB' %in% hetest) {

  #check that 2 of the prior parameters are specified, otherwise
  #return an error
  countarg <- 0
  if (!is.numeric(lambda)) countarg <- countarg + 1
  if (!is.numeric(eta)) countarg <- countarg + 1
  if (!is.numeric(tau2prior)) countarg <- countarg + 1
```

```
  if (countarg==1) {

    name[esti] <- "AB"

    # set up output data
    require(rmeta)
    require(pscl)

    #calculate both eta and lambda parameters if they are not
    #both specified
    if (is.null(lambda)) lambda <- tau2prior*(eta-1)
    else if (is.null(eta)) eta <- (lambda/tau2prior)+1

    # Compute approximate Bayes estimate of heterogeneity variance
    # from DerSimonian-Laird estimate and prior distribution
    tau2DL <- meta.summaries(xi,sei,method="random")$tau2
    AB_est <- tau2[esti] <- tau2AB <-
      max((2 * lambda + K * tau2DL)/(2 * eta + K - 2), 0)
    esti <- esti + 1
  }
  else {cat("ERROR: AB estimate cannot be calculated as prior
    parameters have been specified incorrectly \n")}
}

#Rukhin (zero prior)
if ('B0' %in% hetest) {
  name[esti] <- "B0"
  #just assume fixed effects mean, this is the way Kontopantelis
  #did it also doesnt specify what n_i is given that the estimator
  #is proposed in the context where there isn't 2 treatment groups
  #per study, N=nci + nti as used by Kontopantelis
  #not exactly the same formula as in Rukhin 2012, because there
  #is a mistake, this is the corrected
  #formula similar to that used by Konto 2012
  B0theta <- theta[esti] <- FEtheta
  B0_est <- tau2[esti] <- ( sum((xi-B0theta)^2)/(K+1) ) -
    (( (sum(nci+nti)-K)*(K-1)*sum(vi) ) /
       ( K*(K+1)*sum(nci+nti-K+2) ))
  #this is possible if the denominator is zero (rare)...
  if (is.infinite(B0_est)) B0_est <- tau2[esti] <- NA
  esti <- esti + 1
}

#Rukhin (simple prior)
```

```
  if ('BP' %in% hetest) {
    name[esti] <- "BP"
    #just assume fixed effects mean for now, paper doesn't specify
    BPtheta <- theta[esti] <- FEtheta
    BP_est <- tau2[esti] <- sum((xi-BPtheta)^2)/(K+1)
    esti <- esti + 1
  }
  #Bayes Modal estimator
  if ('BM' %in% hetest) {
  }

  ####data frame for reporting all output
  #first round off the estimate to specified number of decimal
  #places by signiftau2 arguament
  tau2 <- signif(tau2, digits=signiftau2)
  out <-data.frame(name,tau2)
  if (output==TRUE) print(out)

  ####output that can be used after function has been run
  #create an output frame that can be used when iterating through this
  #function multipe times
  #the above dataframe is better when only calculating estimates for
  #one meta-analysis
  res <- list(CA_est,DL_est,PM_est,IPM_est,CA2_est,DL2_est,DLp_est,
    DLb_est,HM_est,HS_est,SJ_est,SJ2_est,MBH_est,ML_est,REML_est,
    ARML_est,AB_est,B0_est,BP_est)
  names(res) <- c("CA","DL","PM","IPM","CA2","DL2","DLp","DLb","HM",
    "HS","SJ","SJ2","MBH","ML","REML","ARML","AB","B0","BP")
  return(res)
  #we can refer to the estimates outside of this function by <funct name>$<est name>

}
```

D.3 R code for confidence interval methods of the summary effect

```
#########################################################################
#     R code for calculating all confidence interval estimates        #
#########################################################################

########################################
#list of arguments and their meanings #
########################################
```

```
#
#hetests - vector of heterogeneity estimates
#xi - effect estimates of the studies in the meta-analysis
#sei - standard errors of the effect estimates
#CIest - names of the confidence intervals to be calculated
#signif - significance level set for the confidence intervals,
#   the default is 0.05 (i.e. 95% CI)
#output - if TRUE then output results into the R console
#signifCI - numer of significant figues to round off the CI bounds
#hetnames - names of the heterogeneity estimators (corresponding to
#   the heterogeneity estimates in hetests argument)
#
########################################
#list of estiamtors and their acornyms#
########################################
#####so far just includes the CI methods that are used in the
#####simulation study
#
#Z   - Z-type confidence interval
#T   - t-distribution confidence interval
#HK  - Hartung-Knapp confidence interval

#make sure I keep this update and consistent with the estimators
#available in the het_est.R program

CIests <- function(hetests, xi, sei, CIest=c("Z","T","HK"),
  signif=0.05, output=TRUE, signifCI=4, hetnames=c("CA","DL","PM",
  "IPM","CA2","DL2","DLp","DLb","HM","HS","SJ","SJ2","MBH","ML",
  "REML","ARML","AB","B0","BP")) {

  if (length(hetests)!=length(hetnames))
    stop("Number of tau2 estimator names doesn't match the number
         of estimates given")

  #calculate mean effects
  thetaests <- rep(NA,times=length(hetests))
  for(i in 1:length(hetests)) {
    thetaests[i] <- sum(xi*(1/(sei^2 + hetests[i])))/
      sum(1/(sei^2 + hetests[i]))
  }
  #number of studies
  nostudies <- length(sei)

  #create a blank matrix where all the results will go...
  CImat<-matrix(NA,nrow=length(CIest)*2,ncol=length(hetnames))
```

```
colnames(CImat) <- hetnames
#define row names...
rownammat <- rep(NA,times=nrow(CImat))
for (i in 1:nrow(CImat)) {
  #if i is even
  if (i %% 2 == 0) rownammat[i] <- paste(CIest[i/2],"_ub",sep="")
  else rownammat[i] <- paste(CIest[(i/2)+0.5],"_lb",sep="")
}
rownames(CImat) <- rownammat

#Z-type CI
if ('Z' %in% CIest) {
  for (i in 1:length(hetnames)) {
    CImat["Z_lb",hetnames[i]] <- thetaests[i] -
      qnorm(1-(signif/2))*sqrt(1/sum(1/(hetests[i]+(sei^2))))
    CImat["Z_ub",hetnames[i]] <- thetaests[i] +
      qnorm(1-(signif/2))*sqrt(1/sum(1/(hetests[i]+(sei^2))))
  }
}

#t-type CI
if ('T' %in% CIest) {
  for (i in 1:length(hetnames)) {
    CImat["T_lb",hetnames[i]] <- thetaests[i] -
      qt(1-(signif/2), df=nostudies-1)*
        sqrt(1/sum(1/(hetests[i]+(sei^2))))
    CImat["T_ub",hetnames[i]] <- thetaests[i] +
      qt(1-(signif/2), df=nostudies-1)*
        sqrt(1/sum(1/(hetests[i]+(sei^2))))
  }
}

#Hartung-Knapp CI
if ('HK' %in% CIest) {
  for (i in 1:length(hetnames)) {
    varHK <- sum((1/(hetests[i]+(sei^2)))*((xi-thetaests[i])^2))/
      ((nostudies-1)*sum(1/(hetests[i]+(sei^2))))
    CImat["HK_lb",hetnames[i]] <- thetaests[i] -
      qt(1-(signif/2), df=nostudies-1)*sqrt(varHK)
    CImat["HK_ub",hetnames[i]] <- thetaests[i] +
      qt(1-(signif/2), df=nostudies-1)*sqrt(varHK)
  }
}

if (output==TRUE) print(CImat)
```

```
  CImat <<- round(CImat,digits=signifCI)

}
```

Appendix E: Simulation study protocol

Protocol amendments

- Generic meta-analyses are no longer included because the results are almost identical to SMD meta-analyses. The only difference between the two meta-analysis types is that within-study variances are based on the underlying study effects, as opposed to the observed study effects.

- I simulate 5,000 meta-analyses per scenario, not 10,000 as stated in the original protocol.

- Scenarios 4 and 5 for the study sample size distributions have been changed. Originally, these two scenarios were:

  (4) $n_1 = 20$ and $n_2, ..., n_k \sim U(1000, 2000)$
  (5)$n_1 \sim U(1000, 2000)$ and $n_2, ..., n_k \sim 20$.

  I changed them to:

  (4) Small and large studies: $n_{11}, ..., n_{1m} = 20$ and $n_{1m}, ..., n_{1k} \sim U(1000, 2000)$ where m is the integer half way between 1 and k (when k is odd, one study is be generated from one of the two distributions at random)
  (5) Large studies: $n_{1i} \sim U(1000, 2000)$

  I took this decision because the original mixture distributions were dependent on the number of studies k.

- The analysis plan originally stated "Typical confidence intervals for τ^2 will also be presented, which will help show scenarios where τ^2 estimates are imprecise". This will not be included as there is no way of effectively displaying these on the graphs and would divert attention away from the main results.

The protocol that follows in this appendix is the original protocol agreed between collaborators.

PROTOCOL

A Comprehensive Simulation Study to Compare Methods of Estimating Heterogeneity Variance in Meta-analysis

Version 3.0, 23/12/2014

Dean Langan (dl790@york.ac.uk) [1]

Julian PT Higgins (julian.higgins@bristol.ac.uk) [2]

Mark Simmonds (mark.simmonds@york.ac.uk) [1]

Dan Jackson (dan.jackson@mrc-bsu.cam.ac.uk) [3]

Jack Bowden (jack.bowden@mrc-bsu.cam.ac.uk) [3]

Areti Angeliki Veroniki (veronikia@smh.ca) [4]

Evangelos Kontopantelis (e.kontopantelis@manchester.ac.uk) [5]

Wolfgang Viechtbauer (wolfgang.viechtbauer@maastrichtuniversity.nl) [6]

[1] Centre for Reviews and Dissemination, University of York, York, YO10 5DD, UK

[2] School of Social and Community Medicine, University of Bristol, Bristol, UK

[3] MRC Biostatistics unit, Cambridge, UK

[4] Li Ka Shing Knowledge Institute, St. Michael's Hospital, 209 Victoria Street, East Building. Toronto, Ontario, M5B 1T8, Canada

[5] Centre for Health Informatics, Institute of Population Health, University of Manchester, Manchester, UK

[6] Department of Psychiatry and Psychology, Maastricht University, The Netherlands

Background

Heterogeneity is an important consideration in any meta-analysis, as its presence can have a considerable impact on the conclusions reached. All meta-analyses should examine the extent of heterogeneity across studies. There are various ways of doing this, and of addressing heterogeneity in the statistical synthesis. A common approach, often advocated when an adequate explanation cannot be identified for a moderate amount of between-study variation, is to use a random-effects model. As part of a random-effects meta-analysis, a heterogeneity variance parameter (τ^2) is estimated. This parameter may be used to quantify heterogeneity even if a synthesis is not performed.

Several methods have been proposed to estimate τ^2, the most commonly used being the DerSimonian- Laird method [25]. However, simulation studies suggest this method underestimates heterogeneity variance in dichotomous outcome meta-analyses [78, 102]. Other heterogeneity estimation methods include Paule-Mandel's method [24, 80], which falls under the same method of moments approach as DerSimonian-Laird. Estimators have also been proposed based on maximum-likelihood (ML) [37] and restricted maximum likelihood (REML) [124] approaches. These estimators allow for negative estimates of heterogeneity variance and must be truncated to zero in such cases. Hartung-Makambi [40], Sidik-Jonkman [101] and the improved Sidik-Jonkman [102] estimators are designed to provide only positive estimates so that truncation is not required.

A recent empirical study (at the Centre for Reviews and Dissemination in York) on meta-analyses from the Cochrane Database of Systematic Reviews (CDSR) compared five methods including DerSimonian-Laird method, Paule-Mandel, Hartung-Makambi, Sidik-Jonkman and REML [67]. For each method, I^2 statistics were calculated as an estimate of the level of heterogeneity using the generalised formula proposed by Higgins and Thompson [47]. The study found that estimates of I^2 derived from different methods can be highly discordant. Differences between τ^2

estimates can lead to I^2 values that differ in absolute terms by more than 50%. Discrepancies between heterogeneity variance estimates can also lead to discordant conclusions on the summary effect; the empirical study showed that 7.4% of Cochrane meta-analyses have discordant statistical significance at the 5% level depending on which heterogeneity estimation method is used (out of five methods that were compared). The disagreement between heterogeneity variance estimates can be partly attributed to the low number of studies typically found in meta-analyses of healthcare interventions: those in the Cochrane Database of Systematic Reviews contain a median of just 3 studies (inter-quartile range 2 to 6)(Davey et al., 2011). When conducting a meta-analysis, it is therefore important to make an informed decision when choosing which heterogeneity estimation method to use and whether to rely on one point estimate alone.

Twelve simulation studies assessing the performance of heterogeneity variance estimators have been identified from a preliminary review of the literature, which are summarised in chapter 5. Of these studies, four compare a wide selection of estimators over a wide range of simulated scenarios (Novianti et al., 2014, Viechtbauer, 2005, Sidik and Jonkman, 2007, Kontopantelis et al., 2013). One study (Viechtbauer, 2005) recommends REML, one study (Novianti et al., 2014) recommend Paule-Mandel and Sidik and Jonkman (Sidik and Jonkman, 2007) recommend their own methods. Kontopantelis et al. suggested that the bootstrap alternative to the standard DerSimonian-Laird method performs better but mainly recommends conducting sensitivity analyses; in most meta-analyses, there is insufficient data to rely on one estimate of heterogeneity alone. DerSimonian-Laird, Cochran's ANOVA and REML were the only estimators included in all four of the main studies. The study by Kontopantelis et al. [64] was the only main study to include Bayesian estimators from Rukhin [93]. Only the study by Novianti et al. [78] included the Paule-Mandel method in its comparisons. Aside from the four most comprehensive studies, other simulations have been identified in the literature in which only a small selection of heterogeneity variance estimators are compared [2, 3, 11, 61, 74, 79, 93, 96, 101].

Previous simulation studies have made conflicting recommendations for heterogeneity estimation and as such there is currently no overall consensus. This can be partly attributed to differences between studies in the effect sizes used to determine what makes a good heterogeneity variance estimator. Performance of heterogeneity variance estimators in previous studies have been assessed mainly in terms of bias, variance or some measure of precision. Bias is a measure of how much the heterogeneity parameter is under or overestimated. Measures of precision include the mean squared error (MSE) and efficiency, which quantify the expected deviation from the true parameter value. Recommendations made by Viechtbauer [124] were based a trade-off between minimising bias and maximising efficiency, as is the case in most other simulation studies. In contrast, the study by Novianti et al. [78] made recommendations based only on which method has the lowest bias. Previous studies have aimed to provide a simple recommendation that can easily be applied in practice, yet their results suggests no estimator is clearly best under all conditions.

Comprehensive simulation studies are needed to examine methods for estimating heterogeneity variance and for incorporating these into random-effects meta-analyses. The current study primarily addresses the first part of this, namely the choice of variance estimator. There are conflicting recommendations about this issue in the current literature, and a number of specific questions need answering to inform a consensus recommendation. Existing simulation studies compare the performance of heterogeneity variance estimators mainly on meta-analyses with uniformly distributed study sample sizes. No simulation study compares multiple distributions of study sample sizes. IntHout et al. [55] suggests the performance of heterogeneity methods in meta-analysis varies depending on the distribution of study sizes. Research where true treatment effects are simulated from a non-normal distribution is also limited to a comparison between effect (θ) estimation methods, but not specifically heterogeneity variance estimators [62, 63]. The two main limitations of current recommendations are: (1) they are only based on a comparison of a small subset of all heterogeneity estimation methods available (2) they do not address sufficiently the practical situation that in many meta-analyses all heterogeneity variance estimates

are very imprecise.

This study aims to address the limitations of the current research, comparing the performance of heterogeneity variance estimators using simulated meta-analysis data that resemble conditions that may occur in practice. Empirical meta-analysis data taken from a 2008 snapshot of the *Cochrane Database of Systematic Reviews* (CDSR) will be used to inform parameter values for simulations [21]. A comprehensive selection of heterogeneity estimation methods will be compared, as identified in a recent review yet to be published [122].

Aims

The principal aim of this study is to make clear recommendations for meta-analyses in a wide range of realistic situations about which heterogeneity variance estimator (if any) is most appropriate to use. Recommendations will be informed by how heterogeneity variance estimators perform in simulated meta-analyses and agreed between collaborators of the study. Parameter values used to simulate meta-analyses through a random-effects model will cover the full range of possible scenarios observed in practice. The study will answer the following questions:

- In what situations do all estimators perform poorly, where relying on one point estimate of heterogeneity is not recommended?
- In what situations does one estimator outperform all others and perform well enough to provide a reasonable point estimate of heterogeneity?
- In what situations do most heterogeneity variance estimators perform equally well?
- Are there any estimators that we can exclude entirely?
- Are there characteristics of the meta-analysis that explain the poor estimation?

Heterogeneity variance estimators

Methods for estimating the heterogeneity variance (τ^2) in a random-effects model were identified from a comprehensive review of heterogeneity methods [122]. For each simulated meta-analysis, estimates of τ^2 will be calculated from the following 14 methods:

- Cochran's ANOVA [18] (also known as Hedges-Olkin [42])
- DerSimonian-Laird [25]
- Paule-Mandel [80]
- Two-step DerSimonian-Laird [24]
- Two-step Cochran's ANOVA [24]
- Hunter-Schmidt [53]
- Maximum likelihood [37]
- Restricted maximum likelihood [124]
- Hartung-Makambi [40]
- Sidik-Jonkman [101]
- Improved Sidik-Jonkman [102]
- Rukhin with zero prior [93]
- Rukhin's simple estimator [93]

Details of how to estimate τ^2 from all these methods are given in chapter 2. A small number of methods were excluded, as listed on page113 with reasons for exclusion.

Confidence intervals for the summary effect

In order to perform a comprehensive comparison of heterogeneity variance estimators, their performance must be evaluated not only as a point estimate of heterogeneity (τ^2). This study also investigates the impact using a given heterogeneity variance estimator to calculate a summary effect and confidence interval of the summary effect. Many methods of calculating the confidence interval have been proposed. Because inference on the mean of the random-effects distribution is not the main focus of the simulation study, only a small subset of confidence interval methods available will be included:

- Wald-type [25]
- t-distribution (with number of studies -1 degrees of freedom) [28]
- Hartung-Knapp [38]

All these methods are independent of the choice of heterogeneity variance estimator. Therefore, any combination of methods can be applied in practice and all combinations are considered in this simulation study. 95% confidence intervals will be calculated for all analyses. Details of each confidence interval method are given in chapter 3.

Many methods are also available to calculate a confidence interval for τ^2. These methods will not be considered as the scope of this study is limited to the impact of using point estimates of heterogeneity.

Performance measures

Heterogeneity variance estimators will be compared in terms of the following 11 performance measures:

- Median and mean absolute bias in estimate of τ^2 *
- Median and mean squared error of estimate of τ^2 *
- Proportion of zero estimates of τ^2
- Mean absolute bias in estimate of the mean treatment effect ǂ
- Mean squared error of estimate of the mean treatment effect ǂ
- Coverage of the 95% confidence interval for the mean treatment effect** (i.e. the proportion of times the underlying mean treatment effect falls inside the 95% confidence interval)
- Power to detect a significant summary effect
- Mean and variance of the error-interval estimation of effect §

*Previous studies have used the mean squared error to measure performance, but τ^2 and estimation errors do not conform to the normal distribution. Performance measures based on the mean make heterogeneity variance estimators with negative bias appear better as underestimates are more likely to be negative and therefore truncated at zero. Therefore, median bias and squared errors will also be presented.

ǂ Mean treatment effect calculated by the weighted inverse variance method

**Confidence interval coverages will also be compared against the coverage of each confidence interval method based on the true value of the heterogeneity variance.

§ Error-interval estimation is a ratio between the width of the estimated confidence interval and the true confidence interval, as defined by a previous simulation study [64]. The formula is given in the original paper.

A good estimator is unbiased, has a low MSE and a summary effect confidence interval with coverage close to 95%. Details of how to calculate all performance measures are given in appendix C.2

Simulation methods

Analysis will be undertaken using simulated data produced from the following steps:

1. A meta-analysis dataset is generated for specified parameter values using the methods outlined on page 248.
2. Heterogeneity variance estimates are calculated for the given meta-analysis using methods on page 244.
3. Steps 1 and 2 are repeated 10,000 times and performance measures are calculated (see page 245)
4. Steps 1-3 are repeated for all combinations of parameter values. The parameter values are given on page 252.

All steps will be carried out in R [85]. The *metafor* package in R [126] will be used to calculate estimates of heterogeneity from methods coded in this package, and bespoke code for those that are not. Methods available in metafor to estimate τ^2 include Cochran's ANOVA, DerSimonian-Laird, Hunter-Schmidt, Sidik-Jonkman, maximum likelihood and REML. Wald-type and Hartung-Knapp confidence interval methods for the mean treatment effect are also available in *metafor*. Heterogeneity methods will be compared using the same simulated datasets to eliminate some of the sampling error. Maximum likelihood and REML heterogeneity variance estimators are iterative and fail to converge to a solution in a small number of cases [64], but this is primarily due the chosen iteration algorithm rather than the estimator [126]. In this study, the default iteration algorithm in *metafor* will be used - Fishers scoring method with Cochran's ANOVA the initial estimate [126]. Simulated meta-analyses that cause such failures will not be replaced, and instances recorded so that the characteristics of the simulated data can be examined. Heterogeneity variance estimates for each meta-analysis and performance measures for each combination of parameter values will be stored for the analysis. The simulation code has already

been written according to the methods described in this version of the protocol and is available on request.

Simulating meta-analyses

For studies $i = 1, \ldots, k$ in each meta-analysis, true treatment effects (θ_i such as a log odds ratio or standardized mean difference), are simulated from some distribution D_1:

$$\theta_i \sim D_1\left(\theta, \tau^2\right)$$

where θ is the mean parameter and τ^2 the heterogeneity variance parameter of D_1 Defined distributions for D_1 with parameter values θ and τ^2 used to simulate meta-analyses are detailed on page 252. For each θ_i sampled from D_1, estimates of θ_i (denoted by $\hat{\theta}_i$) are then generated to simulate within-study sampling error. The process for doing so depends on the type of outcome of studies in each meta-analysis. In this study, three types of meta-analyses will be simulated: (1) Generic effect sizes with known variance; (2) continuous outcome meta-analyses with a standardised mean difference effect measure; and (3) dichotomous outcome meta-analyses with an odds ratio effect measure, as detailed on pages 248 to 250.

Generic effect sizes (with known variance)

Estimates of θ_i are simulated from a normal distribution:

$$\hat{\theta}_i \sim N(\theta_i, \sigma_i^2)$$

where θ_i is the mean parameter and σ_i^2 is the sampling variance parameter.

$\sigma_i^2$are calculated based on the sample size of each study. We assume equal sample sizes between intervention groups and denote the sample size for each group by n_i:

$$\sigma_i^2 = \frac{2}{n_i} + \frac{\theta_i^2}{4n_i}$$

n_i are generated from one of a number of distributions as detailed on page 252. The formula for σ_i^2 above is the approximate variance of a standardised mean difference with sample sizes equal between arms and $\theta = 0$ [8]. By using this formula to derive σ_i^2, simulation results from outcome-independent meta-analyses can be directly compared with standardised mean difference meta-analyses where n_i are simulated from the same distributions. Results from outcome-independent meta-analyses will represent the performance of heterogeneity variance estimators under ideal conditions, where standard errors are known and performance is not affected by the choice of effect measure.

Standardised mean difference meta-analyses

For each study i with simulated true effect θ_i:

1. Generate sample sizes for each intervention group, denoted by n_i, from one of a number of distributions as detailed on page 252.

2. Generate n_i observations from $N(0, \sigma_{Ti}^2)$ and n_i observations from $N(\theta_i, \sigma_{Ci}^2)$, to represent patient-level data in the treatment and control groups respectively. Without loss of generality, we set $\sigma_{Ti}^2 = \sigma_{Ci}^2 = 1$.

3. Calculate the sample mean and standard deviation of these observations for the treatment and control groups.

4. Calculate the sample SMD and standard error of the study, denoted by $\hat{\theta}_i$, using Hedge's g method [8]:

$$\hat{\theta}_i = \frac{\overline{Z_{Ti}} - \overline{Z_{Ci}}}{s_i} J$$

where s_i is the pooled variance and J is an adjustment to correct for bias:

$$s_i = \sqrt{\frac{(n_i - 1)\,\hat{sd}^2_{Ti} + (n_i - 1)\,\hat{sd}^2_{Ci}}{2n_i - 2}}$$

$$J = 1 - 3/\left(8n_I - 9\right)$$

5. Calculate the variance of $\hat{\theta}_i$:

$$\hat{\sigma}^2_i = \left(\frac{2}{n_i} + \frac{\hat{\theta}^2_i}{4n_i}\right) \cdot J^2$$

Hedge's g method calculates a pooled standard deviation assuming equal variances between treatment groups, which is the case in this simulation study.

Odds ratio meta-analyses

Meta-analyses with an odds ratio effect measure are simulated for two reasons: (1) Previous simulation studies [78] suggest heterogeneity variance estimators perform worse in this setting compared with SMD meta-analyses and (2) the odds ratio is the most common outcome measure in binary outcome meta-analyses [120]. It has been suggested this is partly because σ^2_i are estimated poorly when the event of interest is rare [3]. For each study i with simulated true effect θ_i:

1. Generate the true average probability of an event across the control and treatment groups, denoted by $\overline{p_i}$. $\overline{p_i}$ is drawn from the distributions detailed on page 252. p_{Ti} and p_{Ci} are found from solutions to the simultaneous equations:

$$\overline{p_i} = \frac{p_{Ti} + p_{Ci}}{2}$$

$$OR_i = \frac{p_{Ti}\left(1 - p_{Ci}\right)}{p_{Ci}\left(1 - p_{Ti}\right)}$$

Solving the equations leads to the following formula for p_{Ti} and p_{Ci}:

$$p_{Ci} = \frac{1}{2\left(OR_i - 1\right)} \cdot \left[2\overline{p_i}OR_i - OR_i - 2\overline{p_i} - 1 - \sqrt{\left(OR_i - 2\overline{p_i}OR_i + 2\overline{p_i} + 1\right)^2 + 8\overline{p_i}\left(OR_i - 1\right)}\right]$$

$$p_{Ti} = 2\overline{p_i} - p_{Ci}$$

2. Generate sample sizes for each intervention group, denoted by n_i, from one of a number of distributions as detailed on page 252.

3. The numbers of events in the control and treatment groups are generated from the binomial distributions $B(n_{Ci}, p_{Ci})$ and $B(n_{Ti}, p_{Ti})$. n_{Ci} and n_{Ti} are the sample sizes of the treatment and control groups and will be assumed equal for this study ($n_{Ci} = n_{Ti} = n_i$). Cell counts in a 2x2 contingency table can then be derived.

4. Add 0.5 to all cell counts if there is any zero in the table. If there are zero events in both arms then exclude this study from the synthesis. If there are fewer than 2 studies remaining after exclusions then the meta-analysis will be withdrawn from the simulations without replacement.

5. Calculate the sample OR on the log scale, $\hat{\theta}_i$ and its variance [8].

Parameter values

Performance of the heterogeneity variance estimators will be assessed for all combinations of parameter values and distributions given in table E.1 for outcome-independent, standardised mean difference and odds ratio meta-analyses. There will be a total of 840 simulated scenarios for outcome-independent meta-analyses, 840 standardised mean difference meta-analyses and 3360 simulated scenarios for OR meta-analyses. Each scenario will have 10,000 simulated datasets. It is estimated that each scenario will take 5 minutes to simulate given the computing power available and a total of 17 days to simulate all data. Justification of each parameter and distribution is given in this section.

Parameter values were chosen to represent the range of values observed in published meta-analyses.

(1) Figure E.1 is taken from a recent empirical study at the Centre for Reviews and Dissemination (CRD) in York and shows the number of studies in meta-analyses from the *Cochrane Database of Systematic Reviews* up to 2008. Note, the figure excludes meta-analyses with fewer than 3 studies, calculating an estimate of heterogeneity was considered inappropriate in meta-analyses with two studies. The distribution in figure E.1 was used to inform parameter values for k. The number of studies per meta-analysis in simulations will range between 2 and 100, although 95% of meta-analyses in the Cochrane database contain fewer than 16 studies; this is to account for meta-analyses with a higher number of studies in other fields.

(2) Parameter values of τ^2 were chosen based on the distribution of τ^2 predicted in meta-analyses from Cochrane reviews. Two empirical studies used Bayesian methods derive a distribution for τ^2 based on meta-analyses with an odds ratio outcome [120] and standardised mean difference outcome [88]. The studies showed similar distributions of τ^2 between the two outcomes types. Therefore, the same parameter values will be used to simulate meta-analyses of all outcomes. Using these predictive distributions for τ^2 is more appropriate than a distribution of τ^2 estimates which

Parameter			Value/distribution
1	k	Number of studies in the meta-analysis	2, 3, 5, 10, 20, 30, 50, 100
2	I^2	Level of heterogeneity	0%, 15%, 30%, 45%, 60%, 75%, 90%, 95%
3	θ	Mean of the random-effects	0.5
4	θ_i	Distribution of true study effects	*(1)* $\theta_i \sim N(\theta, \tau^2)$(standard random-effects model) *(2)* Normal distribution with moderate skew: $\theta_i \sim SN(\theta, \tau^2, \gamma = 0.7)$ *(3)* Normal distribution with high skew: $\theta_i \sim SN(\theta, \tau^2, \gamma = 0.95)$ τ^2 takes parameter values that satisfy the I^2 values above and $\theta = 0.5$
5	n_i	Sample size in each intervention group (1:1 allocation ratio)	*(1)* $n_i = 20$ *(2)* $n_i \sim U(20, 200)$ *(3)* $n_i = 200$ *(4)* $n_1 = 20$ and $n_2, ..., n_k \sim U(1000, 2000)$ *(5)*$n_1 \sim U(1000, 2000)$ and $n_2, ..., n_k \sim 20$ In all scenarios, sample sizes are equal between groups ($n_{1i} = n_{2i}$)
Parameters only applying to **odds ratio meta-analyses**			
6	$\overline{p_i}$	Average probability of event across treatment and control groups	*(1)* $\overline{p_i} = 0.5$ *(2)* $\overline{p_i} \sim U(0.1, 0.5)$ *(3)* $\overline{p_i} = 0.05$ *(4)* $\overline{p_i} = 0.01$

Table E.1: Set of parameter values and distributions to simulate meta-analyses

is more dependent on which heterogeneity estimation method is used. Preliminary simulations show that the chosen parameter values for τ^2 result in generic meta-analyses with I^2 values that span the full I^2 range from 0% up to nearly 100%. This

dataset of Cochrane reviews is described in detail elsewhere [21].

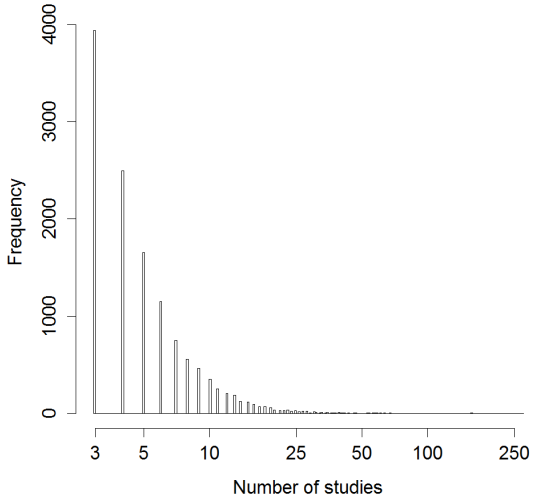


Figure E.1: Histogram of the numbers of studies in meta-analyses in the Cochrane Database of Systematic Reviews (CDSR)

(3) To reduce the number of scenarios, we will only simulate meta-analyses where $\theta = 0.5$. The choice is immaterial for simulation of generic effect sizes, and previous simulation studies have suggested that the parameter value chosen for true summary effect (θ) has little bearing on performance measures including bias and variance of heterogeneity variance estimators [78, 102].

(4) True treatment effects θ_i will be generated from three distributions. In all scenarios, θ_i are sampled from distributions with mean θ and variance τ^2. First, θ_i will be generated from the normal distribution which is assumed in the standard random-effects model and represents optimal conditions where estimators may perform best (scenario 1). Some heterogeneity estimation methods such as Paule-Mandel do not assume normality of true effects and therefore it is hypothesised such estimators will be more robust under non-normal conditions [24]. Second, θ_i will be sampled from a skew normal distribution (scenario 2) with a 0.8 skew parameter value; this represents a moderate negative skew as illustrated in figure E.2. A simulation study [62, 63] previously looked at performance of heterogeneity variance estimators under skew-normal conditions, and defined this similar level of skew as 'moderate'. This

distribution is defined elsewhere [68]. Third, θ_i will be generated from a bimodal distribution with studies drawn from two normal distributions of unequal means; this represents a scenario where some dichotomous factor is responsible for some of the heterogeneity present. The means of the two normal distributions were chosen so that the resulting mixture distribution has mean $\theta = 0$ and variance τ^2. The level of bi modality is dependent on the parameter value of τ^2 [98]. Kontopantelis et al. [64] conducted a simulation study of non-normal treatment effects and showed, of the heterogeneity variance estimators compared, that the performance of methods were moderately robust to various distributions of effects.

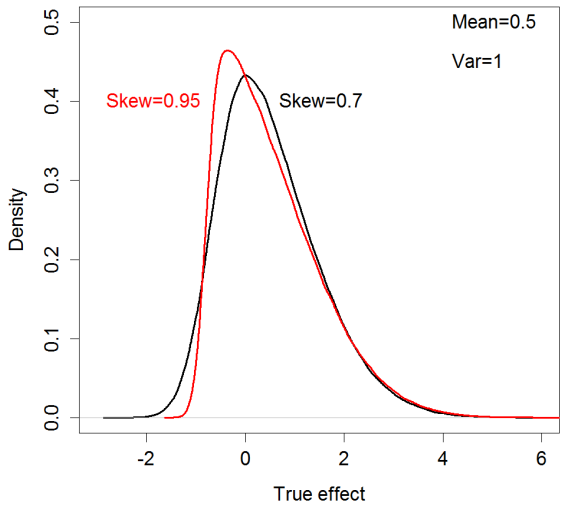


Figure E.2: Probability density function of skew-normal distribution
Note that the variance differs depending on the simulation scenario, but this is only the scaling parameter.

(5) Study sample sizes (n_i) will be generated from five different distributions to represent a wide range of distributions. Distributions include small equally sized studies (scenario 1), medium equally sized studies (scenario 2), uniform variation of small to medium size studies (scenarios 3), one large study with all other studies small (scenario 4) and one small study with all other studies large (scenario 5). A recent simulation study suggests that the performance of heterogeneity estimation methods may be dependent on the distribution of study sizes; this study however, only compared two heterogeneity methods [55]. Other previous studies simulated sample sizes from one distribution only; namely a uniform [78, 102], normal [124] or

χ^2 distribution [64].

(6) Binary outcome meta-analyses will be generated from a range of underlying event rates. In scenario 1, the underlying average event rate will be 0.5 to represent the ideal scenario with event rates sampled as far from the asymmetric tails of the binomial distribution as possible. Scenario 2 represents a situation where event rates are variable between studies but not so rare as to be considered a big contributing factor to poor estimates of treatment effects and standard errors. Scenarios 3 and 4 represent situations where the average underlying event rate is homogeneous and rare. It is not necessary to simulate meta-analyses where the event of interest is extremely common (i.e. $\overline{p_i} = 0.95$ or 0.99) as the resulting odds ratios are the inverse of those obtained with extremely uncommon event rates.

Analysis

Primary analysis

Heterogeneity variance estimators will be compared in terms of the performance measures defined on page 245 and will be presented graphically for each simulated scenario. Graphs will be produced for each performance measure to compare the results of the 14 heterogeneity variance estimators. For each scenario, typical I^2 values will be presented to show the level of heterogeneity the scenario represents. Typical confidence intervals for τ^2 will also be presented, which will help show scenarios where τ^2 estimates are imprecise. 95% confidence intervals will be calculated by the Q-profile method [9]. If estimators are judged to have similar performance, they may be grouped together to simplify results. Also, if the results are similar between different parameter values, the results from such scenarios will be combined together.

Maximum likelihood and REML estimators may in some meta-analyses fail to converge [62, 63]. Therefore for each scenario, the percentage of failures will be tabulated

and results of which will be taken into account when making recommendations. If other iterative heterogeneity variance estimators fail to converge, results of such failures will also be presented.

Secondary analysis

The range of simulation scenarios in this study aims to be representative of all meta-analyses from Cochrane systematic reviews. A secondary analysis will identify which scenarios from this study occur most in reviews published in the Cochrane Database of Systematic Reviews up to 2008 [21]. This will identify the importance of results of each scenario and also help inform recommendations. As the level of heterogeneity can only be estimated in real meta-analyses, we will use distributions of the underlying level of heterogeneity derived using Bayesian techniques in two empirical studies [88, 120] to assess how frequently each τ^2 parameter value occurs in practice. Distributions derived from the two studies were based separately on OR outcome [120] and SMD outcome meta-analyses [88]. Both studies assume that $\tau^2 = 0$ are untenable in real meta-analyses and therefore the distribution has zero probability of such values. As a consequence, we make this assumption in our analysis. Only scenarios where true treatment effects have been simulated from the normal distribution will be included in this analysis (scenario 1, see page 252); identifying non-normal effects in real meta-analyses would be difficult given most meta-analyses contain few studies. Results will be presented as a list of scenarios in the order of scenarios most likely to occur in practice to the least.

Simulated meta-analyses will be selected for further investigation where heterogeneity variance estimates are particularly discordant. These meta-analyses will be explored to identify characteristics that may have caused discordance.

Recommendations

We aim to make recommendations about the best choice of heterogeneity variance estimator in a meta-analysis with any given observed characteristics and make recommendations as a general strategy. Recommendations based on the results of this study are likely to be made mainly on a subjective compromise between the results from all performance measures and the practicality of such recommendations. Simple estimators will be recommended over iterative estimators where the difference in performance in negligible. An estimator will be recommended in all scenarios to provide a point estimate of heterogeneity in the primary random effects model. However, making conclusions based a single point estimate will only be recommended when the estimate is sufficiently precise, alternative approaches will be recommended otherwise. A decision tree may be formulated if it is appropriate to do so that can be used in a given meta-analysis. The process of interpreting and summarising the results of this study will involve all collaborators to make this study as systematic as possible. Any relevant discussions will also be documented as part of the results.

Appendix F: Supplementary material from chapter 7

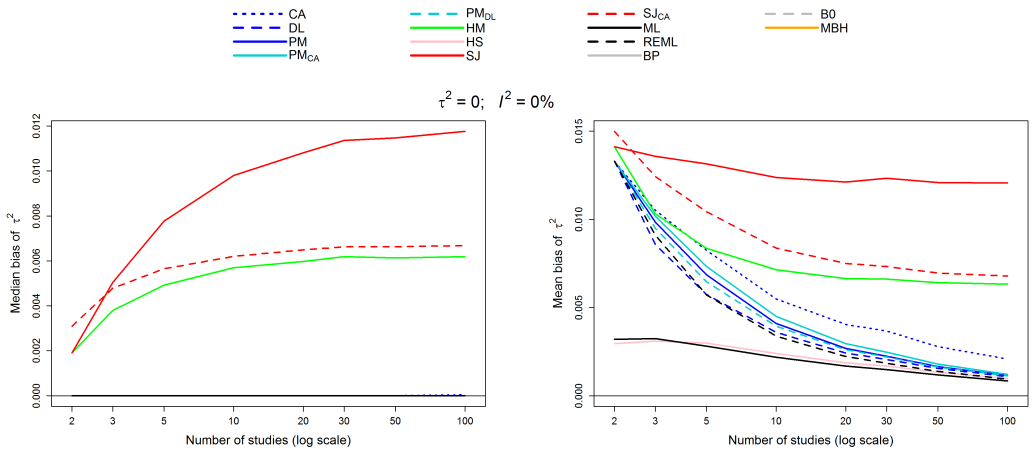


Figure F.1: Proportional median bias (left-hand-side) and proportional mean bias (right-hand-side) of the heterogeneity variance in selected scenarios to show why median bias is excluded from the main results.
Scenarios containing standardised mean difference meta-analyses with small-to-medium study sizes and I^2 of 0%.

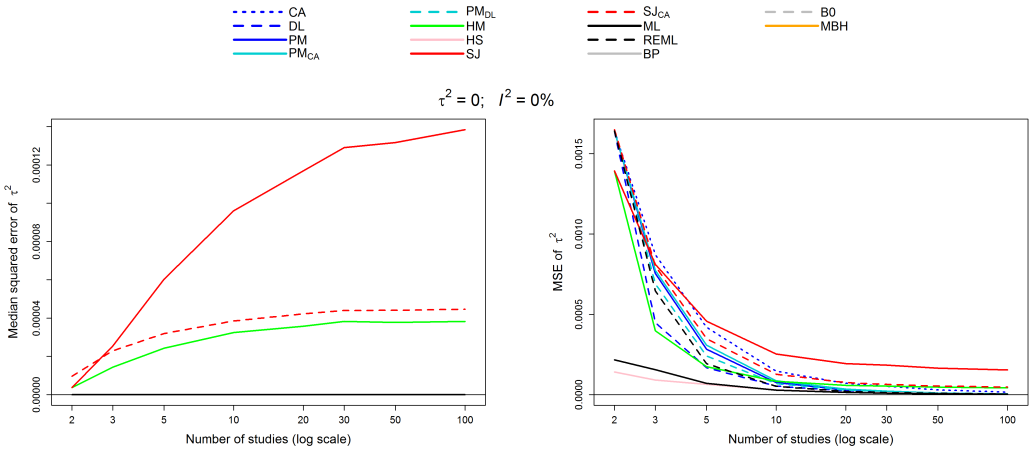


Figure F.2: Proportional median squared error (left-hand-side) and proportional mean squared error (right-hand-side) of the heterogeneity variance in selected scenarios to show why median squared bias is excluded from the main results.
Scenarios containing standardised mean difference meta-analyses with small-to-medium study sizes and I^2 of 0%.

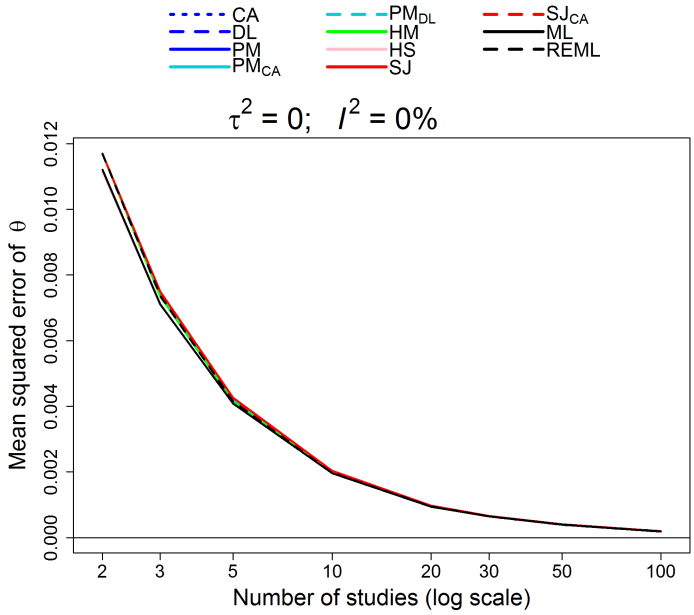


Figure F.3: Mean squared error of the summary effect in selected scenarios to show why this measure is excluded from the main results.
Scenarios containing standardised mean difference meta-analyses with small-to-medium study sizes and I^2 of 0%.

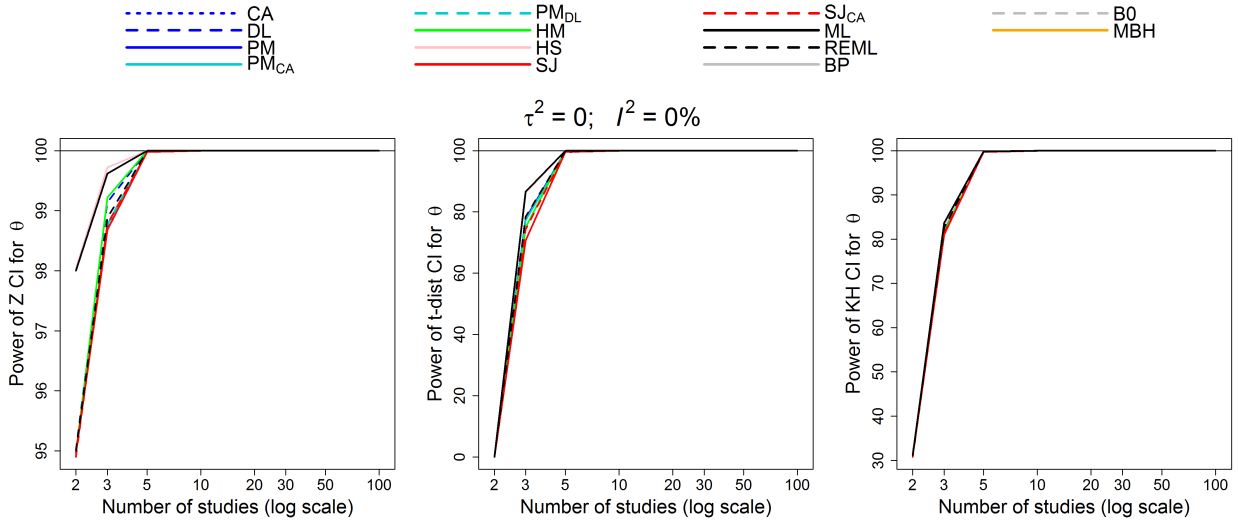


Figure F.4: Power to detect a statistically significant summary effect in selected scenarios to show why this measure is excluded from the main results.
Scenarios containing standardised mean difference meta-analyses with small-to-medium study sizes and I^2 of 0%.

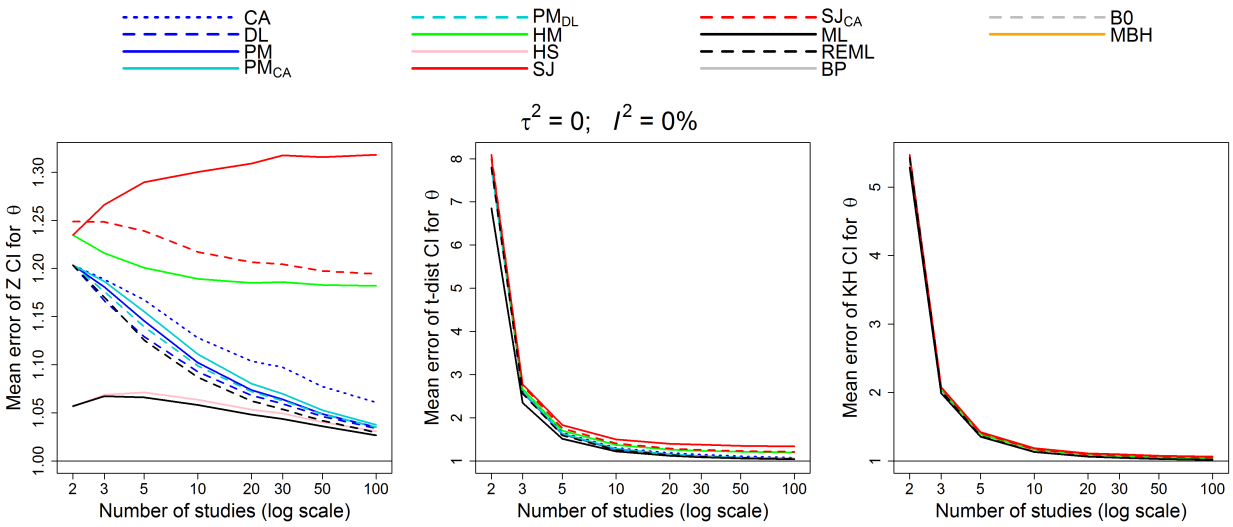


Figure F.5: Mean error of the error-interval estimation of effect in selected scenarios to show why this measure is excluded from the main results. *Scenarios containing standardised mean difference meta-analyses with small-to-medium study sizes and I^2 of 0%.*

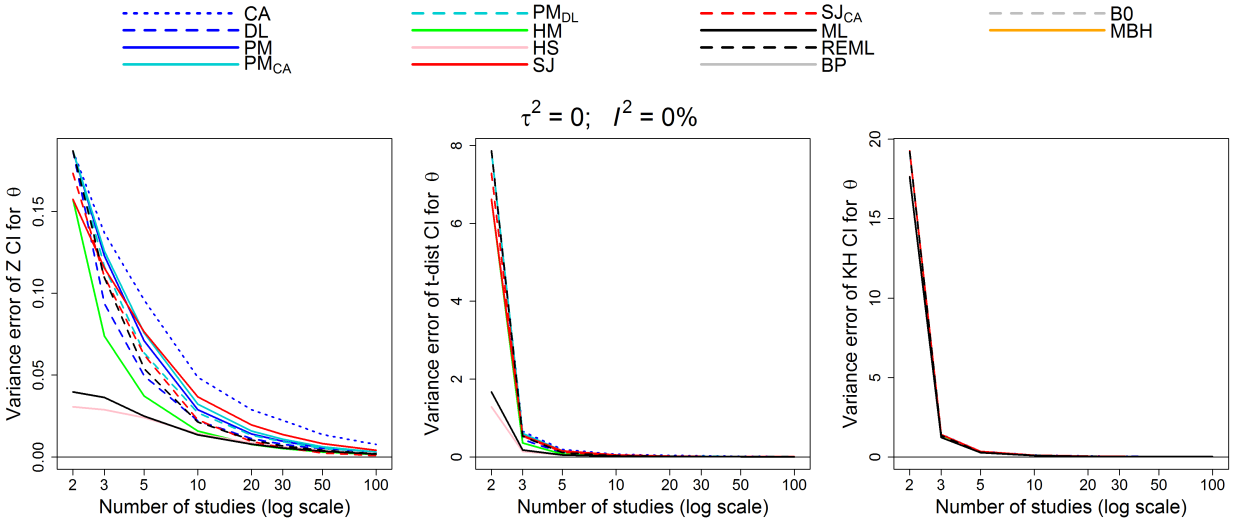


Figure F.6: Variance error of the error-interval estimation of effect in selected scenarios to show why this measure is excluded from the main results. *Scenarios containing standardised mean difference meta-analyses with small-to-medium study sizes and I^2 of 0%.*

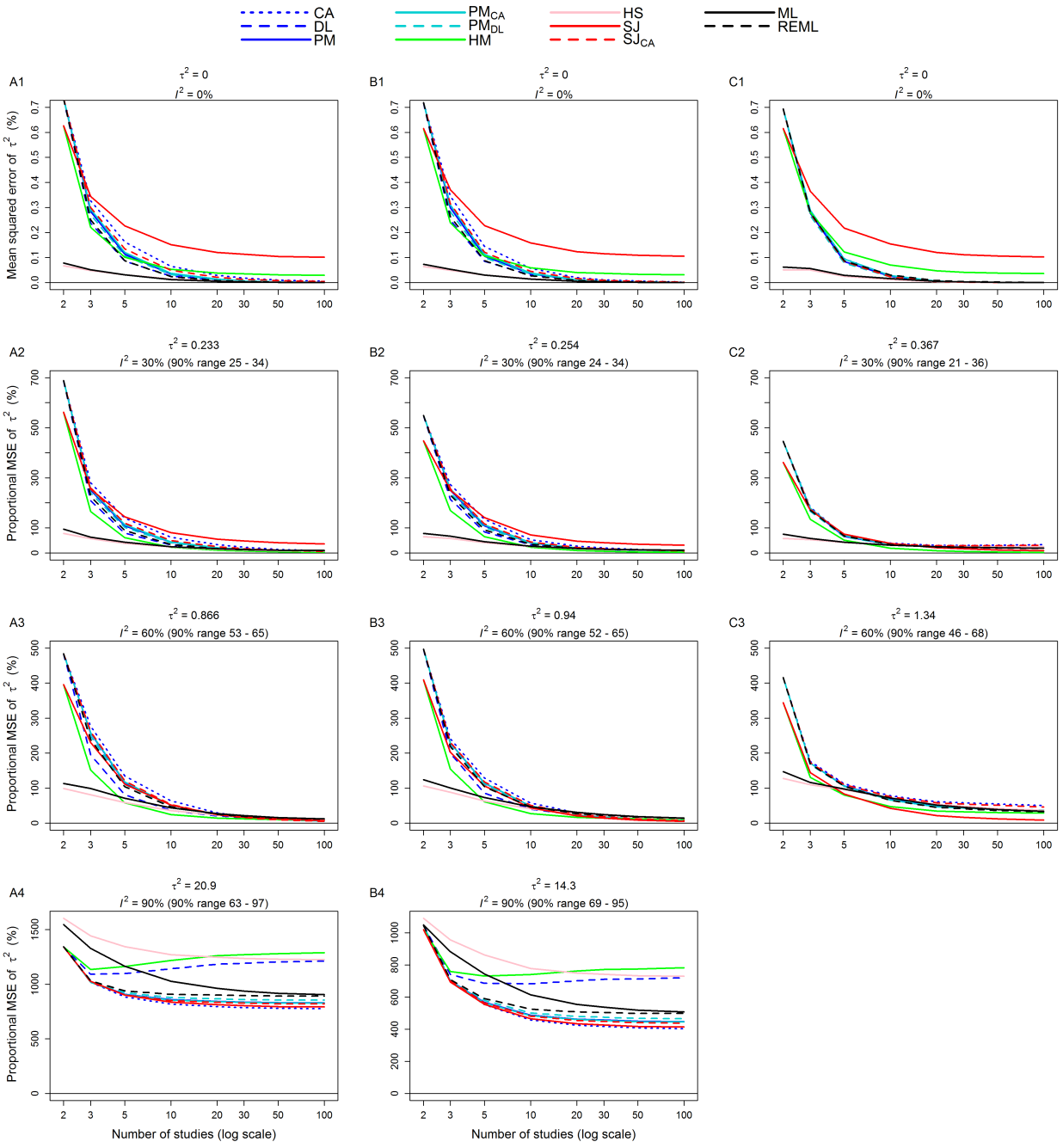


Figure F.7: Mean squared error of heterogeneity variance estimates in odds ratio meta-analyses containing small studies and with event probability 0.1 to 0.5

Scenarios with an underlying summary odds ratio of 1.65 (A1-A4), 3 (B1-B4) and 10 (C1-C4).

MSE is presented on the proportional scale only when $\tau^2 > 0$.
There was no such τ^2 that produced a mean I^2 of 90% when $\theta = 2.3$, so these scenarios are not presented.

		Study sizes				
		small	**small to medium**	**medium**	**small and large**	**large**
Number of studies (k)	**2**	0 (0)	0 (0)	0 (0)	4.2 (<0.01)	0 (0)
	3	0 (0)	0 (0)	0 (0)	4.2 (<0.01)	0 (0)
	5	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	10	0 (0)	4.2 (<0.01)	2.1 (<0.01)	0 (0)	0 (0)
	20	0 (0)	8.3 (<0.01)	0 (0)	0 (0)	0 (0)
	30	0 (0)	2.1 (<0.01)	0 (0)	0 (0)	0 (0)
	50	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	100	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)

Table F.2: The percentage of scenarios and meta-analyses in which ML failed to converge

Numbers outside the round brackets represent the percentage of scenarios where at least one meta-analysis failed to converge to a ML estimate (out of a total of 48). Numbers inside the round brackets represent the percentage of failed ML convergence in scenarios that contain failures.

		Study sizes				
		small	**small to medium**	**medium**	**small and large**	**large**
Number of studies (k)	**2**	0 (0)	0 (0)	0 (0)	31 (0.02)	0 (0)
	3	0 (0)	6 (<0.01)	2 (<0.01)	83 (0.02)	0 (0)
	5	2 (<0.01)	2 (<0.01)	0 (0)	15 (<0.01)	0 (0)
	10	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	20	0 (0)	4 (<0.01)	0 (0)	0 (0)	0 (0)
	30	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	50	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	100	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)

Table F.3: The percentage of scenarios and meta-analyses in which REML failed to converge

Numbers outside the round brackets represent the percentage of scenarios where at least one meta-analysis failed to converge to a REML estimate (out of a total of 48). Numbers inside the round brackets represent the percentage of failed REML convergence in scenarios that contain failures.

Performance measure	**Reason for exclusion**	**Figures of selected results**
Median bias of the heterogeneity variance estimate	The median is not meaningful in the many scenarios when >50% of heterogeneity variance estimates are zero.	Figure F.1, shown in comparison with results of mean bias.
Median squared error of the heterogeneity variance estimate		Figure F.2, shown in comparison with results of mean squared error.
Mean squared error of estimate of the summary effect	Preliminary analysis showed all heterogeneity variance estimators have almost identical mean squared errors. Therefore, the only observation that can be made is that mean squared error decreases as the number of studies/size of studies increase, but this is to be expected and trivial.	Figure F.3
Power to detect a significant summary effect	Preliminary analysis showed all heterogeneity variance estimators have almost identical power. Therefore, the only observation that can be made is that power increases as the number of studies/size of studies increase, but this is to be expected and trivial.	Figure F.4
Mean of the error-interval estimation of effect	Preliminary analysis showed this measure is highly correlated with mean squared error of the heterogeneity variance. This is perhaps because a high mean squared error causes more variability in confidence interval widths.	Figure F.5
Variance of the error-interval estimation of effect		Figure F.6

Table F.1: Excluded performance measures and reasons for exclusion

		Study sizes:	Small				Small-to-medium				Medium				Small and large				Large			
		θ parameter:	0	0.5	1.1	2.3	0	0.5	1.1	2.3	0	0.5	1.1	2.3	0	0.5	1.1	2.3	0	0.5	1.1	2.3
OR meta-analyses with event probability:	0.5	$I^2 = 0\%$	<0.001	<0.001	<0.001	0.008	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.004	<0.001	<0.001	<0.001	<0.001
		$I^2 = 30\%$	<0.001	<0.001	0.001	0.013	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.004	<0.001	<0.001	<0.001	<0.001
		$I^2 = 60\%$	<0.001	0.001	0.002	0.028	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.005	<0.001	<0.001	<0.001	<0.001
		$I^2 = 90\%$	0.032	0.029	0.061	0.15	<0.001	<0.001	<0.001	0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.001	0.009	<0.001	<0.001	<0.001	<0.001
	0.1 to 0.5	$I^2 = 0\%$	0.028	0.035	0.064	0.21	0.008	0.009	0.018	0.068	<0.001	<0.001	<0.001	0.001	0.026	0.032	0.059	0.184	<0.001	<0.001	<0.001	<0.001
		$I^2 = 30\%$	0.034	0.041	0.073	0.223	0.008	0.01	0.019	0.069	<0.001	<0.001	<0.001	0.001	0.026	0.033	0.06	0.184	<0.001	<0.001	<0.001	<0.001
		$I^2 = 60\%$	0.054	0.063	0.1	0.251	0.009	0.011	0.02	0.074	<0.001	<0.001	<0.001	0.002	0.027	0.033	0.061	0.184	<0.001	<0.001	<0.001	<0.001
		$I^2 = 90\%$	0.274	0.355	0.35	-	0.018	0.021	0.033	0.103	<0.001	<0.001	<0.001	0.009	0.038	0.046	0.075	0.193	<0.001	<0.001	<0.001	<0.001
	0.05	$I^2 = 0\%$	0.588	0.609	0.679	0.848	0.076	0.088	0.138	0.389	<0.001	<0.001	0.005	0.138	0.326	0.337	0.371	0.455	<0.001	<0.001	<0.001	<0.001
		$I^2 = 30\%$	0.63	0.644	0.69	0.803	0.084	0.098	0.153	0.401	<0.001	0.001	0.009	0.164	0.328	0.338	0.371	0.453	<0.001	<0.001	<0.001	<0.001
		$I^2 = 60\%$	0.705	0.707	0.726	0.771	0.113	0.128	0.193	0.42	0.001	0.004	0.023	0.207	0.332	0.341	0.372	0.45	<0.001	<0.001	<0.001	<0.001
		$I^2 = 90\%$	-	-	-	-	-	-	-	-	0.088	0.108	0.166	-	0.374	0.378	0.406	-	<0.001	<0.001	<0.001	0.005
	0.01	$I^2 = 0\%$	0.967	0.969	0.975	0.989	0.613	0.633	0.697	0.856	0.25	0.283	0.396	0.698	0.651	0.652	0.655	0.685	<0.001	<0.001	0.001	0.071
		$I^2 = 30\%$	0.97	0.971	0.972	0.974	0.654	0.666	0.713	0.817	0.299	0.329	0.424	0.658	0.651	0.652	0.656	0.699	<0.001	<0.001	0.002	0.092
		$I^2 = 60\%$	-	-	-	-	0.727	0.732	0.75	0.794	0.404	0.421	0.482	0.639	0.653	0.656	0.664	0.726	<0.001	0.001	0.007	0.128
		$I^2 = 90\%$	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.02	0.03	0.073	0.27

Table F.4: Proportion of studies with zero events in either study arm in simulated odds ratio meta-analyses

Percentages represent where θ_i are normally distributed.

Missing values indicate scenarios where no simulations were run. In these scenarios, it was not possible to derive a τ^2 parameter value that results in meta-analyses with the required average I^2.

Abbreviations

CDSR	Cochrane Database of Systematic Reviews
FDA	Food and Drug Administration
HK	Hartung-Knapp (confidence interval method)
IPD	Individual Participant Data
MD	Mean Difference
NICE	National Institute for Clinical Excellence
OR	Odds Ratio
RD	Risk Difference
ROC	Receiver Operating Curve
RR	Relative Risk
SMD	Standardised Mean Difference

The following are abbreviations used for heterogeneity variance estimation methods:

CA	Cochran's ANOVA
DL	DerSimonian-Laird
DL_P	Positive DerSimonian-Laird
PM	Paule-Mandel
PM_{CA}	Two-step Cochran's ANOVA
PM_{DL}	Two-step DerSimonian-Laird
HM	Hartung-Makambi
HS	Hunter-Schmidt

SJ	Sidik-Jonkman
SJ_{CA}	Improved Sidik-Jonkman
ML	Maximum Likelihood
REML	Restricted Maximum Likelihood
ARML	Approximate Restricted Maximum Likelihood
FB	Full Bayes
AB	Approximate Bayes
BM	Bayes modal
B0	Rukhin (zero prior)
BP	Rukhin (simple)
SB	Rukhin (alternate)
DL_B	Bootstrap DerSimonian-Laird
MBH	Malzahn, Böhning and Holling

Bibliography

[1] Ahn, S., Ames, A. J., and Myers, N. D. 2012. A review of meta-analyses in education: Methodological strengths and weaknesses. *Rev Educ Res 82,* 4, 436–476.

[2] Berkey, C. S., Hoaglin, D. C., Mosteller, F., and Colditz, G. A. 1995. A random-effects regression model for meta-analysis. *Stat Med 14,* 4, 395–411.

[3] Bhaumik, D. K., Amatya, A., Normand, S.-L. T., Greenhouse, J., Kaizar, E., Neelon, B., and Gibbons, R. D. 2012. Meta-analysis of rare binary adverse event data. *J Am Stat Assoc 107,* 498, 555–567.

[4] Biondi-Zoccai, G., Lotrionte, M., Landoni, G., and Modena, M. 2011. The rough guide to systematic reviews and meta-analyses. *HSR Proc Intensive Care Cardiovasc Anesth 3,* 3, 161.

[5] Bland, M. and Altman, D. 1986. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet 327,* 8476, 307–310.

[6] Böhning, D. and Jr., J. S. 2000. Estimating risk difference in multicenter studies under baseline-risk heterogeneity. *Biometrics 56,* 1, pp. 304–308.

[7] Böhning, D., Malzahn, U., Dietz, E., Schlattmann, P., Viwatwongkasem, C., and Biggeri, A. 2002. Some general points in estimating heterogeneity variance with the DerSimonian-Laird estimator. *Biostatistics 3,* 4, 445–457.

[8] Borenstein, M., Hedges, L. V., and Higgins, J. P. T. 2009. *Introduction to Meta-Analysis.* Wiley, Hoboken, NJ, USA.

[9] Bowden, J., Tierney, J., Copas, A., and Burdett, S. 2011. Quantifying, displaying and accounting for heterogeneity in the meta-analysis of RCTs using standard and generalised Q statistics. *BMC Med Res Methodol 11,* 1, 41.

[10] Bradburn, M. J., Deeks, J. J., Berlin, J. A., and Russell Localio, A. 2007. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. *Stat Med 26,* 1, 53–77.

[11] Brockwell, S. E. and Gordon, I. R. 2001. A comparison of statistical methods for meta-analysis. *Stat Med 20,* 6, 825–840.

[12] Brockwell, S. E. and Gordon, I. R. 2007. A simple method for inference on an overall effect in meta-analysis. *Stat Med 26,* 25, 4531–4543.

[13] Centre for Reviews and Dissemination. 2009. *Systematic review: CRD's guidance for undertaking reviews in health care.* York: University of York.

[14] Cheung, M. W.-L. 2013. Implementing restricted maximum likelihood estimation in structural equation models. *Struct Equ Modeling 20,* 1, 157–167.

[15] Chinn, S. 2000. A simple method for converting an odds ratio to effect size for use in meta-analysis. *Stat Med 19,* 22, 3127–3131.

[16] Chung, Y., Rabe-Hesketh, S., and Choi, I.-H. 2013. Avoiding zero between-study variance estimates in random-effects meta-analysis. *Stat Med 32,* 23, 4071–4089.

[17] Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A., and Liu, J. 2013. A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika 78,* 4, 685–709.

[18] Cochran, W. G. 1954. The combination of estimates from different experiments. *Biometrics 10,* 1, 101–129.

[19] Cooper HM and Hedges LV (editors). 1994. *The Handbook of Research Synthesis.* Russell Sage Foundation.

[20] Cornell, J. E. 2014. Random-effects meta-analysis of inconsistent effects: a time for change. *Annals of Internal Medicine 160,* 4, 267–270.

[21] Davey, J., Turner, R. M., Clarke, M. J., and Higgins, J. P. 2011. Characteristics of meta-analyses and their component studies in the Cochrane Database of Systematic Reviews: a cross-sectional, descriptive analysis. *BMC Med Res Methodol 11,* 1, 1.

[22] Deeks, J. and Higgins, J. P. T. 2010. Statistical algorithms in review manager 5.

[23] Deeks, J. J. 2001. Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests. *BMJ 323,* 7305, 157.

[24] DerSimonian, R. and Kacker, R. 2007. Random-effects model for meta-analysis of clinical trials: An update. *Contemp Clin Trials 28,* 2, 105–114.

[25] DerSimonian, R. and Laird, N. 1986. Meta-analysis in clinical trials. *Control Clin Trials 7,* 3, 177–188.

[26] Efron, B. 1981. Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika 68,* 3, 589–599.

[27] Egger, M., Smith, G. D., Schneider, M., and Minder, C. 1997. Bias in meta-analysis detected by a simple, graphical test. *BMJ 315,* 7109 (9), 629–634.

[28] Follmann, D. A. and Proschan, M. A. 1999. Valid inference in random effects meta-analysis. *Biometrics 55,* 3, 732–737.

[29] Food and Drug Administration Centre for Drugs Evaluation Research. 2009. Summary of published research on the beneficial effects of fish consumption and omega-3 fatty acids for certain neurodevelopmental and cardiovascular endpoints. FDA Maryland.

[30] Friedrich, J. O., Adhikari, N. K., and Beyene, J. 2007. Inclusion of zero total event trials in meta-analyses maintains analytic consistency and incorporates all available data. *BMC Med Res Methodol 7,* 1, 5.

[31] Glass, G., McGaw, B., and Smith, M. 1981. *Meta-analysis in social research.* Sage Library of Social Research. Sage Publications.

[32] Glass, G. V. 1976. Primary, secondary, and meta-analysis of research. *Educ Res 5,* 10, 3–8.

[33] Glasziou, P. and Sanders, S. 2002. Investigating causes of heterogeneity in systematic reviews. *Stat Med 21,* 11, 1503–1511.

[34] Glynn, L. G., Murphy, A. W., Smith, S. M., Schroeder, K., and Fahey, T. 2010. Interventions used to improve control of blood pressure in patients with hypertension. *Cochrane Libr.*

[35] Hamada, M. S., Wilson, A. G., Reese, C. S., and Martz, H. F. 2008. Using degradation data to assess reliability. *Bayesian Reliability*, 271–317.

[36] Handoll, H. H., Gillespie, W. J., Gillespie, L. D., and Madhok, R. 2008. The cochrane collaboration: A leading role in producing reliable evidence to inform healthcare decisions in musculoskeletal trauma and disorders. *Indian J Orthop 42,* 3, 247.

[37] Hardy, R. J. and Thompson, S. G. 1996. A likelihood approach to meta-analysis with random effects. *Stat Med 15,* 6, 619–629.

[38] Hartung, J. 1999. An alternative method for meta-analysis. *Biom J 41,* 8, 901–916.

[39] Hartung, J. and Knapp, G. 2001. A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Stat Med 20,* 24, 3875–3889.

[40] Hartung, J. and Makambi, K. H. 2003. Reducing the number of unjustified significant results in meta-analysis. *Commun Stat Simul Comput 32,* 4, 1179–1190.

[41] Harville, D. A. 1977. Maximum likelihood approaches to variance component estimation and to related problems. *J Am Stat Assoc 72,* 358, 320–338.

[42] Hedges, L. and Olkin, I. 1985. *Statistical Method for Meta-Analysis.* Acad. Press.

[43] Hedges, L. V. 1981. Distribution theory for glass's estimator of effect size and related estimators. *J Educ Behav Stat 6,* 2, 107–128.

[44] Hedges, L. V. and Pigott, T. D. 2001. The power of statistical tests in meta-analysis. *Psychol Methods 6,* 3, 203.

[45] Henmi, M. and Copas, J. B. 2010. Confidence intervals for random effects meta-analysis and robustness to publication bias. *Stat Med 29,* 29, 2969–2983.

[46] Higgins, J. P. and Spiegelhalter, D. J. 2002. Being sceptical about meta-analyses: a Bayesian perspective on magnesium trials in myocardial infarction. *Int J Epidemiol 31,* 1, 96–104.

[47] Higgins, J. P. T. and Thompson, S. G. 2002. Quantifying heterogeneity in a meta-analysis. *Stat Med 21,* 11, 1539–1558.

[48] Higgins, J. P. T., Thompson, S. G., and Spiegelhalter, D. J. 2009. A re-evaluation of random-effects meta-analysis. *J R Stat Soc Ser A Stat Soc 172,* 1, 137–159.

[49] Higgins, J. P. T. and Whitehead, A. 1996. Borrowing strength from external trials in meta-analysis. *Stat Med 15,* 24, 2733–2749.

[50] Higgins, J. P. T., Whitehead, A., and Simmonds, M. 2008. Sequential methods for random-effects meta-analysis. *Stat Med 30,* 9, 903–921.

[51] Higgins, JPT and Green S (editors). 2011. *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011].* The Cochrane Collaboration, The Cochrane Collaboration.

[52] Hooper, L., Thompson, R. L., Harrison, R. A., Summerbell, C. D., Ness, A. R., Moore, H. J., Worthington, H. V., Durrington, P. N., Higgins, J. P., Capps, N. E., et al. 2004. Omega 3 fatty acids for prevention and treatment of cardiovascular disease. *Cochrane Libr.*

[53] Hunter, J. and Schmidt, F. 2004. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings.* SAGE Publications.

[54] IBM Corp [computer program] Version 22.0. 2013. IBM SPSS Statistics for Windows. Armonk, NY: IBM Corp.

[55] IntHout, J., Ioannidis, J. P., and Borm, G. F. 2014. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Med Res Methodol 14,* 1, 25.

[56] Ioannidis, J. P. A., Patsopoulos, N. A., and Evangelou, E. 2007. Uncertainty in heterogeneity estimates in meta-analyses. *BMJ 335,* 7626, 914–916.

[57] Jackson, D. 2006. The power of the standard test for the presence of heterogeneity in meta-analysis. *Stat Med 25,* 15, 2688–99.

[58] Jennrich, R. I. and Sampson, P. 1976. Newton-Raphson and related algorithms for maximum likelihood variance component estimation. *Technometrics 18,* 1, 11–17.

[59] Kacker, R. N. 2004. Combining information from interlaboratory evaluations using a random effects model. *Metrologia 41,* 3, 132.

[60] Knapp, G., Biggerstaff, B. J., and Hartung, J. 2006. Assessing the amount of heterogeneity in random-effects meta-analysis. *Biom J 48,* 2, 271–85.

[61] Knapp, G. and Hartung, J. 2003. Improved tests for a random effects meta-regression with a single covariate. *Stat Med 22,* 17, 2693–2710.

[62] Kontopantelis, E. and Reeves, D. 2012a. Performance of statistical methods for meta-analysis when true study effects are non-normally distributed: A comparison between Dersimonian-Laird and restricted maximum likelihood. *Stat Methods Med Res 21,* 6, 657–659.

[63] Kontopantelis, E. and Reeves, D. 2012b. Performance of statistical methods for meta-analysis when true study effects are non-normally distributed: A simulation study. *Stat Methods Med Res 21,* 4, 409–426.

[64] Kontopantelis, E., Springate, D. A., and Reeves, D. 2013. A re-analysis of the cochrane library data: the dangers of unobserved heterogeneity in meta-analyses. *PLoS ONE 8,* 7.

[65] Kremers, H. M., Myasoedova, E., Crowson, C. S., Savova, G., Gabriel, S. E., and Matteson, E. L. 2011. The Rochester Epidemiology Project: exploiting the capabilities for population-based research in rheumatic diseases. *Rheumatology 50,* 1, 6–15.

[66] Lambert, P. C., Sutton, A. J., Burton, P. R., Abrams, K. R., and Jones, D. R. 2005. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Stat Med 24,* 15, 2401–2428.

[67] Langan, D., Higgins, J. P. T., and Simmonds, M. 2015. An empirical comparison of heterogeneity variance estimators in 12894 meta-analyses. *Res Synth Methods 6,* 2, 195–205.

[68] Lee, K. J. and Thompson, S. G. 2008. Flexible parametric models for random-effects distributions. *Stat Med 27,* 3, 418–434.

[69] Lee, W.-L., Bausell, R. B., and Berman, B. M. 2001. The growth of health-related meta-analyses published from 1980 to 2000. *Eval Health Prof 24,* 3, 327–335.

[70] Leeflang, M. M., Deeks, J. J., Takwoingi, Y., and Macaskill, P. 2013. Cochrane diagnostic test accuracy reviews. *Systematic Reviews 2,* 1, 82.

[71] Littell, J., Corcoran, J., and Pillai, V. 2008. *Systematic Reviews and Meta-Analysis.* Oxford University Press, USA.

[72] Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. 2000. Winbugs-a bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing 10,* 4, 325–337.

[73] Makambi, K. H. 2004. The effect of the heterogeneity variance estimator on some tests of treatment efficacy. *Journal of biopharmaceutical statistics 14,* 2, 439–449.

[74] Malzahn, U., Böhning, D., and Holling, H. 2000. Nonparametric estimation of heterogeneity variance for the standardised difference used in meta-analysis. *Biometrika 87,* 3, 619–632.

[75] Mantel, N. and Haenszel, W. 1959. Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst. 22*, 719–748.

[76] Moher, D., Liberati, A., Tetzlaff, J., and Altman, D. G. 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ 339,* 4, 264–269.

[77] Morris, C. N. 1983. Parametric empirical bayes inference: Theory and applications. *J Am Stat Assoc 78,* 381, 47–55.

[78] Novianti, P. W., Roes, K. C., and van der Tweel, I. 2014. Estimation of between-trial variance in sequential meta-analyses: A simulation study. *Contemp Clin Trials 37,* 1, 129–138.

[79] Panityakul, T., Bumrungsup, C., and Knapp, G. 2013. On estimating residual heterogeneity in random-effects meta-regression: A comparative study. *J. Stat. Theory Appl 12,* 3, 253–265.

[80] Paule, R. and Mandel, J. 1982. Consensus values and weighting factors. *J Res Natl Bur Stand 87,* 5, 377–385.

[81] Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., Mantel, N., McPherson, K., Peto, J., and Smith, P. G. 1976. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. *Br J Cancer 34,* 6, 585–612.

[82] Pittler, M., Schmidt, K., and Ernst, E. 2008. Hawthorn extract for treating chronic heart failure. *Cochrane Database Syst Rev 1.*

[83] Preuss, M. and Ziegler, A. 2014. A simplification and implementation of random-effects meta-analyses based on the exact distribution of Cochran's Q. *Methods Inf Med 53,* 1, 54–61.

[84] Qamar, N., Bray, E. P., Glynn, L. G., Fahey, T., Mant, J., Holder, R. L., and McManus, R. 2013. Self-monitoring for improving control of blood pressure in patients with hypertension. *The Cochrane Library.*

[85] R Development Core Team. 2008. R: A language and environment for statistical computing. ISBN 3-900051-07-0.

[86] Raghunathan, T. et al. 1993. Analysis of binary data from a multicentre clinical trial. *Biometrika 80,* 1, 127–139.

[87] Review Manager (RevMan) [computer program] Version 5.3. 2014. Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration.

[88] Rhodes, K. M., Turner, R. M., and Higgins, J. P. 2015. Predictive distributions were developed for the extent of heterogeneity in meta-analyses of continuous outcome data. *J Clin Epidemiol 68,* 1, 52–60.

[89] Riley, R. D., Gates, S., Neilson, J., and Alfirevic, Z. 2011. Statistical methods can be improved within cochrane pregnancy and childbirth reviews. *J C 64,* 6, 608–618.

[90] Riley, R. D., Lambert, P. C., and Abo-Zaid, G. 2010. Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ 340,* 7745, 521–525.

[91] Ritchie, L. D., Campbell, N. C., and Murchie, P. 2011. New nice guidelines for hypertension. *BMJ 343.*

[92] Rothstein, H. R., Sutton, A. J., and Borenstein, M. 2006. *Publication bias in meta-analysis: Prevention, assessment and adjustments.* John Wiley & Sons.

[93] Rukhin, A. L. 2013. Estimating heterogeneity variance in meta-analysis. *J R Stat Soc Series B Stat Methodol 75*, 451–469.

[94] Rukhin, A. L., Biggerstaff, B. J., and Vangel, M. G. 2000. Restricted maximum likelihood estimation of a common mean and the Mandel-Paule algorithm. *J Stat Plan Inference 83,* 2, 319–330.

[95] Rukhin, A. L. and Vangel, M. G. 1998. Estimation of a common mean and weighted means statistics. *J Am Stat Assoc 93,* 441, 303–308.

[96] Sanchez-Meca, J. and Marín-Martínez, F. 2008. Confidence intervals for the overall effect size in random-effects meta-analysis. *Psychol Methods 13,* 1, 31.

[97] SAS [computer program] Version 6.0. 1990. Cary, North Carolina: SAS Institute.

[98] Schilling, M. F., Watkins, A. E., and Watkins, W. 2002. Is human height bimodal? *Am Stat 56,* 3, 223–229.

[99] Schwarzer, G. 2015. *meta: General Package for Meta-Analysis.* R package version 4.3-0.

[100] Sidik, K. and Jonkman, J. N. 2002. A simple confidence interval for meta-analysis. *Stat Med 21,* 21, 3153–3159.

[101] Sidik, K. and Jonkman, J. N. 2005. Simple heterogeneity variance estimation for meta-analysis. *J R Stat Soc Ser C Appl Stat 54,* 2, 367–384.

[102] Sidik, K. and Jonkman, J. N. 2007. A comparison of heterogeneity variance estimators in combining results of studies. *Stat Med 26,* 9, 1964–1981.

[103] Simmonds, M. C. and Higgins, J. P. 2014. A general framework for the use of logistic regression models in meta-analysis. *Stat Methods Med Res*.

[104] Simmonds, M. C., Higginsa, J. P., Stewartb, L. A., Tierneyb, J. F., Clarke, M. J., and Thompson, S. G. 2005. Meta-analysis of individual patient data from randomized trials: a review of methods used in practice. *Clin Trials 2,* 3, 209–217.

[105] Smith, C. T., Williamson, P. R., and Marson, A. G. 2005. Investigating heterogeneity in an individual patient data meta-analysis of time to event outcomes. *Stat Med 24,* 9, 1307–1319.

[106] Smith, T. C., Spiegelhalter, D. J., and Thomas, A. 1995. Bayesian approaches to random-effects meta-analysis: A comparative study. *Stat Med 14,* 24, 2685–2699.

[107] Starr, M., Chalmers, I., Clarke, M., and Oxman, A. D. 2009. The origins, evolution, and future of The Cochrane Database of Systematic Reviews. *Int J Technol Assess Health Care 25*, 182–195.

[108] StataCorp [computer program] Release 14. 2015. Stata statistical software. College Station, TX: StataCorp LP.

[109] Sterne, J. A., Gavaghan, D., and Egger, M. 2000. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *J Clin Epidemiol 53,* 11, 1119–1129.

[110] Sterne, J. A., Smith, G. D., and Cox, D. 2001. Sifting the evidence-what's wrong with significance tests? Another comment on the role of statistical methods. *BMJ. 322,* 7280, 226–231.

[111] Sutton, A. J. and Abrams, K. R. 2001. Bayesian methods in meta-analysis and evidence synthesis. *Stat Methods Med Res 10,* 4, 277–303.

[112] Sutton, A. J. and Higgins, J. P. T. 2008. Recent developments in meta-analysis. *Stat Med 27,* 5, 625–650.

[113] Sweeting, M. J., Sutton, A., and Lambert, P. 2004. What to add to nothing? use and avoidance of continuity corrections in meta-analysis of sparse data. *Stat Med 23,* 9, 1351–1375.

[114] Teo, K. K., Yusuf, S., Collins, R., Held, P. H., and Peto, R. 1991. Effects of intravenous magnesium in suspected acute myocardial infarction: overview of randomised trials. *BMJ 303,* 6816, 1499–1503.

[115] Thompson, S. G. and Higgins, J. P. T. 2002. How should meta-regression analyses be undertaken and interpreted? *Stat Med 21,* 11, 1559–1573.

[116] Thompson, S. G. and Sharp, S. J. 1999. Explaining heterogeneity in meta-analysis: a comparison of methods. *Stat Med 18,* 20, 2693–2708.

[117] Thorlund, K., Wetterslev, J., Awad, T., Thabane, L., and Gluud, C. 2011. Comparison of statistical inferences from the Dersimonian-Laird and alternative random-effects model meta-analyses - an empirical assessment of 920 Cochrane primary outcome meta-analyses. *Res Synth Methods 2,* 4, 238–253.

[118] Tierney, J. F., Stewart, L. A., Ghersi, D., Burdett, S., and Sydes, M. R. 2007. Practical methods for incorporating summary time-to-event data into meta-analysis. *Trials 8,* 1, 16.

[119] Tierney, J. F., Vale, C., Riley, R., Smith, C. T., Stewart, L., Clarke, M., and Rovers, M. 2015. Individual Participant Data (IPD) meta-analyses of randomised controlled trials: Guidance on their use. *PLoS Med 12,* 7.

[120] Turner, R. M., Davey, J., Clarke, M. J., Thompson, S. G., and Higgins, J. P. 2012. Predicting the extent of heterogeneity in meta-analysis, using

empirical data from the Cochrane Database of Systematic Reviews. *Int J Epidemiol 41,* 3, 818–827.

[121] Van Houwelingen, H. C. and Zwinderman, K. H. 1993. A bivariate approach to meta-analysis. *Stat Med 12,* 24, 2273–2284.

[122] Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J. P., Langan, D., and Salanti, G. 2015. Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Res Synth Methods 7*, 55–79.

[123] Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Knapp, G., Kuss, O., and Langan, D. 2015. Recommendations for quantifying the uncertainty in the summary intervention effect and estimating the between-study heterogeneity variance in random-effects meta-analysis. *Cochrane Methods 13*, 25.

[124] Viechtbauer, W. 2005. Bias and efficiency of meta-analytic variance estimators in the random-effects model. *J Educ Behav Stat 30,* 3, 261–293.

[125] Viechtbauer, W. 2007. Confidence intervals for the amount of heterogeneity in meta-analysis. *Stat Med 26,* 1, 37–52.

[126] Viechtbauer, W. 2010. Conducting meta-analyses in R with the metafor package. *J Stat Softw 36,* 3, 1–48.

[127] Wetterslev, Jorn Thorlund, K., Brok, J., and Gluud, C. 2009. Estimating required information size by quantifying diversity in random-effects model meta-analyses. *BMC Med Res Methodol 9,* 1, 86.

[128] Whitehead, A. and Wiley, J. 2002. *Meta-analysis of controlled clinical trials.* John Wiley & Sons West Sussex, UK.

[129] Zapfe, G. 2001. Clinical efficacy of crataegus extract ws® 1442 in congestive heart failure nyha class ii. *Phytomedicine 8,* 4, 262–266.

[130] Zhang, W. W., Speare, S., Churilov, L., Thuy, M., Donnan, G., and Bernhardt, J. 2014. Stroke rehabilitation in china: a systematic review and meta-analysis. *Int J Stroke 9,* 4, 494–502.